1.3 Human genome variation and why it matters.

*In the last chapter, we discussed the standard human Reference Genome. But in practice **everyone's genome is unique, and differs from the Reference at millions of sites**. Here we introduce the concept of genome variation and how it can change the information encoded in genomes.*

SNPs. The most abundant type of genetic variation [a] are **SNPs** (Single Nucleotide Polymorphisms, pronounced *snips*). These are simple sequence differences that affect a single nucleotide: for example in a short stretch of genome your maternal chromosome might read ATCGAAGCC, and your paternal chromosome ATCGGAGCC. Although four nucleotides would be possible at any given position, the vast majority of SNPs only have two alleles [43]:

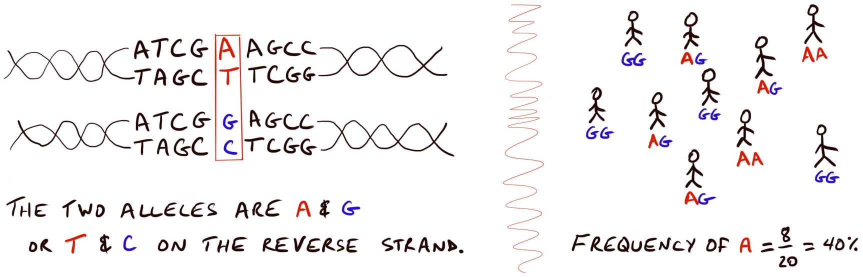

ATCG A AGCC
TAGC T TCGG
ATCG G AGCC
TAGC C TCGG
THE TWO ALLELES ARE A & G
OR T & C ON THE REVERSE STRAND.
GG
AG
AA
AG
GG
GG
AG
AA
GG
AG
FREQUENCY OF A = 8/20 = 40%

Figure 1.24: **Illustration of an A/G SNP.** *A has a sample frequency of 40%.*

Allele frequency. In the figure above, we show the genotypes for ten individuals at an A/G SNP (i.e., a SNP where the two possible alleles are A and G). Since each person carries two copies of this sequence, they can either be AA, AG, or GG. In this examples there are 8 copies of A out of 20: this gives a frequency of p=0.4 for A, and q=0.6 for G.

We will often use p and q to indicate the frequencies of two alternative alleles, where $p + q = 1$.

If you're analyzing data it's important to be keep track of **which strand of the DNA** the SNP refers to; in the example above we would consider this an A/G SNP if we're looking at one strand, but a T/C SNP on the other strand. This is especially tricky for transition mutations (A/T versus T/A or C/G vs G/C as both alleles are found on both strands). SNPs are usually labeled with respect to the strand used in the Reference Genome, or occasionally with respect to the direction of translation if the focus is on protein-coding variation.

Once we've solved the strand issue, it's also useful to have generic ways of referring to alleles so that we don't have to remember that this particular SNP is A/G and that G is more common. When you read papers, you'll see this done using one of three different naming conventions [44]:

- **Reference/Alternate allele**: The *reference* allele is the allele listed at that position in the Human Reference Genome.
- **Minor/Major allele**: The *minor* allele is the less common allele in a population (frequency $< 50\%$). **MAF** stands for *Minor Allele Frequency*.

[a] *In this chapter we'll see many different types of ways that sites or regions of the genome can differ among individuals. We can refer to these genetic differences as **variants**.*

- **Ancestral/Derived allele**: The *ancestral* allele is the allele that was present in the common ancestor of humans (this can be inferred if one allele matches the nucleotide found at this position in other great apes), while the *derived* allele is inferred to have arisen by mutation within the human population. **DAF** stands for *Derived Allele Frequency*.

Some authors reserve the term SNP for variants where both alleles are fairly "common" – often defined as MAF>1% – and use the term **SNV (single nucleotide variant)** to include sites with rarer variation. But since the cutoff is arbitrary, here we use the term SNP throughout.

Genotype frequencies and the Hardy-Weinberg model. Aside from allele frequencies, we also need to know about **genotype frequencies**: the fraction of people who have each possible pair of alleles, in this case AA, AG, and GG. People with AA or GG are **homozygotes**, while people with AG are **heterozygotes**.

Part of the power of population genetics is that we can often predict fundamental properties of populations using mathematical models. In most cases, the starting assumptions are quite simple (although sometimes the math is complicated). This brings us to the first important model of this book [b].

To keep things simple, we'll first assume that we have distinct generations, so that there is a population of parents in one generation who mate to produce kids in the next generation. Among the parents, let p represent the frequency of the A allele, and q the frequency of the G allele. We'll also assume that the parents don't choose their reproductive partners based on genotype at this locus.

So for a kid in the next generation, what is the probability that this kid will have an AA genotype?

Answer: There's a probability p that she gets an A from the mum; and similarly from the dad. So the probability that she gets AA from both parents is $p \times p$ or p^2. (The probability of two independent events occurring is the product of the probabilities.)

Using the same logic, the probability that she gets GG from both parents is q^2. Lastly, she could get A from mum and G from dad (probability pq) or G from mum and A from dad (probability qp)–those both result in her being an AG heterozygote, with total probability $2pq$.

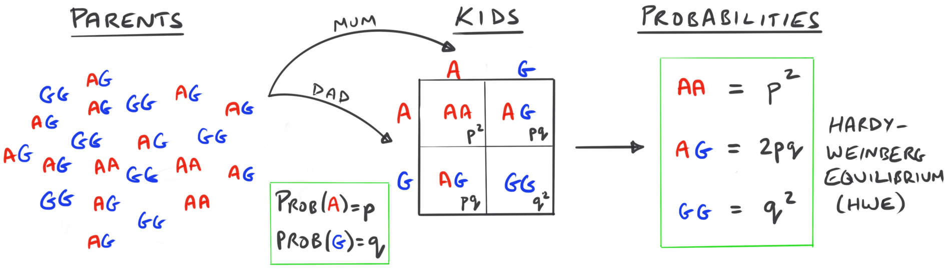


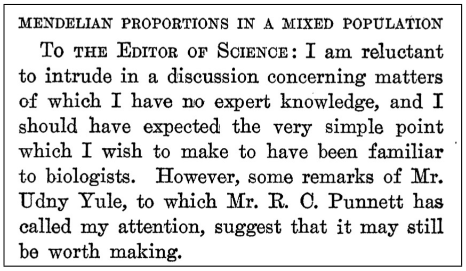
Figure 1.25: **Hardy-Weinberg.** *The genotype of a kid can be modeled as a random draw of two alleles from the population of possible parents. This means that if the allele frequencies in the parents are* **p** *and* **q***, then the genotype frequencies in the kids are* $\mathbf{p^2}$: **2pq** : $\mathbf{q^2}$.

[b] *The Hardy-Weinberg model gives us the fundamental relationship between allele frequencies and genotype frequencies.*

The key prediction here is that the expected genotype frequencies in the kids' generation is given by the proportions $\mathbf{p^2} : \mathbf{2pq} : \mathbf{q^2}$. This result is known as **Hardy-Weinberg Equilibrium (HWE)**. This result gives us the fundamental relationship between allele frequencies and genotype frequencies.

If you learned about the Hardy-Weinberg rule at school, you might have been taught a whole host of assumptions that this relies upon: for example, random mating, no selection, non-overlapping generations, etc. But in practice **Hardy-Weinberg is usually extremely accurate**, except sometimes with strong population structure. In large part, this is because just **one generation of random mating will restore a population to HW proportions** [45]. Indeed, in data analysis, if a SNP is out of Hardy-Weinberg proportions within a population, this is often taken as an indication that the genotyping may not have worked properly for that SNP [46].

Lastly, the Hardy-Weinberg result has an amusing backstory. It was first published independently in 1908 by GH Hardy, a famous English mathematician, and a German gynecologist Wilhelm Weinberg. Hardy was told about the problem of genotype frequencies by the geneticist Reginald Punnett, with whom he played cricket at Cambridge University. (You may be familiar with "Punnett Squares", used to predict the outcomes of genetic crosses.) Hardy's 1-page paper is written in a bashful tone because he thought the main result was rather beneath him (..."A little mathematics of the multiplication-table type is enough to show...") though it is now seen as the first fundamental result in population genetics. To add further insult, the article was initially rejected by the leading British journal Nature for being "tainted" with Mendelism (which was unpopular in Britain at the time, see Chapter 4.4), and it was eventually published in Science instead [47].



MENDELIAN PROPORTIONS IN A MIXED POPULATION

To THE EDITOR OF SCIENCE: I am reluctant to intrude in a discussion concerning matters of which I have no expert knowledge, and I should have expected the very simple point which I wish to make to have been familiar to biologists. However, some remarks of Mr. Udny Yule, to which Mr. R. C. Punnett has called my attention, suggest that it may still be worth making.

Figure 1.26: **The start of Hardy's 1908 paper** *on genotype frequencies.*

How many SNPs are there? Popular culture often refers to "the human genome" – but of course in practice everyone's genome is unique (unless you have an identical twin [48]).

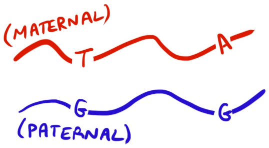


Figure 1.27: **Heterozgygosity** *measures the fraction of sites that differ between the homologous copies of someone's genome.*

Since you inherited one copy of each chromosome from your mum, and one from your dad, one way to measure genetic diversity is to count up how many differences you have between these two genome copies. Any difference between the two genomes – for example, you got an A from mum and a G from dad at a particular position – is a heterozygous SNP. So we can ask, how frequently would you find heterozygous SNPs between the homologous copies of your genome?

The answer is that you can expect to find a heterozygous SNP about once every 1,000–2,000 basepairs, depending on your ancestry. The fraction of heterozygous sites is referred to as **heterozygosity**, and is a useful measure of genetic diversity [49]. Here's a table of heterozygosity estimates from a variety of human populations:

Region	Population	Heterozygosity $\times$ 1000
Africa	San	0.95
	Yoruba	0.96
	Maasai	0.93
	Mbuti	0.91
Near East	Palestinian	0.73
	Iranian	0.71
Europe	Spanish	0.69
	Polish	0.67
South Asia	Punjabi	0.71
	Bengali	0.72
East Asia	Thai	0.69
	Japanese	0.67
Oceania	Australian	0.63
	Papuan	0.58
Americas	Inuit	0.63
	Surui	0.50

Table 1.2: Heterozygosity estimates by population, *reported as the mean number of heterozygous sites per 1000 bp. Populations such as 'Australian' refer to indigenous groups.*

Data from Supp. Table 1:AH of Swapan Mallick et al (2016) [Link].

The most striking pattern in these data above is that **heterozygosity is highest in Africa, and decreases with migration distance from Africa.** This reflects the fact that modern humans evolved in Africa; as we will discuss later in the book, modern humans spread out of northeast Africa during the last 100,000 years and eventually colonized most of the world. These population movements caused a steady loss of diversity with distance.

Another way to think about variation is in terms of the total numbers of SNPs. Since your genome is about 3.2 billion basepairs, this table implies that **you have about 1.5–3.0 million heterozygous sites, depending on your ancestry**.

What if we look at SNPs in a larger number of individuals? For example, the 1000 Genomes Project sequenced the genomes of 2500 individuals from a diverse set of global populations. They reported 85 million SNPs, most of which were very rare: 65 million were below frequency 0.5%; 12 million were between 0.5%–5%, and 8 million SNPs were above 5% [50]. In other words, **there is a common SNP with frequency$>$5% about once per 400 bp**.

It's important to note that, while a study of modest size like 1000 Genomes Project can identify essentially all common SNPs, large sequencing studies continue to find many more novel, rare SNPs [51]. Indeed, we should expect to find that **nearly every possible SNP allele exists somewhere in the world**. The world population is nearly 10^{10} people and, as we'll see in Chapter 1.5, the mutation rate is around 10^{-8} per nucleotide per generation. This implies that nearly every possible single nucleotide variant must occur many times each generation, somewhere in the world. (There are a few exceptions: a tiny fraction of possible mutations would not be observed because they disrupt biological processes so severely that they prevent embryonic or fetal development.)

Single nucleotide differences between human and chimpanzee. We don't need to limit ourselves to looking at genetic diversity within humans – we can also examine the number of sequence differences between humans and other species [c]. For example, our closest relative is the chimpanzee; our two species evolved from a common ancestor about 6.5 million years ago [52]. The average sequence divergence between the human and chimpanzee genome is about 1.37%. This is only about 15-fold higher than the divergence between the two copies of your own genome. I'll explain a bit later how to think about this 15-fold ratio. We still know very little about specifically which variants are responsible for the major phenotypic differences between humans and chimpanzees, such as our remarkable fondness for cell phones.

Genotypes and haplotypes. So far we have just been *counting* SNPs, but it will also be important to consider how they are arranged along chromosomes. The identities of alleles along one copy of a chromosome are referred to as a **haplotype**. For example, the first haplotype in the plot below is A-T-A-C-G-A, and the second is A-A-C-C-G-C. As we go through this book, we'll learn a lot from studying the structure of haplotypes:

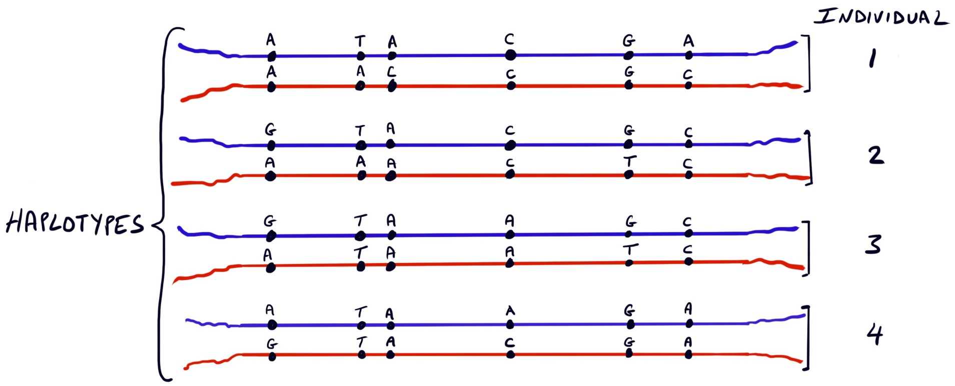


Figure 1.28: **Haplotypes for 4 individuals** *in a small region of the genome. The term* ***haplotype*** *refers to the arrangement of alleles along a homologous chromosome, or sometimes to a pattern seen in a genomic region in multiple individuals.*

Now, one challenge is that standard technologies for genome sequencing are very good at telling us the genotype at any given location, but for a heterozygous site we can't tell which allele is on which haplotype. (We'll cover sequencing in Chapter 1.4, but the issue is that standard DNA sequence reads are much shorter than the usual spacing between heterozygous sites. This is starting to change with the arrival of new long-read sequencing techniques.). So traditional sequencing does not give us the actual haplotypes, but instead genotypes like this:

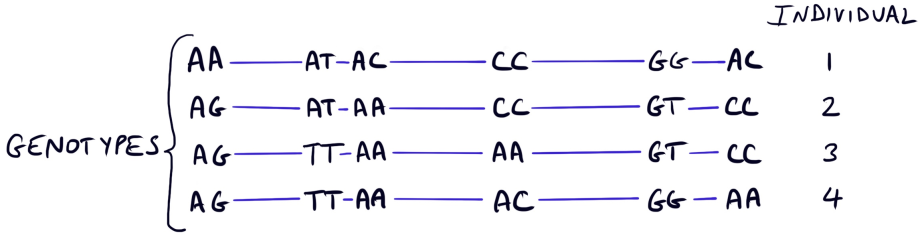

INDIVIDUAL
GENOTYPES
AA — AT-AC — CC — GG — AC 1
AG — AT-AA — CC — GT — CC 2
AG — TT-AA — AA — GT — CC 3
AG — TT-AA — AC — GG — AA 4

Figure 1.29: **Genotypes for 4 individuals** *in the same region as above. For heterozygous sites we usually do not get a direct measurement of which allele comes from which haplotype so the data for individual must be represented as genotypes.*

[c] *For thinking about human genetic variation,* ***there are a few numbers that are useful to remember****: Heterozygosity in humans is* $0.5–1 \times 10^{-3}$*, depending on population. There are* ∼*10M common SNPs. The human and chimpanzee genomes differ by 1.4% – about 40M single nucleotide differences.*

The assignment of alleles to haplotypes is referred to as **haplotype phase**; haplotypes can be estimated using statistical techniques that we will cover later.

Lastly, it's often convenient to replace the genotype matrix with a simpler version that recodes the genotypes with the allele counts: 0, 1, 2, depending on the number of minor alleles at each SNP:

$$\begin{array}{cc} & \text{SNPs} \\ \text{INDIVIDUALS} & \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 2 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 2 \end{bmatrix} \end{array}$$

Figure 1.30: **Genotype Matrix representation** *of the data above. The entries in the matrix show the numbers of minor alleles at each position: i.e., the columns show the numbers of G,A,C,A,T,A alleles, respectively).*

Example from real data. One of the first groups to study human sequence variation at large scale was Debbie Nickerson's lab at the University of Washington, in the early 2000s [53]. This plot shows the genotype matrix they obtained for the IGF1 locus on Chromosome 12, using colors to represent the genotype counts 0, 1, and 2:

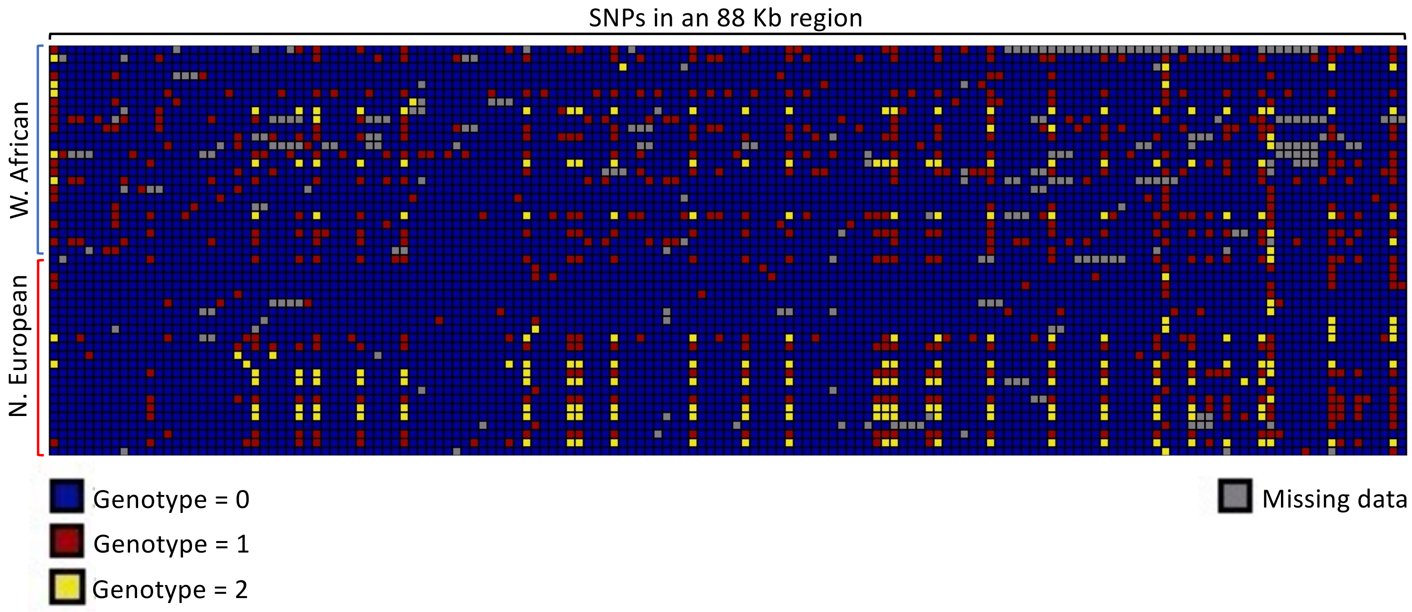


Figure 1.31: Genotype Variation at IGF1. *Each row shows the genotype for a single individual; columns are SNPs within an 88 Kb region containing the IGF1 gene, ordered by position within the region. Sequencing was PCR-based and included gaps within the region. Credit: Debbie Nickerson's lab/ Seattle SNPs project [Link].*

This example illustrates several typical features of genomic data:
- Most common variants are shared between human populations, though allele frequencies may differ.
- As expected, many of the SNPs are at low frequency. Indeed this pattern becomes increasingly stronger in larger samples, as nearly all the common SNPs are identified in a small sample like this one, while the new variants discovered with additional individuals would be rare;
- Variants at different sites often co-occur together – for example notice

the block of yellow genotypes on the left-hand side of the region where individuals who are yellow (or red) at one site are usually yellow (or red) at multiple other sites as well. This pattern of genotype correlations across sites is called **linkage disequilibrium (LD)** and will be an important topic for us later.

Beyond SNPs: Other types of inherited variation. While SNPs are the most common type of variation, there are many other ways that genomes can differ. These include small-scale events such as indels and STRs (short tandem repeats), as well as a diverse variety of larger structural elements. Collectively, we refer to all these different forms of variation as **variants**.

In Medieval books, a bestiary was a compendium of beasts (animals), both real and imagined, with pictures and descriptions; by analogy, here I show a collection of some of the main types of genetic variation. Many of the large repeats that we'll talk about soon remain slightly mysterious: they are very difficult to measure using current DNA sequencing technology, and variation in complex regions is still largely uncharacterized.

Figure 1.32: **A gryphon and a greyhound:** *two beasts from an illustrated bestiary (England ~1520). The wide array of possible types of variation – some of them difficult to glimpse with current methods – reminds me of a tableau of mysterious beasts.* *From the Tudor Pattern Book in the Bodleian Digitized Collection.*

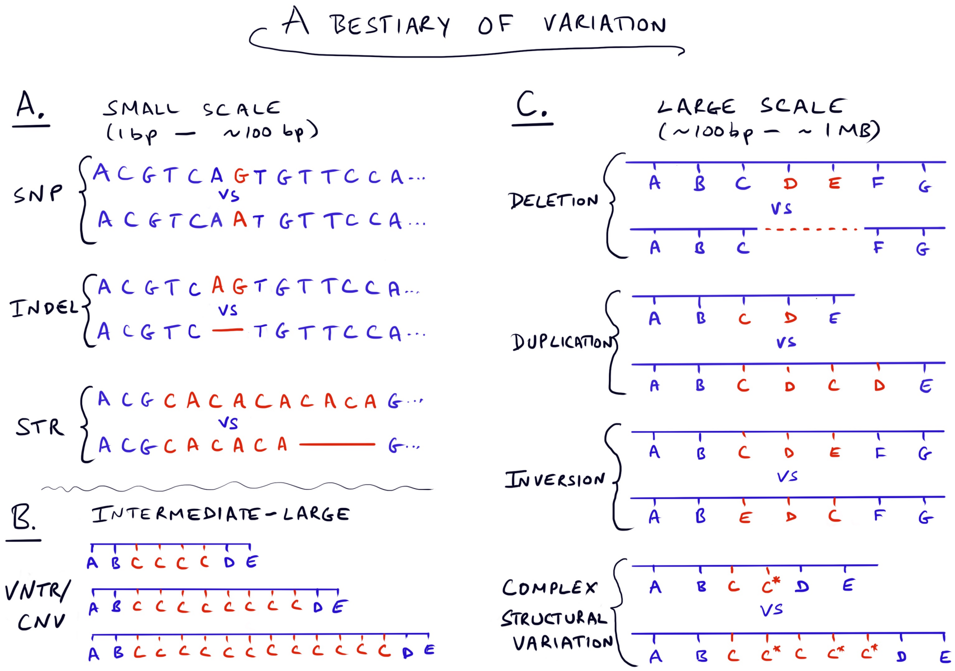


Figure 1.33: **Major types of variation. A.** *These variants affect short stretches of DNA sequence.* **B.** *and* **C.** ***Structural variants:*** *Here* ***the letters represent large blocks of DNA sequences****. These categories are often blurred, and complex structural variants often contain multiple types of events. There are many more SNPs than other kinds of variants, but because they are so large, the structural variants cover more total genome.*

Short-scale variation. The first type of short-length variation are the **indels** as shown in the picture above. This term is a mashup of the phrase **insertions/deletions**, reflecting the fact that without comparing to another genome such as chimpanzee, it is difficult to know whether an indel represents a gain, or loss, of nucleotides relative to the ancestor. These are about one tenth as many indels as SNPs, and most are very short, between 1 and 5 nucleotides in length [54]. As we mentioned above, indels are of special importance in exons, as they result in frameshifts (unless the indel length is a multiple of 3).

The next important type of short-length variation is the **STR (short tandem repeat)**. STRs are places in the genome where a short DNA sequence (up to around 6bp) is repeated many times (eg., CACACACA) – often dozens of times. During cell division, the copying of STRs is highly error-prone due to a process known as replication slippage. For this reason, STRs are highly variable within populations [d].

Intermediate-scales: VNTRs. Similar to STRs, there are larger blocks of sequences, known as **VNTRs: Variable Numbers of Tandem Repeats**, that can be duplicated many times in the genome [55]. One example is shown at right, where a block of 57bp is repeated between 22–29 times within an exon of the ACAN gene.

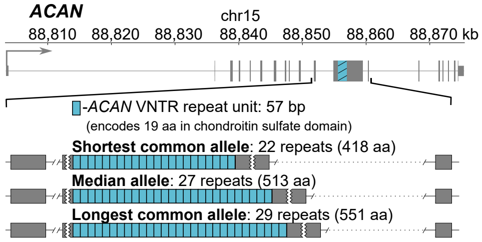

ACAN
chr15
88,810
88,820
88,830
88,840
88,850
88,860
88,870 kb
-*ACAN* VNTR repeat unit: 57 bp
(encodes 19 aa in chondroitin sulfate domain)
Shortest common allele: 22 repeats (418 aa)
Median allele: 27 repeats (513 aa)
Longest common allele: 29 repeats (551 aa)

Figure 1.34: **VNTR in the ACAN gene.** *A 57 bp (19 amino acid) repeat is present 22–29 times on different haplotypes. The ACAN protein is part of the extracellular matrix of cartilage, and larger repeat numbers are associated with higher average height.* *Figure 1a from Ronen Mukamel et al (2021) [Link]. Used with permission from the authors.*

Large-scale variation. The last major class of variation are various types of **large-scale structural variation**. Some of these are simple rearrangements of the DNA sequence, including **deletions** and **duplications**. As shown above, deletions are events that cut a segment out of a chromosome, while duplications copy a segment. An individual with a deletion would then carry one copy of any genes that lie inside that deletion (i.e., the copy on the unaffected chromosome); or three copies of those genes in the case of a duplication (two copies on the duplicated chromosome, plus one on the unaffected chromosome). Together, deletions and duplications can be referred to as **copy number variation (CNVs)**.

For reasons we'll discuss shortly, large CNVs – at megabase scale or larger – are usually highly deleterious, and can cause severe genetic syndromes in children. In contrast, smaller CNVs are often benign, especially if they do not overlap genes or key regulatory regions. A typical person carries more than 200 deletions with a median size of 7 Kb, and with 20% deleting more than 25 Kb. Numbers for duplications are similar [56]. While CNVs do not overlap genes, a small fraction – about 10% – of the deletions remove either entire genes or parts of genes (i.e., exons) [e].

Meanwhile, **inversions** take a segment of chromosome and flip it around. Inversions are much less abundant than deletions and duplications, but can be very large. The two best-characterized human inversions include one on Chromosome 8, which is a huge 4.5 Mb at around 50% global frequency, and a very interesting 900 Kb inversion on Chromosome 17 that is at 20% frequency in Europeans and may be associated with fertility and other phenotypes [57].

Lastly, some regions of the genome become crucibles of **complex structural variation**, in which the processes of replication slippage, deletion, duplication, inversion, become layered upon one another, to the extent that there may be huge numbers of different alleles. These regions are very difficult to sequence, difficult to visualize, and generally not well-characterized. However, new technologies for getting very long sequencing reads are starting to open up these regions to study. We'll see more of these topics in Chapters 1.4 and 1.5.

[d] *Because STRs are highly variable they are used in paternity testing and* ***DNA fingerprinting*** *for forensics.*

[e] *Common deletions are less likely to overlap genes than you would expect if they occurred randomly in the genome. This, and other evidence, implies that natural selection preferentially removes genic deletions.*

How do SNPs affect the information encoded in genomes? [f] Genomes are a molecular system for storing information; SNPs can alter that encoded information. Roughly speaking, SNPs and other kinds of genetic variation can have two main types of effects: they can change protein coding sequences, or they can change gene regulation.

While SNPs that affect function are of special interest for understanding disease and evolution, it's important to bear in mind that less than ∼10% of all possible SNPs have any meaningful impact on the encoded information.

1. Protein-coding variants. The figure below illustrates four important types of variants within part of a coding exon:

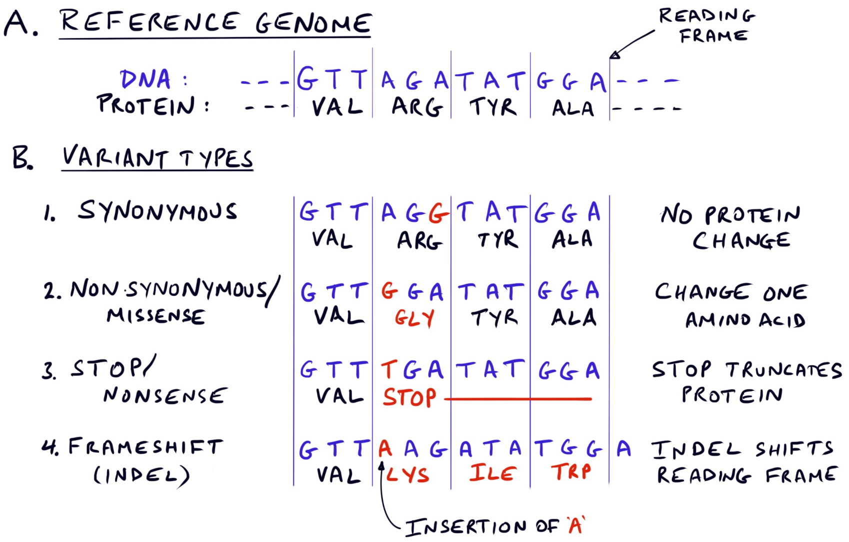


Figure 1.35: **Four types of genic variants** *in a small part of a coding region.* **A.** *Reference sequence: DNA in blue, with corresponding protein sequence in black.* **B.** *Four important categories of exonic mutations. Changes to the DNA and protein sequences are indicated in red. Frameshifts are caused by indels, which we have not covered yet:* ***indels*** *are short insertions or deletions of DNA sequence and they can shift the protein reading frame.*

- **Synonymous.** Remember that DNA encodes proteins using a genetic code in which three consecutive DNA letters correspond to amino acids: 3 DNA triplets (or codons) encode STOP signals and, together, the other 61 possible triplets code for 20 amino acids. This means that about 1/4 of possible one-step mutations simply convert between equivalent triplets. For example, in the illustration, AGA and AGG both encode Arginine. Hence, such a variant is referred to as *synonymous* in the sense that it encodes an identical protein. Synonymous variants generally do not have phenotypic effects [58].

- **Nonsynonymous/Missense.** However, many exonic mutations do change the encoded amino acid: for example AGA → GGA swaps Arginine → Glycine. Such mutations are referred to as *nonsynonymous* or *missense* as they change the meaning of the information encoded in the DNA. The functional impact of missense mutations can range from lethal to no effect, depending on what the gene is, whether the amino acid lies in a key functional domain of the protein, and the chemical properties of the original amino acid and its replacement.

- **Stop/Nonsense.** Three codons (TAA, TAG, TGA) encode the protein Stop signal. Thus for example changing AGA → TGA causes protein

[f] *We'll cover phenotypic effects of variation in Section 4 of the book, but it will be helpful for you to have this preview in mind for the upcoming chapters.*

translation to terminate. Unless a stop mutation is near the 3′ end of the coding region, it will usually obliterate protein function. Most mRNAs with premature stop codons are degraded through a process called Nonsense Mediated Decay (NMD) to prevent translation. [g]

• **Frameshift.** So far we have been talking about SNP changes, but as we'll discuss below, it's also possible to have insertions and deletions of DNA sequences. Unless these are in multiples of three nucleotides, they cause the reading frame of the protein to shift. Like stop mutations, unless these are near the end of the protein sequence, these would also generally destroy protein function.

• **Splice Site Disruption.** The next type of variant is a little different. Recall from Chapter 1.2 that genes contain introns within the coding region; these must be spliced from the transcript to produce a functional protein. The positions of exon-intron boundaries are encoded in the DNA: this code includes a required GT at the start, and AG at the end of most human introns, as well as other sequences that help position the splicing machinery. Mutation of either the GT or AG usually prevents splicing or moves the location of splicing. We'll see an example of this shortly.

As a broad generalization, **single nucleotide mutations with large effects, such as in monogenic diseases and cancer are primarily driven by effects on coding sequences**.

2. Effects on gene regulation. As we discussed in Chapter 1.2, the second important function of DNA is to encode gene regulation: how much mRNA (and protein) from any given gene should be produced in a particular cell type, or phase of development.

While the DNA encodes precise patterns of gene regulation, there is no analog to the "genetic code" for proteins. As we discussed in the previous chapter, gene regulation is achieved through a complex dance of DNA-protein and protein-protein interactions that stabilize RNA Pol-II at the promoter and enable transcription. Much of the regulatory information is encoded through sequences that control the ability of transcription factors to bind at particular sites.

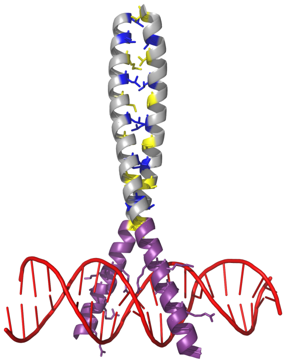


Figure 1.36: **Transcription factor binding to DNA.** *Recall that gene regulation is largely controlled by sequence-specific TF binding to DNA sequences. Thus, sequence changes can increase or decrease TF binding strength, thereby potentially changing expression.* Credit: Houq [Link] CC-BY-SA-3.0

Thus, SNP alleles can affect expression by changing the encoded regulatory information – for example by increasing or decreasing the strength of transcription factor binding at a particular site. But, because TF binding is generally controlled by multiple sites, and because expression of any single gene is usually controlled by the interplay of many proteins, **the effects of individual SNPs tend to be quantitative – slightly increasing or decreasing expression – rather than turning expression on or off.**

As a result, it is unusual for individual SNPs in gene regulatory elements to result in single-gene diseases. However, genome-wide there are hundreds of thousands of SNPs with regulatory effects, and **in aggregate regulatory SNPs are the main drivers of most common phenotypic variation, and probably evolutionary change.** [h]

[g] *Variants that destroy the functional protein are called* ***Loss of Function (LOF)****: these would include most stop mutations, splice site disruptions and frameshifts.* ***LOF mutations are usually at very low frequencies*** *due to selection against them.*

[h] *We'll come back to regulatory variation in much more detail in Section 4 of the book.*

We close out this section by reviewing the story of one of the most famous mutations in history, and its interesting mechanism.

Example: hemophilia in the royal families of Europe. One famous example of a SNP mutation comes from the inheritance of hemophilia in the European royal families of the 19th and 20th Centuries.

Hemophilia is a genetic disease, caused by mutations in either of two X chromosome genes that are essential for normal blood clotting. Since males only have one X chromosome, any male with the mutation will have the disease. In contrast, females with one copy of the mutation do not have the disease, but can transmit to their children. Prior to modern treatments, affected individuals often died at young ages, but hemophilia can now be treated using clotting factors.

The 19th century British queen, **Queen Victoria** (1819-1901), is the first person in her family known to have carried the mutation. One of her three sons had the disease; he and two of Victoria's daughters passed it into the royal families of Spain, Germany and Russia. Ultimately, eleven male-line descendants of Victoria had hemophilia, spread across 4 generations. Victoria's last known descendent with hemophilia died in 1945.

Even though Victoria's mutation no longer exists in any living person, recent genetic analysis was able to determine the causal variant.

Figure 1.37: **Tsarevich Alexei of Russia in 1916** *(front right), with his family and cossacks. Anastasia front left.* Credit: Beinecke Rare Book and Manuscript Library, Yale University; [Link].

We pick up the story with one of Victoria's great-grandsons, the **Tsarevich Alexei Nikolaevich**, born in 1904 as heir apparent to the Russia throne. Alexei inherited the hemophilia mutation from his mother, the Tsarina Alexandra. He almost died from blood loss at birth, and suffered throughout his life from dangerous hemorrhages resulting from the minor bumps and bruises of childhood. After the Russian Revolution of 1917, Alexei and his family were exiled to Siberia. The following year the Bolsheviks executed the entire family. Much later, amid persistent rumors that Alexei and one of his sisters had escaped, the remains were exhumed and eventually subjected to genetic analysis in the mid-1990s that confirmed their identities [59].

But what about the hemophilia mutation? In 2009, a Russian team sequenced the main hemophilia genes in DNA recovered from Alexei and his sister Anastasia. They identified a causal mutation in the Factor IX gene that was present on Alexei's one X chromosome, and heterozygous in Anastasia [60].

This mutation has a very interesting mechanism. Recall that the 3′ ends of introns are indicated, in part, by an AG dinucleotide. In this case, the mutation creates a new AG near the 3′ end of the intron, two base pairs upstream from the original wildtype AG splice site. The new AG is preferred by the splicing machinery, and this results in the exon being extended by two base pairs. This in turn creates a frameshift in the reading frame, leading to a nonfunctional protein:

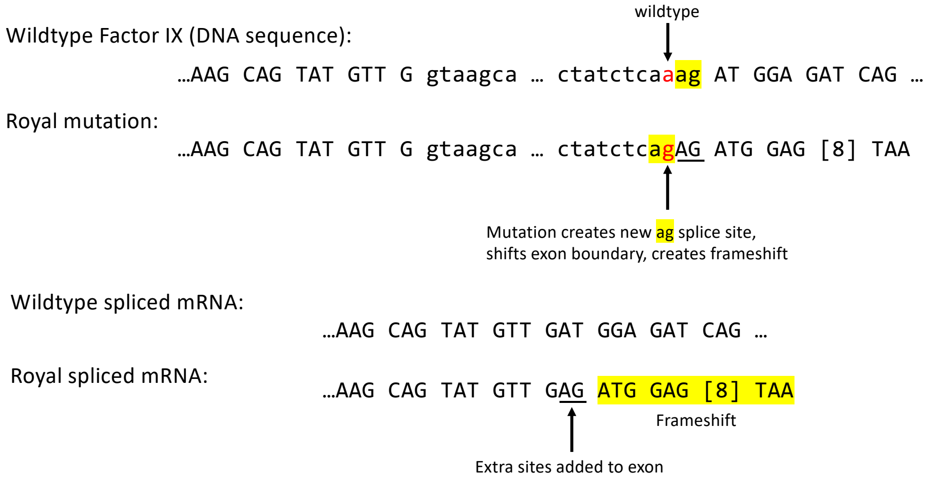

wildtype
Wildtype Factor IX (DNA sequence):
…AAG CAG TAT GTT G gtaagca … ctatctcaaag AT GGA GAT CAG …
Royal mutation:
…AAG CAG TAT GTT G gtaagca … ctatctcagAG ATG GAG [8] TAA
Mutation creates new ag splice site,
shifts exon boundary, creates frameshift
Wildtype spliced mRNA:
…AAG CAG TAT GTT GAT GGA GAT CAG …
Royal spliced mRNA:
…AAG CAG TAT GTT GAG ATG GAG [8] TAA
Frameshift
Extra sites added to exon

Figure 1.38: **Mechanism of the royal hemophilia mutation.** *Exonic nucleotides are in upper case letters and intronic in lowercase. The a→g mutation is marked in red; it creates a new 3′ ag splice site, which shifts the position of the second exon to the left by two base pairs. Thus the underlined G becomes part of the exon in the mutated gene. The lower panel shows the spliced mRNA with coding triplets indicated. Addition of G to the exon creates a frameshift (highlighted), which extends another 10 amino acids before terminating in TAA (STOP).* Credit: Redrawn from Rogaev et al 2009 [Link]

The royal mutation is just one of many different mutations that can cause hemophilia: the global prevalence of hemophilia is about 1 per 5,000 male births, caused by more than 1,000 different mutations in the two main hemophilia genes. A database of mutations in the Factor IX gene provides a sense of the relative importance of different disease mechanisms at this gene:

Mutation type	Number	% severe
missense	558	39%
frameshift	130	78%
nonsense	77	75%
splice site	83	41%
promoter	18	17%

Table 1.3: Mutation types in the Hemophilia B disease database *(Factor IX). Notice that most frameshift and nonsense mutations cause severe disease (unless they are near the end of the transcript), while other mutation types are less often severe.* Simplified data from Table 2 of Tengguo Li et al (2013) [Link]. For brevity, minor categories including structural variants are not shown.

As you can see above, most of the Hemophilia mutations affect either the protein coding sequence (missense) or entirely rewrite the protein sequence (frameshifts, nonsense, and splice site mutations).

Just a tiny fraction of the mutations are located in the promoter; these presumably change gene expression. This distribution of mutation mechanisms is typical of monogenic diseases. In contrast, we'll see later that regulatory variants are the major drivers for complex traits.

How does structural variation affect the information in genomes? As we discussed for SNPs, structural variants can affect both protein coding sequences and expression, but with a wide diversity of possible outcomes.

Changes in protein-coding sequences. Some smaller-scale variants such as STRs and VNTRs sit inside protein coding sequences; hence changes in copy number alter the protein coding sequence, as we described above for the ACAN gene which affects bone growth (and therefore height). Another famous example, Huntington's disease, is caused by an STR repeat (CAG, coding the amino acid glutamine). Alleles with large numbers of CAG produce long glutamine tracks; these form aberrant protein clumps in neurons, leading to a severe neurological symptoms. We'll come back to Huntington's Disease in Chapter 1.5.

But while this type of mechanism is important in a handful of genes, it impacts relatively few genes across the genome [61].

Copy number changes. In contrast, changes in copy number usually act through a completely different mechanism: they **alter the expression levels** of any genes contained within the affected segments: typically to 50% of wildtype for a heterozygous deletion, or 150% for a duplication.

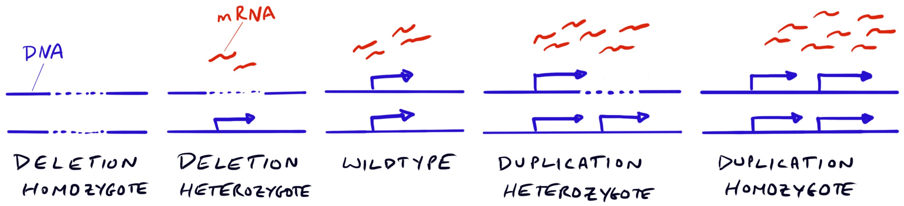


Figure 1.39: **Expression reflects copy number.** *The cartoon shows what's known as an* ***allelic series*** *in which the copy number of a particular gene (marked by the arrow) ranges from 0 to 4 copies in different individuals. mRNA (and protein) expression is usually roughly proportional to the gene's copy number.*

Does this matter?

It turns out that cellular systems are often sensitive to the precise expression levels of genes. For example, cellular differentiation is controlled by transcriptional regulatory networks, and small changes in expression of key genes can lead to widespread changes in expression. Secondly, many proteins act as components of multi-protein complexes that must be formed in precise ratios. Under- or over-expression of any component of a complex can have deleterious consequences.

Reflecting the importance of expression levels, some genes are described as **haploinsufficient**, meaning that a single functioning copy of that gene would not be sufficient for normal development or health. There are several hundred genes with known haploinsufficiency, resulting in major phenotypic effects.

Meanwhile, a much larger number of genes are **copy-number sensitive**: i.e., copy-number changes have measurable effects on survival or reproduction: it's estimated that, for about 20% of genes, loss of one copy results in a 10% loss of fitness [62]. (Here, "fitness" is a combined measure of survival and reproduction.) Thus, most large deletions or duplications (1 Mb or more, say) are extremely likely to overlap one or more copy-sensitive genes. Such large events often cause severe genetic syndromes in heterozygotes and are usually very rare in the population.

Adaptation through copy-number changes. However, copy number changes are not universally negative. Very occasionally copy number expansions evolve as a mechanism for increasing expression. For example, the amylase gene, AMY1, which is responsible for breaking down dietary starch, is present in our genome with a variable number of copies, ranging from around 2-16 copies per person [63]. These copies appear as **tandem repeats** within a single genomic region ("tandem" here meaning adjacent, rather than scattered around the genome), as you can see in the FISH image below [64]. (FISH is a technique in which DNA probes with fluorescent tags are hybridized to specific DNA sequences so that they can be imaged with microscopy.)

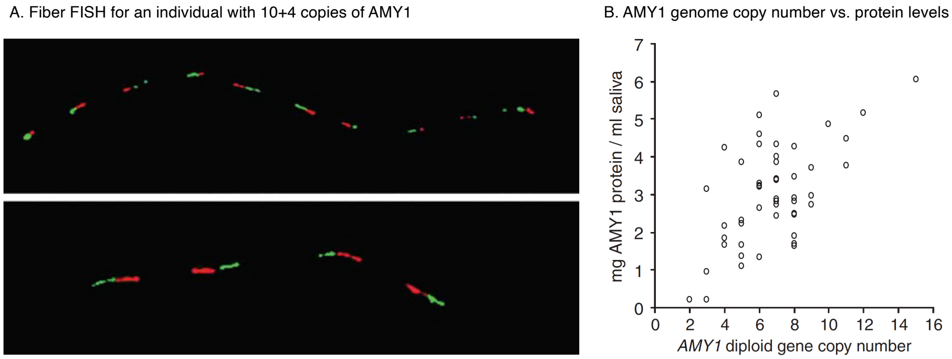


Figure 1.40: **Genome and protein variation of salivary amylase.** **A.** *Fiber FISH has been used to label individual copies of the AMY1 locus in this microscopic image. The two images show the two homologous copies of this region; each AMY1 copy is marked by one green and one red block, showing* 10 *copies on one homolog and* 4 *on the other.* **B.** *Genomic copy number of AMY1 is highly correlated with Amylase protein levels in the saliva, in a sample of* 50 *individuals.* *Credit: Modified from Figure 3a and 1c of George Perry et al (2007) [Link] Used with permission.*

As you can see above, variation in amylase copy number also has functional consequences: higher copy number is correlated with higher protein levels in the saliva, which may enhance starch digestion.

There's one more fascinating aspect to the amylase story: The copy number of amylase in humans is greatly expanded relative to other great apes (e.g., the copy number in chimpanzees is around 3 per haploid genome). This has led to the hypothesis that the copy number expansion is an evolutionary adaptation to the higher levels of starch in human diets compared with our evolutionary ancestors. Remarkably, some other species that are associated with humans (and may also have high-starch diets) also have expansions of the amylase locus, including dogs compared to wolves, and mice and rats compared to other rodents [65].

Chromosome inheritance errors: aneuploidy. Lastly, we'll talk briefly about **a completely different kind of genetic variation that is usually not inherited**: the cases where the fertilized egg inherits too many, or too few chromosomes, a situation known as **aneuploidy**. In most cases, aneuploidy has major phenotypic consequences, and is generally not transmitted to subsequent generations [i].

Recall from our last chapter that we discussed the process of **meiosis**. Most human cells contain two sets of 23 chromosomes, but eggs and sperm each carry half the usual number: 1 set of 23 chromosomes each. Meiosis is the reduction process in which the number of chromosomes is cut from 46 to 23. A fertilized egg then receives 23 chromosomes from each parent to bring it back to the correct number of 46 total chromosomes.

Sometimes there are errors in meiosis where two homologs stick together during cell division and both wind up in the same cell. When this happens, the fertilized egg ends up with either an extra chromosome (3 copies of one of the chromosomes for a total of 47) or missing a chromosome (1 copy of that chromosome for a total of 45). These errors occur most of-

[i] *Like the discussion of copy number variation above, aneuploidy illustrates a fundamental principle:* ***organisms are very sensitive to changes in the precise ratios of expression across genes****. Small variations in expression, spread across many genes, are major drivers of human phenotypic variation, evolution, and disease* [66].

ten in older mothers, for reasons we'll explain in the next chapter. (It's outside our scope in this chapter, but aneuploidy can also arise during mitosis and is a common feature of cancer genomes.)

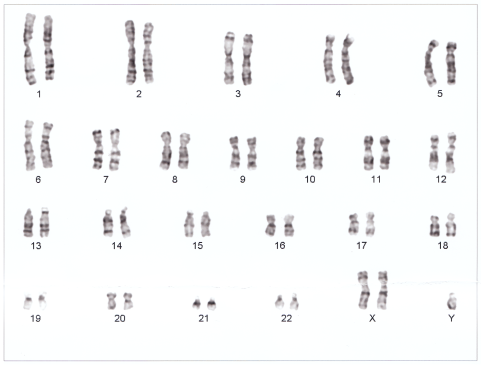


Figure 1.41: **Karyotype of a boy with Klinefelter Syndrome**: *he has two X and one Y chromosome. The chromosomes were imaged while condensed during mitosis, and then positioned in order.* Credit: Nami-ja [Link] Public Domain

Aneuploidy where a chromosome is present in single copy is referred to as **monosomy**, while aneuploidy with three copies of a chromosome is a **trisomy**. Most of the possible aneuploidies have very severe effects on global gene regulation, and the embryos do not survive to birth.

One exception is that embryos with an extra copy of the smallest chromosome, Chromosome 21, do sometimes survive to birth. These children have Down Syndrome, and usually have developmental delays and may suffer from health issues including heart problems. Children with trisomies of two other chromosomes (18 and 13) can also survive to birth, but they have severe disabilities and low survival.

Additionally, embryos with extra, or missing copies of the X and Y chromosomes also often survive to birth. Although these individuals often exhibit developmental problems, the X/Y aneuploidies are generally much less severe than autosomal aneuploidies for reasons we'll describe below. Individuals with at least one Y are usually assigned male sex at birth, regardless of the total number of Xs and Ys. This is because the SRY gene, which encodes a transcription factor that turns on male developmental programs, is located on the Y chromosome. The side image shows the **karyotype** – the number and identities of the chromosomes – in a boy with Klinefelter Syndrome (two X and one Y).

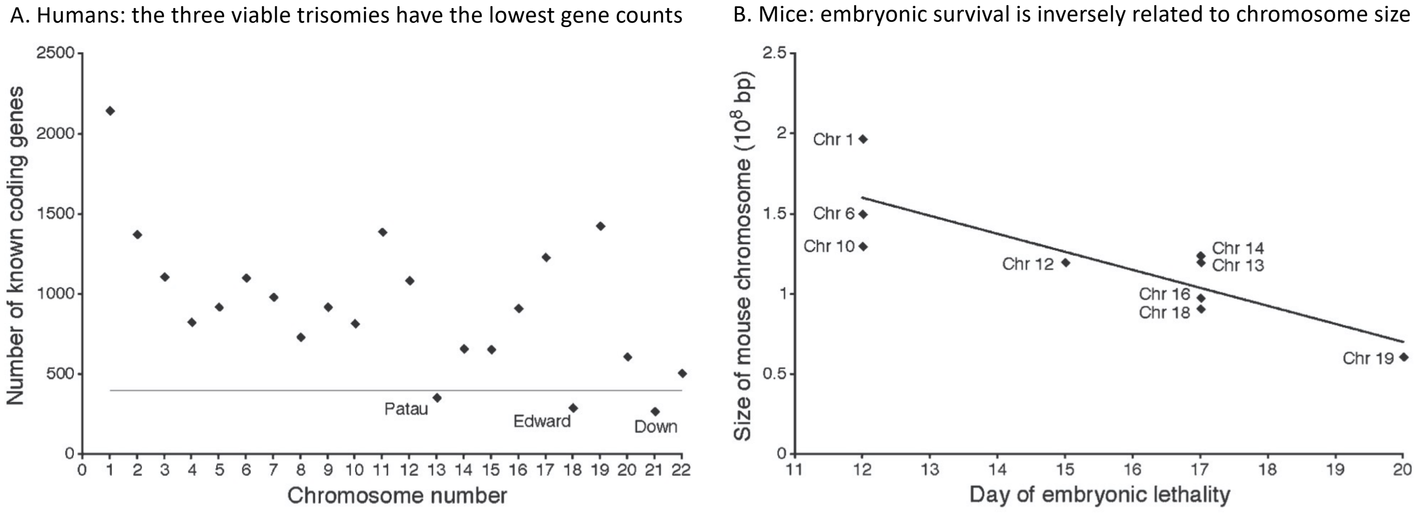
The most frequent types of aneuploidy in humans that survive to birth are listed below. As we'll discuss shortly, these include the three chromosomes with the fewest genes, and unusual combinations of the X and Y chromosomes.

Syndrome	Frequency	Notes
Trisomy 13 (Patau Syndrome)	1/5,000	Severe developmental issues; 5-10% survival at first year
Trisomy 18 (Edwards' Syndrome)	1/5,000	Severe developmental issues; 5-10% survival at first year
Trisomy 21 (Down Syndrome)	1/1,000	Mild to moderate intellectual disability, low fertility
X (Turner Syndrome)	1/2,000	Female, extensive developmental issues, infertile
XXX (Triple X Syndrome)	1/1,000	Female, frequent physical/learning issues, often fertile
XXY (Klinefelter Syndrome)	1/500	Male, may have some feminized features, low fertility
XYY	1/1,000	Male, symptoms usually absent, normal fertility

Table 1.4: *The main types of aneuploidy that can survive to birth. Frequency estimates reflect rates among live births; birth rates for some of these conditions are declining due to prenatal screening.*

Why does aneuploidy affect development? An individual with a monosomy or trisomy can still make the full complement of proteins. But as we discussed above, cells depend on having a precise balance of all their proteins. Individuals with an extra (or a missing) chromosome produce either too much, or too little of all the proteins on that chromosome, and these imbalances lead to major developmental problems. By and large, these are not due to the impact of a few genes that are particularly sensitive to copy number, but instead the accumulated effects of hundreds to thousands of genes with one extra copy each.

One indication of this is that the severity of trisomies reflects the number of genes on each chromosome: it's no coincidence that Trisomy 21 (Down Syndrome) is the least severe trisomy, and that 18, and 13 are next in line, as these are the three chromosomes with the fewest genes (among the autosomes). A similar result can be seen in mice, where there is a tight inverse correlation between chromosome length and how long the corresponding trisomies can survive [67]:


A. Humans: the three viable trisomies have the lowest gene counts

B. Mice: embryonic survival is inversely related to chromosome size

Figure 1.42: Trisomies of chromosomes with fewer genes are more viable. A. *The plot shows gene number for each autosome (y-axis), ordered by chromosome number* 1–22. *The three autosomes with viable trisomies are those with the fewest genes.* **B.** *In a study of trisomies of different mouse chromosomes, there was a strong relationship between chromosome size and how long the mice could survive.* *Credit: Modified from Figure 2 of Eduardo Torres et al (2008) [Link] Used with permission.*

The situation for the sex chromosomes is a little different, but reflects the same principles.

Remember that usually females have 2 X chromosomes, and males have an X and a Y. Ordinarily in females one X chromosome in each cell is silenced by a process known as **X-inactivation**: in other words, only one chromosome is used for gene expression. This is important so that there is no major mismatch between the gene expression levels of X chromosome genes in males and females. This same mechanism rescues people who have three or more X chromosomes, because X-inactivation ensures that there is only one active X, regardless of the actual number of Xs.

Meanwhile, there are only about 50 protein-coding genes on the Y chromosome, and many of these are involved in development of the male reproductive system and of sperm, so individuals with an extra Y chromosome (XYY) are generally healthy.

The fact that individuals with unusual X/Y karyotypes often do have some degree of symptoms is because around 100 genes escape X inactivation [68]. Consequently, individuals with unusual XY karyotypes do not maintain the correct dosages for these genes: this can lead to developmental and health issues, especially for Turner's Syndrome (X-) patients, as outlined in Table 1.4 [69].

Karyotypes evolve rapidly over evolutionary time! Given the strong constraints within the human population in maintaining correct chromosome numbers, it may come as some surprise to learn that closely-related species often evolve quite different karyotypes [70] [j].

For example, humans have 23 pairs of chromosomes, while other great apes have 24 pairs: this is because human chromosome 2 is a fusion of two ancestral ape chromosomes [71].

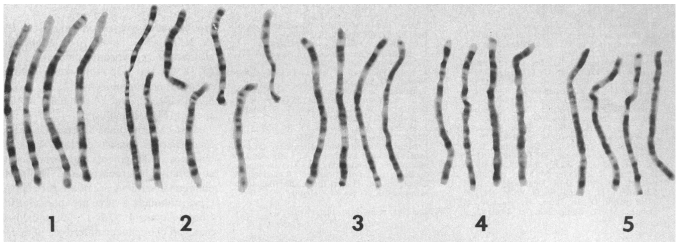


Figure 1.43: **Partial karyotypes of the great apes.** *The image shows human chromosomes 1–5, pictured alongside the corresponding great ape chromosomes. Left to Right: human, chimpanzee, gorilla, orangutan. Human Chromosome 2 is a fusion of two chromosomes that are separate in the other apes.*Credit: From Figure 1 of Jorge Yunis and Om Prakash (1982) [Link]Used with permission.

While the overall structure of the great ape genomes are largely similar aside from this fusion event, our next-nearest relatives, the gibbons, have undergone extraordinarily rapid chromosome evolution. For example, the genome of the northern white-cheeked gibbon has been dramatically reorganized: it has 26 pairs of chromosomes, but even more strikingly it differs from humans by 96 different rearrangements of large chromosomal blocks [72]!

Ordinarily, one might expect major chromosomal rearrangements to have deleterious effects, and these are almost vanishingly rare within human populations [73]. Thus it's surprising that closely-related species can evolve dramatically different karyotypes. One potential explanation is that this may be driven in part by the evolution of new centromeres that can hijack meiosis to gain a selective advantage and spread through populations – but this is an exciting area that is not yet well understood [74].

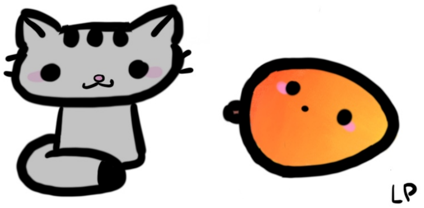


Figure 1.44: *We like to tease our pet cats that they are one chromosome short of a mango! (Cats have 19 pairs of chromosomes – and mangoes have 20.)* Credit: Lucy Pritchard

In summary, the genomes of any two people differ at millions of positions, including SNPs, as well as a variety of more complicated types of sequence differences. In the next chapter we will discuss how these differences can be detected by DNA sequencing.

[j] *For an interesting table of chromosome numbers in different species see: Link.*

Notes and References.

[43] To be more precise, the vast majority of SNPs only have two alleles at any appreciable frequency. However, as we discuss below, virtually every possible allele that is one step away from the reference genome exists somewhere in the world (excluding alleles that would be incompatible with life).

[44] You can imagine that there are pros and cons to each naming system. The *reference allele* is rather arbitrary, because it depends on whether the allele happens to match the individual who was sequenced at that position for the Human Genome (and sometimes that individual had a super rare allele). The *minor allele* label is particularly useful for rare alleles, but it can lead to inconsistent labeling across different samples if the allele frequency is near 0.5. The *derived allele* label is attractive in having a clearer evolutionary interpretation, but it involves an inference about which allele is ancestral that may be uncertain or even incorrect for some SNPs.

[45] For autosomal loci, one generation of random mating (i.e., random with respect to the SNP in question) immediately restores HW proportions regardless of the starting allele frequencies. This means that a process like selection must be implausibly strong to drive meaningful departures from HWE. Note that X-linked loci do not reach HWE immediately (but do converge within a few generations).

[46] Genotyping issues that lead to departures from HWE can occur for various reasons, and the details depend a bit on the specific technology. One common reason for errors is that the sequence surrounding a putative SNP is duplicated elsewhere in the genome and so the sequencing reads or genotyping assay contain a mixture of DNA fragments from two different locations. Suppose that these two duplicated versions of this region differ at exactly one position, and this position has been inferred incorrectly as a SNP. Then all individuals would appear to be heterozygous.

[47] Edwards A. Anecdotal, Historical and Critical Commentaries on Genetics: GH Hardy (1908) and Hardy–Weinberg Equilibrium. Genetics. 2008;179(3):1143.

[48] Genomes of "identical" (monozygous) twins are in fact nearly identical: the genomes of a monozygous pair differ by only $\sim$5 early developmental mutations in non-repetitive sequences, as well as presumably additional STRs and other more-mutable sequences that are more difficult to measure:

Jonsson H, Magnusdottir E, Eggertsson HP, Stefansson OA, Arnadottir GA, Eiriksson O, et al. Differences between germline genomes of monozygotic twins. Nature Genetics. 2021;53(1):27-34

[49] We can generalize the concept of heterozygosity to consider the expected heterozygosity under random mating. The expected heterozygosity is useful if we don't have access to individual-level genomes, and the estimator also has lower variance. For example, if we know the allele frequency p_s at every SNP s in a region of size L, then we can compute the expected heterozygosity as

$$\frac{1}{L}\sum_s 2p_s(1-p_s).$$

(Note that in practice the formula above is slightly biased since we only have estimates of p_s rather than true values; an unbiased formula can be derived by computing the heterozygosity summed over all pairwise comparisons.)

[50] 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526(7571):68

[51] Large sequencing studies continue to find many more novel, rare SNPs: for example the gnomAD Project identified 230M high confidence variants – nearly one every 10 bp – by sequencing about 16,000 genomes. Note that the gnomAD Project had higher sequencing depth than 1000 Genomes, and this accounts for why they detected more new variants per individual. gnomAD Project:

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-43

[52] We'll return to questions about divergence among the great apes in Chapter 2.2.

Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, et al. Insights into hominid evolution from the gorilla genome sequence. Nature. 2012;483(7388):169-75

[53] This was laborious work that relied on PCR amplifying regions of interest, followed by Sanger sequencing. Anna Di Rienzo's lab, at the University of Chicago, also did important work in this area at around the same time.

Frisse L, Hudson R, Bartoszewicz A, Wall J, Donfack J, Di Rienzo A. Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. The American Journal of Human Genetics. 2001;69(4):831-43

Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L, Nickerson DA. Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. nature Genetics. 2003;33(4):518-21

[54] Bhangale TR, Rieder MJ, Livingston RJ, Nickerson DA. Comprehensive identification and characterization of diallelic insertion–deletion polymorphisms in 330 human candidate genes. Human Molecular Genetics. 2005;14(1):59-69;

Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, Mu XJ, et al. The origin, evolution, and functional impact of short insertion–deletion variants identified in 179 human genomes. Genome Research. 2013;23(5):749-61

[55] VNTRs are also sometimes known as minisatellites, while STRs are also microsatellites.

[56] Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, et al. Global diversity, population stratification, and selection of human copy-number variation. Science. 2015;349(6253):aab3761

[57] Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J, et al. A common inversion under selection in Europeans. Nature Genetics. 2005;37(2):129-37

Salm MP, Horswell SD, Hutchison CE, Speedy HE, Yang X, Liang L, et al. The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism. Genome Research. 2012;22(6):1144-53.

One effect of inversions is that they disrupt local recombination in heterozygotes. In some species this enables the evolution of co-adapted gene clusters, but there are no clear examples in humans: Inversion coadapted complexes

Wellenreuther M, Bernatchez L. Eco-evolutionary genomics of chromosomal inversions. Trends in Ecology & Evolution. 2018;33(6):427-40.

[58] The main exceptions where a synonymous variant has a phenotypic effect are usually due to some regulatory function that overlaps with the same positions – for example that the variant is contained with a transcription factor binding site or exonic splicing enhancer.

[59] For a good account of the genetic testing, with quite a bit of historical and forensic context see

Coble MD, Loreille OM, Wadhams MJ, Edson SM, Maynard K, Meyer CE, et al. Mystery solved: the identification of the two missing Romanov children using DNA analysis. PloS One. 2009;4(3):e4838.

[60] Rogaev EI, Grigorenko AP, Faskhutdinova G, Kittler EL, Moliaka YK. Genotype analysis identifies the cause of the "royal disease". Science. 2009;326(5954):817-7

[61] Mukamel RE, Handsaker RE, Sherman MA, Barton AR, Zheng Y, McCarroll SA, et al. Protein-coding repeat polymorphisms strongly shape diverse human phenotypes. Science. 2021;373(6562):1499-505

[62] The most relevant studies test for a depletion of LOF mutations compared with a neutral background. If this is detected it implies that there is at least some degree of selection against heterozygous LOFs. The effects of haploid gene deletions should be roughly functionally similar to haploid LOFs.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285-91

Agarwal I, Fuller ZL, Myers S, Przeworski M. Relating pathogenic loss-of function mutations in humans to their evolutionary fitness costs. bioRxiv. 2022

[63] Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, et al. Diet and the evolution of human amylase gene copy number variation. Nature Genetics. 2007;39(10):1256-60 CITE NOVEMBRE TOO

[64] While the main form of variation at Amylase1 is variation in copy number, it turns out that there is also additional complex structure within the region, as the gene copies appear in several slightly different forms that are variable across individuals:

Usher CL, Handsaker RE, Esko T, Tuke MA, Weedon MN, Hastie AR, et al. Structural forms of the human amylase locus and their relationships to SNPs, haplotypes and obesity. Nature Genetics. 2015;47(8):921-5

[65] Pajic P, Pavlidis P, Dean K, Neznanova L, Romano RA, Garneau D, et al. Independent amylase gene copy number bursts correlate with dietary preferences in mammals. Elife. 2019;8:e44628

[66] It's interesting to note that polyploidy (usually 3 or 4 copies of *all* chromosomes) can be less deleterious than aneuploidy of a single chromosome. Many species, across the tree of life, have evolved polyploid genomes, and it's believed that our own ancestors went through two rounds of whole genome doubling in early tetrapod evolution. Moreover, some human tissues, including liver, placenta, and heart are polyploid. This indicates that problem with aneuploidy is that changes the relative proportions of genes (stoichiometry) relative to one another, not the absolute changes in expression of specific genes.

[67] Torres EM, Williams BR, Amon A. Aneuploidy: cells losing their balance. Genetics. 2008;179(2):737-46

[68] These mainly fall into three categories: (1) There is a pair of *pseudo-autosomal regions*, containing a total of 3 Mb of DNA and 20 genes, that are shared between the X and Y chromosomes and are important for proper chromosomal pairing during meiosis and mitosis; (2) Secondly, there are about 25 genes with essential roles in gene and protein regulation, that have homologs on the X and Y chromosome. These genes have evolved to escape X-inactivation because both XX and XY individuals have two functional copies; (3) genes that are not particularly dosage-sensitive. For estimates of the number of genes that escape X inactivation see Balaton 2015 [Link]

[69] For a more detailed discussion of this see

Posynick BJ, Brown CJ. Escape from X-chromosome inactivation: an evolutionary perspective. Frontiers in Cell and Developmental Biology. 2019;7:241

For analysis of X-Y homologs and their functions see:

Bellott DW, Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Cho TJ, et al. Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. Nature. 2014;508(7497):494-9

[70] Ferguson-Smith MA, Trifonov V. Mammalian karyotype evolution. Nature Reviews Genetics. 2007;8(12):950-62

[71] Yunis JJ, Prakash O. The origin of man: a chromosomal pictorial legacy. Science. 1982;215(4539):1525-30

Ventura M, Catacchio CR, Sajjadian S, Vives L, Sudmant PH, Marques-Bonet T, et al. The evolution of African great ape subtelomeric heterochromatin and the fusion of human chromosome 2. Genome Research. 2012;22(6):1036-49

[72] Carbone L, Harris RA, Vessere GM, Mootnick AR, Humphray S, Rogers J, et al. Evolutionary breakpoints in the gibbon suggest association between cytosine methylation and karyotype evolution. PLoS Genetics. 2009;5(6):e1000538

Carbone L, Alan Harris R, Gnerre S, Veeramah KR, Lorente-Galdos B, Huddleston J, et al. Gibbon genome and the fast karyotype evolution of small apes. Nature. 2014;513(7517):195-201

[73] There are rare examples of balanced translocations that are inherited within families, but I'm not aware of any chromosomes fusions or fissions.

[74] Chmátal L, Gabriel SI, Mitsainas GP, Martínez-Vargas J, Ventura J, Searle JB, et al. Centromere strength provides the cell biological basis for meiotic drive and karyotype evolution in mice. Current Biology. 2014;24(19):2295-300