1.2 A genome owner's starter pack.

A whirlwind introduction to some essentials of the human genome. We emphasize the core function of the human genome as a physical device for storing and replicating biological information.

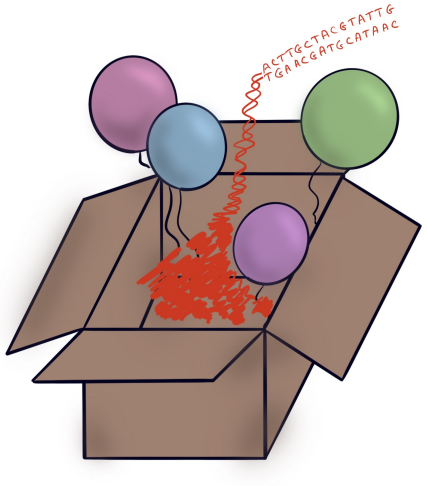


Figure 1.1: *Congratulations on your brand-new human genome! Your custom-made genome has been synthesized for your exclusive use following 4 billion years of evolution.* Lucy Pritchard

A short history. Genetic concepts, and genetic data, are so ubiquitous in modern society that it's easy to forget how recent our understanding of genetics really is.

The origins of modern genetics trace back to the 1850s, when a Moravian monk, Gregor Mendel, used pea plants to learn the most basic rules of genetic inheritance. Mendel's work was published in an obscure journal from Brno (now in the Czech Republic) and was ignored until mainstream scientists rediscovered his manuscript in 1900, thereby kicking off the scientific study of genetics. Thus, genetics had a late start compared to many other scientific fields: for example, cells were first described by Robert Hooke in 1665, and Isaac Newton's theory of universal gravitation was published in 1687.

During the 20th Century, geneticists worked out the basic nuts and bolts of inheritance: phenotypes, mutations, chromosomes and linkage; next, the biochemistry of DNA, RNA and proteins; and ultimately most of the major principles of molecular genetics. Meanwhile, in human genetics, early researchers had learned that certain genetic diseases, like cystic fibrosis or Huntington's disease, are caused by Mendelian mutations in single genes, and by the 1980s they had started to map the genes that are responsible.

At the same time, there was growing realization that most of the ways that humans vary from one another – think of traits like height or weight, diabetes or schizophrenia – are influenced by small contributions from many genes, as well as environmental factors. Last but not least, they had already developed much of the theoretical framework that we use to understand human population genetics today. All of these pieces are central to our story here.

Thus, by the start of the 21st Century, most of the fundamental building blocks of molecular biology and genetics were in place. And yet, at the same time our viewpoint was limited by bottlenecks in measurement: most notably DNA sequencing. The last two decades have seen a revolution in all kinds of biological measurement, but especially of DNA genotyping and sequencing.

Until very recently, it was extravagantly expensive to sequence human genomes: the Human Genome Project completed the DNA sequence for a single genome in 2003, at a cost of $3 billion [2]. With newer, highly efficient technologies, the cost to sequence someone's genome is now well under $1000. These technical advances have ushered in a revolution of human genetics. As of 2022, hundreds of thousands of people have had

their genomes sequenced for research or medical applications. Meanwhile, genome-scale data (SNP genotypes) have been collected for tens of millions of people.

In research settings, high-throughput sequencing is now a universal routine tool. Sequencing has enabled major new insights into the genetic basis of inherited traits, and cancers. Ancient DNA sequencing has revealed important new storylines about the origins and evolution of modern humans. Genome sequencing has also revolutionized our ability to study the *functions* of genomes including our ability to measure which parts of a genome are active in any given cell type. We'll touch on all these topics in later chapters.

The new technologies have also allowed the general public to interface with genetic tools for the first time: millions of people have sent their DNA samples to personal genetics companies that promise insights into customers' family trees, their ancestries, and perhaps even their genetic predispositions. DNA forensics has become an essential part of the criminal justice system, connecting suspects to crime scenes (or exonerating them), and recently using genetic genealogy tools to solve a large number of "cold" cases that had been unsolvable by traditional methods. The use of genetic data in medicine is steadily increasing: millions of mothers have received prenatal genetics screening to provide early detection of chromosomal abnormalities; genome-sequencing is now an important tool in cancer treatment; we are on the cusp of genetic prediction in clinical medicine; techniques like genome-editing with CRISPR and cellular reprogramming promise to transform the role of genetics in medicine. And of course, many aspects of genetic research came together to enable the rapid development and approval of mRNA-vaccines against COVID-19 in 2020.

Genomes, inheritance, and variation. Most of the topics above relate to human genome variation, which is the focus of this book. To understand genetic variation, it's first helpful to think about the genome as a device for storing data, and encoding biological functions. The data stored in your genome (or mine) are inherited from our species' shared ancestors in Africa, via many thousands of generations of mutation, genetic drift, and natural selection. Looking further back into history, your genome is also inherited, albeit with massive modifications, through billions of cell divisions from single-celled ancestors that lived near the beginning of life on earth, some 4 billion years ago.

In the next sections, we look at how DNA stores and encodes biological information [a].

[a] ***Parts of this introductory chapter may be familiar to you already, so feel free to skip over those!***

The DNA molecule. Your genetic data are stored using a molecule called **DNA**, short for deoxyribonucleic acid.

The DNA molecule is shaped like a twisted ladder. Each side of the ladder is called a **strand**, and is made up of four different kinds of chem-

ical building blocks called **bases**: namely **A, C, G, and T** (for adenine, cytosine, thymine and guanine). Along each strand, the bases are linked together by a chemical backbone. A base plus its chunk of backbone is called a **nucleotide**. The distinction between base and nucleotide is not especially important for us here, and we'll use the terms somewhat interchangeably.

The two strands of the DNA molecule fit together with what's called complementary base-pairing: specifically, A on one strand is always matched with T; and C with G. This means that we can have four kinds of rungs: A:T and T:A; C:G and G:C, depending on which base is on which side of the ladder. One rung of the ladder–i.e., 2 bases from opposite strands–is called a **base-pair**.

Another key feature of the strands is that they have a natural direction–analogous to how we always read English from left to right. Each nucleotide is asymmetric, with a so-called 5′ side and a 3′ side (pronounced "5-prime" and "3-prime"). In a DNA strand, all the 5's are oriented in the same direction, so we can label one end of a strand as 5′, and the other end as 3′. Meanwhile, the other strand of the helix is oriented in the opposite direction.

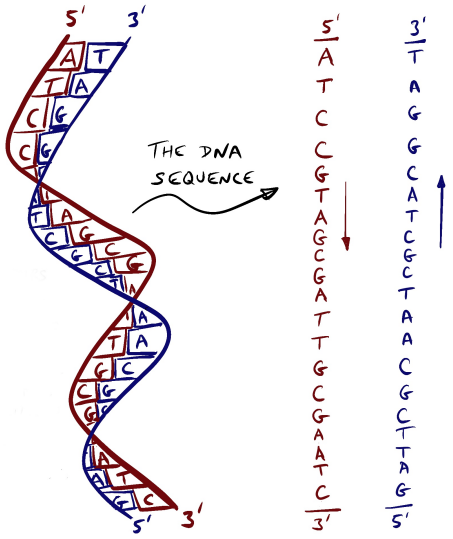


Figure 1.2: *The nucleotides fit together to form complementary strands of DNA, like a twisted ladder. Some of the nucleotides shown in the sequence are obscured at the points where the helix turns perpendicular to the page. Genes can be encoded on either strand, but always in a 5′ → 3′ direction on the relevant strand.*

Again, similar to English which is always written left to right, everything important in genetics happens 5′ to 3′. DNA replication occurs 5′ to 3′. And the copying and decoding of genes – transcription and translation – is from 5′ to 3′. Genes can be encoded on either strand, but since the two strands of a double helix are oriented in opposite directions, genes encoded on one strand are oriented opposite to genes encoded on the other strand.

23 pairs of chromosomes. The DNA in your genome is organized into chromosomes. You have 23 pairs of chromosomes–you got one copy of each chromosome from your mum [b], and one from your dad for a total of 46. That includes 22 regular pairs, creatively named Chromosome 1 to Chromosome 22, roughly in order of size, as well as the sex chromosomes X and Y: biological females have two Xs; males have an X and a Y. The smallest chromosome is actually 21, and not 22, because early studies put them in the wrong order, and the original numbering has been kept.

[b] *In this book, we'll generally use terms relating to sex and parental relationships as a shorthand for referring to biological sex and genetic relationships, while noting that these terms simplify a complicated reality* [3].

Chromosomes 1–22 are referred to as **autosomes** when we want to distinguish them from the sex chromosomes. The two copies of each chromosome pair are referred to as **homologous chromosomes**, or simply **homologs** for short.

Altogether, you have about 3.3 billion base pairs of DNA from each parent: 6.6 billion total in every cell. If we laid out the DNA from a single cell in a straight line, it would be over 2 meters long [c] [4]. Obviously the DNA is not stored in a straight line: instead it's wrapped around small balls of protein called **nucleosomes**, much like spools of thread. The spools themselves are also packaged in an orderly fashion, to fit the whole lot into the cell's nucleus. Together, this highly compact DNA-protein packaging is referred to as **chromatin**, and is the default state for our

[c] *Your body contains ∼40 trillion cells, nearly all of which carry a copy of your genome. If we stretched out all the DNA from all your cells end to end, it would span across the solar system!*

genomes. If you do need to read part of your genome, you'll briefly unspool the bit that you need.

DNA is the world's greatest data storage device. The central role of DNA is to store biological information. Chromosomes are long strings of A, C, G, and T that encode information, sort of the same way a book does, but using 4 letters instead of 26. In a minute, we'll talk about how this information is encoded, but for right now, just pause for a moment to think about the fact that this book of 6 billion base pairs provides complete instructions for all the proteins needed to make each of your cells– and in fact all the proteins you need to assemble a complete person.

As a loose analogy you can think of an organism's genome as being like the computer code (operating system and software) that controls a computer. If you want to make a human or a dog, a worm or a melon, you need to input different strings of DNA. And like computer code, which relies on computer hardware (your phone or laptop) to produce a physical output, a genome relies on elaborate cellular machinery that reads it and interprets it to produce biological functions. Later we'll also talk about the important role of environment in shaping phenotypes.

DNA is an incredibly impressive storage system. If we put this in terms of data storage on a computer, a single human cell carries the equivalent of about 1.5 Gigabytes of computer data [d]! (In computing, a *bit* is a single binary digit that is either zero or one; the basic unit of data storage is a *byte*, which is 8 bits. Since DNA has four possible letters, each base pair is the equivalent of 2 bits, and 4 DNA base pairs carry one byte of data.) So just a few hundred cells carry as much data as your phone – though to be fair, it's the same information repeated in every cell. At the same time, cells are so small that you could actually fit about 100 million cells inside a standard phone.

[d] *Not only is DNA storage physically compact, but it might also seem surprising that the information content of a diploid human genome is also modest compared to modern computing systems, at just 1.5 gigabytes. For comparison, the iPhone operating system is somewhat larger at 2–3 gigabytes, depending on version.*

Indeed, DNA is such an efficient and stable storage system that there is a line of research on how to use DNA to store digital data. DNA is far more compact than computer storage systems, but currently much slower and more expensive for humans to read or write DNA compared to conventional systems [5].

The genomic encoding of biological information. For biological systems, the importance of DNA is that it encodes biological information. One major challenge in genome science is to be able to read the encoded information. What does each of the 3.1 billion base pairs do – if it does anything at all? What would be the impact of a mutation that changes any specific part of the genome sequence?

Soon after the Human Genome Project was completed in 2003, at a cost of 3 billion dollars, one project leader, Eric Lander, famously gave this terse summary of the challenges ahead [6]:

"Genome. Bought the Book. Hard to read."

You can think of the genome as encoding two main types of information: **The first kind of information is contained in genes**. A gene is a stretch of DNA that encodes a protein. (A small fraction of genes encode functional RNAs instead of proteins, but the principles are similar.)

The second kind of information tells each cell how much of each protein to make. This is referred to as **gene regulation**, and is also critically important. For example, the differences between a liver cell, a neuron, or a muscle cell are mainly due to precisely-controlled differences in gene regulation across these cell types.

As we'll see shortly, these two types of information are encoded very differently. Genes encode proteins using a very simple format where each successive block of 3 nucleotides specifies an amino acid.

In contrast, gene regulation is controlled by molecular interactions between DNA sequences and cell-type specific proteins. The language of gene regulation is both highly complex, and highly context-specific: a particular sequence may be interpreted as an important regulatory region in a liver cell, and completely ignored by a neuron. Consequently, while the general principles of gene regulation are fairly well understood, it is still a difficult research problem to predict how a particular DNA sequence will be interpreted in any given cell type. Luckily, it's possible to create accurate maps of regulatory regions using a variety of experimental assays.

Genes and the encoding of proteins. Each gene stores the instructions to make a particular **protein**. If DNA is the information storage device in cells, proteins are the molecules that actually get things done. Much of biology is controlled by different proteins doing different kinds of jobs in cells. (I don't mean to trash-talk the other essential biomolecules, such as lipids – but they are not directly encoded by the genome, and will be a much smaller part of this book's story.)

Even though proteins perform a huge variety of different jobs, they are all made up of the same basic building blocks. These building blocks are small molecules called **amino acids**. Your genome encodes 20 different amino acids, which can be joined together in any order to make a protein. What a protein does is determined by the specific order, and number, of its amino acids. Proteins vary greatly in size, but the average protein in humans is about 400 amino acids long.

Unlike DNA, proteins fold into an enormous diversity of shapes, depending upon their amino acid sequences, and this is part of what determines their biological functions. There is a major field of biology devoted to measuring, and even predicting, 3-dimensional protein folding, and how each protein interacts with other molecules in cells [7].

The genetic code. DNA specifies proteins using a simple code, in which a nucleotide sequence along one strand of the helix encodes a sequence of amino acids.

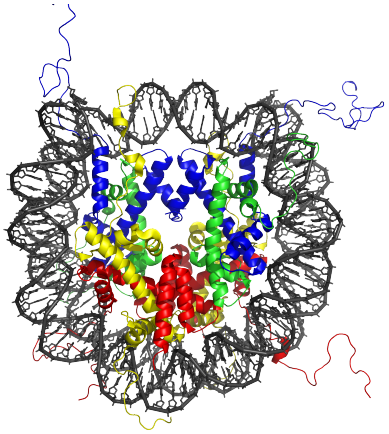


Figure 1.3: **Example of a protein structure.** *Most of your genome (DNA shown here in black) is wrapped around protein complexes called* ***nucleosomes*** *(colors), like thread on spools.* *Credit: Zephyris CC BY-SA 3.0 [Link]*

Remember that DNA is made up of 4 letters: A,C,G,T. So how does DNA encode the 20 amino acids? Just like words in a book, we need more than one letter to encode each amino acid. If we used pairs of adjacent letters (here we mean adjacent on the same strand of the helix), there would be $4^2 = 16$ possibilities: AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT. Hmmm. That still isn't enough to code for 20 amino acids. So we're going to need three adjacent letters for each amino acid, as that gets us to $4^3 = 64$ possibility combinations. So for example, AAA in DNA codes for the amino acid Lysine in the corresponding protein.

Now that gives us 64 combinations when we only really need 20. So what does the cell do with the 44 extra triples? Well, three triples, TAA, TAG, TGA, are STOP signs, marking the end of the protein. And beyond that, there is redundancy, so that most amino acids are encoded by multiple triples: eg TGT and TGC both code for the amino acid Cysteine. The other special signal is ATG, which signals as a START sign when it occurs at the beginning of a protein. ATG also encodes the amino acid Methionine. Each block of three nucleotides is called a **codon**.

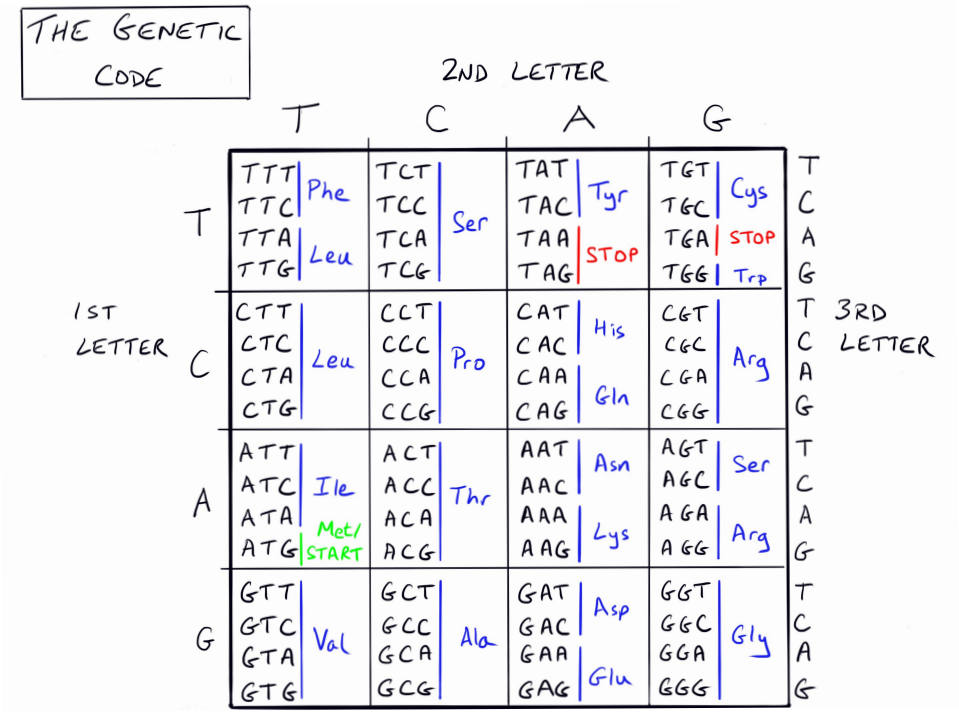


Figure 1.4: *The genetic code: this shows the encoding of DNA triplets for amino acids. The 64 possible DNA codons are shown in black, and the corresponding amino acids are shown in blue using their abbreviations. ATG signals both the protein START and the amino acid Methionine. TAA, TAG, and TGA are protein STOP codes.*

Abbreviations for the amino acids:
Ala: Alanine; Arg: Arginine; Asn: Asparagine; Asp: Aspartic Acid; Cys: Cysteine; Glu: Glutamic Acid; Gln: Glutamine; Gly: Glycine; His: Histidine; Ile: Isoleucine; Leu: Leucine; Lys: Lysine; Met: Methionine; Phe: Phenylalanine; Pro: Proline; Ser: Serine; Thr: Threonine; Trp: Tryptophan; Tyr: Tyrosine; Val: Valine.

This code for translating from DNA to protein is called the **genetic code**. It's interesting that this code is nearly identical in all living things. For example, most bacteria have exactly the same code as humans. There is no fundamental reason why AAA should encode the amino acid Lysine– it just started that way in the first cells to evolve a genetic code, and has been inherited throughout the tree of life ever since, during the last 4 billion years of evolution. Notable exceptions to the "universal" genetic code can be found in the tiny genomes carried by our mitochondria, which encode four of the codons differently [8].

Once we know the genetic code, the encoding from DNA to protein is remarkably simple: it starts with ATG, and then every successive 3-nucleotides encodes a single amino acid until we reach the first STOP. (There's a minor complication, which we'll get to shortly, that blocks of DNA called introns are removed before the protein is decoded.)

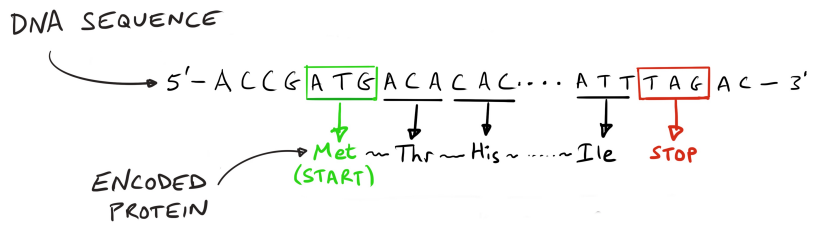


Figure 1.5: **The encoding from DNA to protein.** *The amino acid sequence is interpreted as starting from the first ATG and continuing by threes until the first STOP codon. The DNA sequence shows the coding strand only.*

You can imagine that **this encoding is fragile**, in the sense that just changing a single nucleotide can potentially alter the protein almost completely: for example a mutation that introduces an early stop signal will cause the protein to be immediately terminated; similarly, insertion (or deletion) of a single nucleotide would cause the reading frame of the protein to shift and, from that point on, to encode a completely different amino acid sequence. As we will see in future chapters, both of these types of mutations do occur: they generally cause complete loss-of-function of the affected protein, and depending on the protein, they are often highly deleterious.

DNA → mRNA → Protein. DNA is not interpreted directly into protein, but instead it is first copied into an intermediate called **messenger RNA** (mRNA). RNA is a molecule that is very similar to DNA, but it is usually only one strand of the helix, and is less chemically stable for long-term storage [e]. Note that RNA uses a base called Uracil (U) everywhere that DNA uses Thymine (T). This flow of information from DNA → mRNA → protein is known as the **Central Dogma**.

[e] *Even though RNA is less stable than DNA some viruses, including HIV and the virus that causes COVID-19, use RNA as their main storage molecule instead of DNA.*

Transcription. DNA is stored within the cell's nucleus. In order to make a protein, your cell unwraps the bit of the genome that encodes that gene, and makes mRNA copies of the DNA. This process is known as **transcription**–meaning copying.

Translation. mRNA copies are then transported out of the nucleus into the cell's cytoplasm, where molecular machines called **ribosomes** assemble proteins, using the mRNA sequence as a template. This process converts the biological information from the four-letter alphabets of DNA and RNA into the twenty-amino acid alphabet of proteins. This process is known as **translation**, reflecting the conversion from one type of information (DNA/RNA) into another (protein):

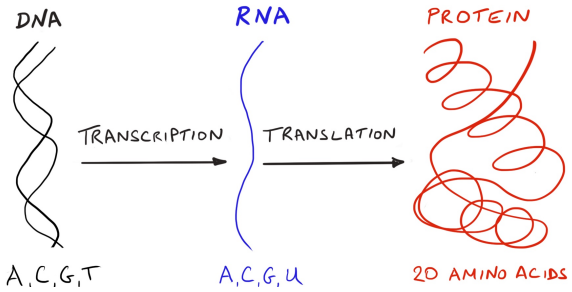


Figure 1.6: **The flow of genetic information.** *DNA provides permanent information storage for cells; mRNA serves mainly as a temporary molecule, used as a template for translation; proteins are highly versatile molecules that perform a wide range of functions. As shown below, the three molecules use different alphabets.*

At this point I should confess that despite the grandiose title of the Central Dogma, a small fraction of genes don't seem to know about this rule, as they produce functional RNAs instead of proteins: for example, some RNA genes encode essential components of the ribosome, and another RNA gene is responsible for inactivating one of the X chromosomes in females [9] [10]. You may be getting the (correct) sense that virtually every rule in biology has exceptions!

Gene structure: UTRs, Exons, Introns, and Splicing. So far, we've been talking about the part of an mRNA that encodes a protein. But this is actually embedded in a much larger transcribed region.

Transcription begins from a location called the **Transcription Start Site**, and terminates at the **Transcription End Site**. The initial immature transcript is referred to as a **pre-mRNA**.

Almost immediately (usually starting during transcription), another key process takes place, in which regions called **introns** are **spliced** (cut) out of the pre-mRNA to produce a shorter, processed mRNA. After being cut out, the introns are trashed, and the nucleotides are recycled. As we'll see shortly, introns are usually much longer than exons, and the final mRNA is usually just a few percent of the initial pre-mRNA [11].

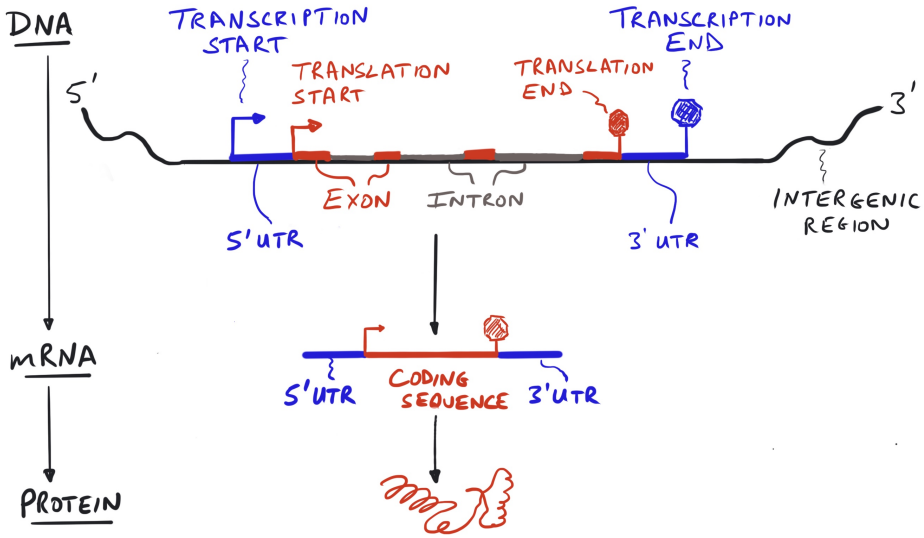


Figure 1.7: **A typical gene structure.** *Transcription initially includes 5′ and 3′ UTRs, coding exons, and introns. The introns are rapidly removed to create the processed mRNA, prior to translation.* ***This is not drawn to scale:*** *typical introns are 10× to 100× larger than exons.*

The final, processed, mRNA is transported from the nucleus into the cytoplasm, where translation takes place. The translation machinery finds the first available start codon (ATG): this is the **Translation Start**. It then

proceeds until it finds the first in-frame STOP signal: the **Translation End**. The regions upstream and downstream of the coding region are known as the **5′** and **3′ Untranslated Regions (UTRs)**. The UTRs often contain information that is used to target the mRNA to particular locations in the cell, or for other forms of regulation.

One advantage of splicing is that it is possible to make different protein products from the same gene, by including or excluding different combinations of exons, or by using different splice sites. This is known as **alternative splicing**, and the different protein products are called **isoforms**. For some genes, distinct isoforms are critical for creating functional diversity of proteins from the same transcripts [12].

Splice site specification. Given that the exons are joined together from a much longer pre-mRNA, this immediately raises another question: how does the splicing machinery know where to cut? What marks the positions of the exon-intron boundaries?

This information is encoded in the DNA (and hence the pre-mRNA, which is what the splicing machinery is actually interacting with) using a variety of signals. First, the splicing code requires a GT at the start, and AG at the end of nearly all human introns (GU and AG in the pre-mRNA). As we'll see in Chapter 1.3, any change in the GT or AG forces splicing to occur at another position – this can dramatically change the encoded protein and can have devastating consequences.

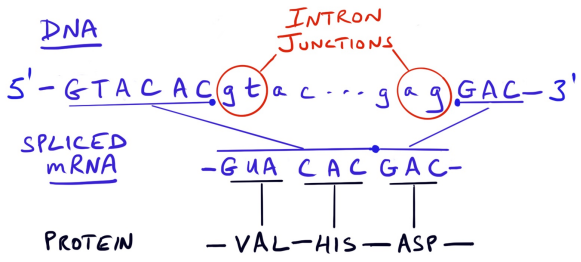


Figure 1.8: **Correct splicing relies on positioning signals encoded in DNA.** *The figure indicates intronic nucleotides with lower-case text. Mutations in the 5′ gt or 3′ ag of the intron disrupt splicing and often result in a nonfunctional protein. T in DNA is U in RNA.*

But of course, there are many GT and AG sites in a typical intron (each of these occurs roughly once every 16 base pairs). So in addition to these required features which help to position the precise splice site, the exon positions are also indicated by a combination of weaker sequence-based signals, where no single nucleotide is fully required for correct splicing: for example the region upstream from the AG usually contains Cs and Ts, as well as an upstream A that is involved in cutting out the intron but can occur at variable distances. These, and other, signals help to position the splice site.

These weaker signals that help position splicing events are very different from the simple and precise algorithmic rules that encode proteins; instead they are more similar to the sequence elements that control gene regulation – which we'll discuss next. The goal of understanding the determinants of splicing is an active research area, using both experiments and machine learning [13].

The encoding of gene regulation. Aside from coding proteins, there's a second kind of information stored in DNA. This information tells a cell which RNAs and proteins to produce, and in what quantities. The production of specific RNAs and proteins is called **gene expression** and the controls of expression are called **gene regulation**. Gene regulation is encoded in the genome, and it turns out that the encoded regulatory information is just as important as the protein-coding sequences themselves.

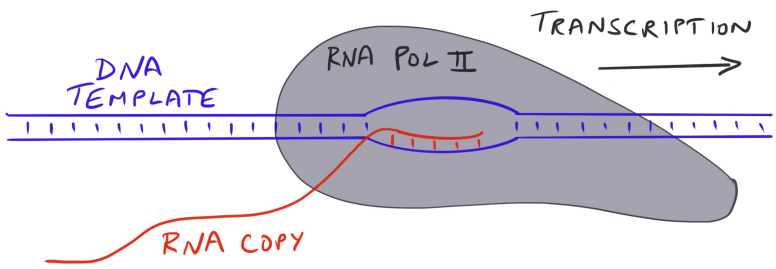
To understand why gene regulation is so important, it's helpful to reflect on the fact that we are immensely complex multicellular organisms. Think about all the different types of cells in a human body: skin cells, heart cells, liver cells, neurons, sperm and eggs, and hundreds of others. These cell types do very different jobs, and look different under a microscope. It turns out that every cell type also expresses a characteristic portfolio of mRNAs and proteins – and this portfolio is a large part of what gives a cell type its identity.

Moreover, cells must produce mRNAs and proteins in very precise proportions (a bit like mixing ingredients for baking). Consistent with this, many genetic diseases are caused by disruptions in the relative proportions of expressed genes [14].

So how is this precise regulatory information encoded in the genome? And, even more strikingly, how does a cell know whether to express the portfolio of genes required for a liver or a neuron or a muscle, even though every cell carries essentially the same genome?

To understand these questions, we first need to detour into some brief details about how gene regulation is encoded in the genome.

The major focus of gene regulation is on controlling transcription. Genes are copied into mRNA – i.e., transcribed – by a protein machine called RNA Polymerase II (**Pol II** to its friends, pronounced "pol-2"). Prior to transcription, Pol II assembles in a region of DNA at the start of the mRNA, known as the **promoter**. The assembly is guided by a set of additional proteins that, along with Pol II form a so-called **Pre-Initiation Complex** within the promoter region. Once Pol II has been assembled at the promoter, it attempts to initiate transcription; if that is successful, Pol II then chugs along the gene, at a speed of ~2 kb/minute [15] to produce an mRNA copy of the DNA [16].



We'll skip over many interesting details about the molecular biology of transcription – but for our story here, the key question is to think about

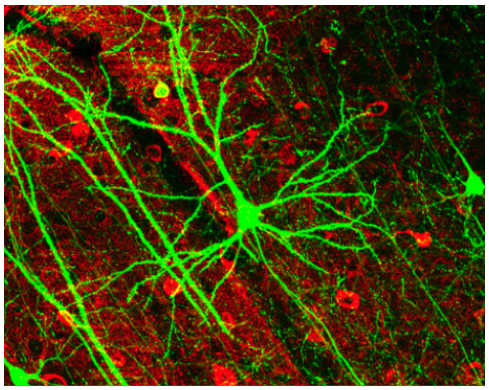


Figure 1.9: **Gene regulation controls the differentiation of cells into many distinct types.** *In this image, two types of neurons in mouse cerebral cortex are stained red and green, depending on whether each cell produces GABA, a key neurotransmitter.* *Credit: Fig. 6F of Wei-Chung Allen Lee et al (2005); [Link] Creative Commons License.*

Figure 1.10: **Transcription:** *RNA Pol II makes an mRNA copy of the DNA. Gene regulatory information is responsible for controlling transcription rates of each gene in specific cell types and conditions.*

the DNA sequences that direct transcription. Crucially, how do DNA sequences position the pre-initiation complex? This determines where transcription will start. And how do DNA sequences control the rate of Pol II assembly and transcription [17]?

These decisions are guided in large part by proteins called **transcription factors (TFs)**. Most TFs have a DNA binding domain that attaches to the genome at specific sequences (**transcription factor binding sites**) [18]. Meanwhile, other parts of the same TF can interact with other proteins to help increase, or sometimes to repress, transcription. As an example, the image below shows the molecular structure of a TF called AP-1, where the purple region of the protein is bound to DNA:

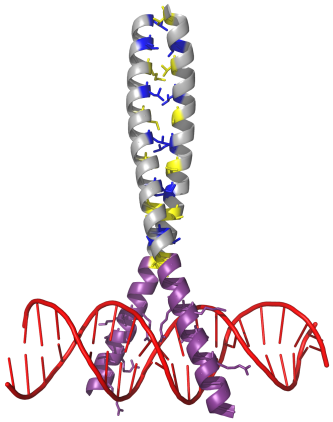


Figure 1.11: **Transcription factor binding to DNA.** *Most TFs have a DNA-binding domain (shown here in purple); other parts of the protein structure can interact with other proteins to control transcription.* *Credit: Houq [Link] CC-BY-SA-3.0*

TF binding usually takes place both within the promoter region itself, as well as at more distant locations called **enhancers**. Enhancers are regions of TF binding that are situated outside the promoter. When TF binding occurs at the enhancers, the DNA can form a loop to bring the enhancer into close physical contact with the promoter. These enhancer-promoter interactions can be essential for assembling the Pre-Initiation Complex which includes Pol II, prior to transcription:

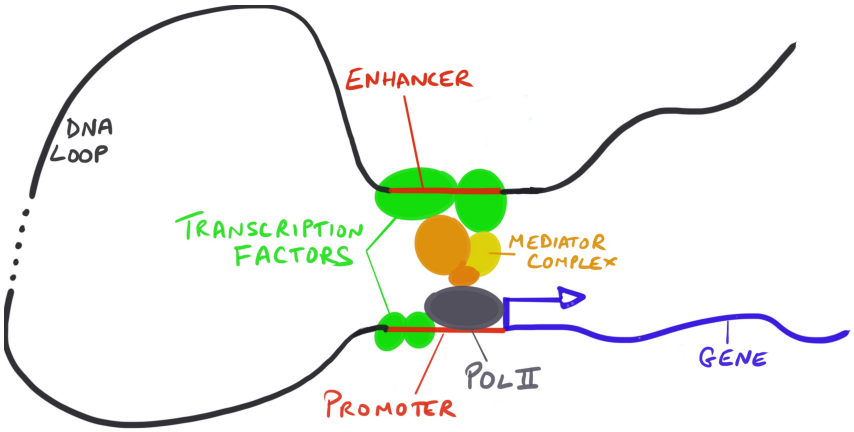


Figure 1.12: **Enhancer-Promoter interactions help drive gene expression.** *Pol II assembles in the core promoter prior to initiating transcription. It is stabilized by protein-protein interactions with TFs bound both within the promoter and at distant enhancers. Promoter-enhancer proteins may attach through protein bridges such as the Mediator Complex.*

Enhancers are often located quite far in DNA distance from the promoters they regulate – usually at distances of tens of thousands up to a million base pairs away, but loop around to create physical proximity.

So what specifies the locations of transcription binding sites, and by extension, of promoters and enhancers? At positions where the protein contacts the DNA directly, each TF has preferred binding sequences that reflect physical interactions in the contact zone between amino acids in the TF, and the nucleotides in the DNA. These preferred sequences are called **binding motifs** – for example, the image below shows the preferred binding sequence, TTTGCAT, for the TF Oct1:

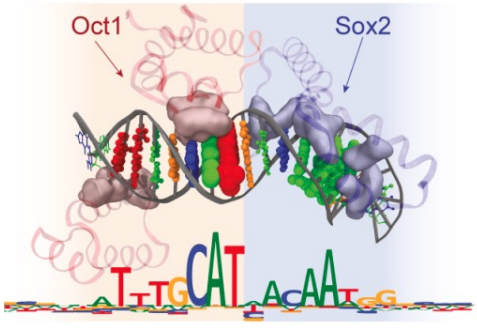


Figure 1.13: **Gene expression is largely controlled by transcription factor (TF) binding of DNA.** *The image shows binding of the Oct1 and Sox2 proteins to DNA (these two factors often bind jointly). The letters below the plot provide a graphical representation of which nucleotides at each position are preferred for binding: larger letters are more important.* *Credit: Fig 3a from Žiga Avsec et al (2020) [Link] CC-BY-NC 4.0.*

However, these binding motifs are neither necessary nor sufficient to predict binding. First, even within TF binding motifs, most nucleotides are not strictly required for binding. In the image above, the sizes of the letters indicate the importance of each position for binding: the largest letters, CAT, are found within most Oct1 binding sites, but the other positions are more variable. Second, since these binding motifs are quite short, they are found found many times in the genome. Most TFs bind only a tiny fraction of all the possible motif matches.

Instead, the specificity of TFs to bind in the correct locations is usually controlled by combinations of factors binding adjacent DNA sequence elements: for example, very often binding is stabilized when ensembles of multiple TFs can bind in a small region [19]. The specific rules that control TF binding are highly complex and vary across cell types; development of computational tools for predicting TF binding sites is an important research area where machine learning techniques have started to make huge progress from around 2015 onward [20].

A related puzzle is that enhancers act by DNA *looping* to create physical interactions with promoters. How do enhancers "decide" where to loop to? While there is a tendency for enhancers to interact with the nearest promoter(s), there are exceptions in which enhancers ignore nearby genes in favor of regulating genes as far as a million base pairs away [21]. The controls of looping are poorly understood at present [22].

Cell type differences in regulation. Lastly, I want to touch briefly on a remarkable feature of genomes. All of your many cell types carry essentially the same genome, and yet they can interpret it differently to produce different portfolios of genes, and these give different cell types their unique identities: for example T cells, or liver cells, or neurons.

The regulatory logic that I've described above starts to hint at how this is possible. Cell type identities are controlled in large part by which en-

hancers are active (and hence which genes are expressed). Enhancer activity, in turn, is controlled by combinations of transcription factors binding. The key point here is that different cell types express different sets of TFs; thereby turning on (or off) different enhancers across the genome.

But how do cells "know" which TFs they should express? In embryonic development, the earliest cells can produce any possible cell type, but as the organisms develops, cells become increasingly specialized. This specialization is controlled in large part by turning off embryonic TFs, and turning on other TFs that are specific to particular cell lineages. The lineage-specific TFs drive programs of gene expression that are appropriate to the corresponding cell types.

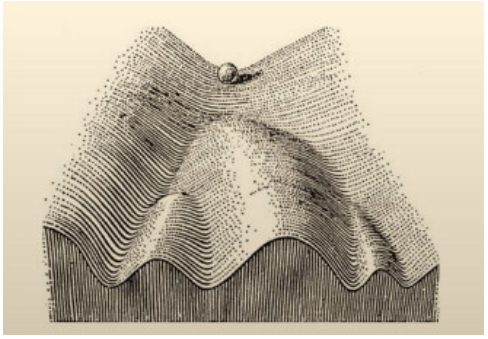


Figure 1.14: **Waddington's landscape metaphor for cellular differentiation (1957).** *Conrad Waddington imagined the increasing specialization of cell types during development as like a ball rolling down a slope. As it rolls it makes random choices that restrict it to increasingly narrow gullies; in cellular terms, we can think of this as turning on lineage-specific transcription factors that drive cell-type relevant programs of gene expression.*

In summary, the encoding of gene regulation works on very different principles than the encoding of proteins. First, gene regulation is analog (i.e., expression level is a continuous variable), unlike protein coding, which is digital (each codon sequence encodes exactly one protein). Secondly, the encoding of expression is controlled by the aggregate effects of many nucleotides and is robust in the sense that single nucleotide changes in the sequence generally have small effects on expression; in contrast, single nucleotide changes such as premature stop codons can completely break protein function.

In the last few pages we have discussed that two major categories of information stored in genomes include the encoding of genes (i.e., mainly proteins); and the encoding of regulatory information (when and how much to make each protein). *How is this information organized in our genomes?*

Bloated genomes: the good, the bad, and the ugly. Remarkably, only about 1% of the genome encodes proteins. A somewhat larger amount codes for regulatory sequences – perhaps around 10% – although the precise amount is uncertain due to the cryptic nature of gene regulatory elements [23]. But most of the remaining ∼90% of the genome sequence shows no clear evidence for biological function. What's there?

To start addressing this question, it'll be useful to have a rough sense of the landscape of genomes and functional elements.

Measurement units of DNA sequence. We'll often need to measure lengths of DNA. The natural units of sequence length are in terms of base pairs but it's convenient to abbreviate different scales with different units (similar to how we use milligrams, grams, and kilograms). So you'll want to remember that:

- 1 bp = 1 basepair
- 1 Kb = 1 kilobase = 10^3 bp
- 1 Mb = 1 megabase = 10^6 bp
- 1 Gb = 1 gigabase = 10^9 bp

Chromosome sizes. The human genome is about 3,100 Mb = 3.1 Gb. Most cells have two copies of the genome (one from mum and one from dad), so that's a total of 6.2 Gb. The chromosomes range in size from 250

Mb (Chromosome 1) down to 47 Mb (Chromosome 21). The mitochondrion has its own genome, contained in a small circular molecule of 16 Kb. For comparison, the genome of SARS-Cov2, the virus that causes COVID-19, is about 29.9 Kb.

Gene numbers and sizes. Meanwhile, the genome contains about 20,000 protein-coding genes [f] (estimates range from about 18,000–22,000 depending on how strict the criteria are that each gene is translated and/or functional). This works out, on average, to about 6.5 genes per Mb, although the distribution of genes is highly uneven.

To give you a sense of scale, the median length of a protein-coding gene, including introns, is 27 Kb [24]. Meanwhile, the coding length is much shorter, with a median length of 1.2 Kb (400 amino acids). A typical gene has 8 exons, and the median size of a coding exon is 122 bp. Introns are more than ten-fold longer, with a median size of 1,600 bp, and a mean of 6,300 bp. Coding exons and UTRs occupy only about 2.5%, each, of the average pre-mRNA before splicing.

You can see an example of a genome region, in a screenshot from the Ensemble Genome Browser. Known protein-coding genes are marked in orange. The vertical bars and boxes are exons or UTRs; horizontal lines are introns. Other possible genes are marked in grey and purple (in practice most of these are likely nonfunctional) [25]. This region is fairly typical, except that the gene density is about twice the genome average.

[f] *Before the human genome was completed, most genome scientists expected that there would be many more genes. In the year 2000, the British scientist Ewan Birney organized a betting pool to guess how many genes there would be. The mean guess was over 60,000; the winner was Lee Rowen who had the lowest guess (24,800) out of more than 460 bets [Link].*

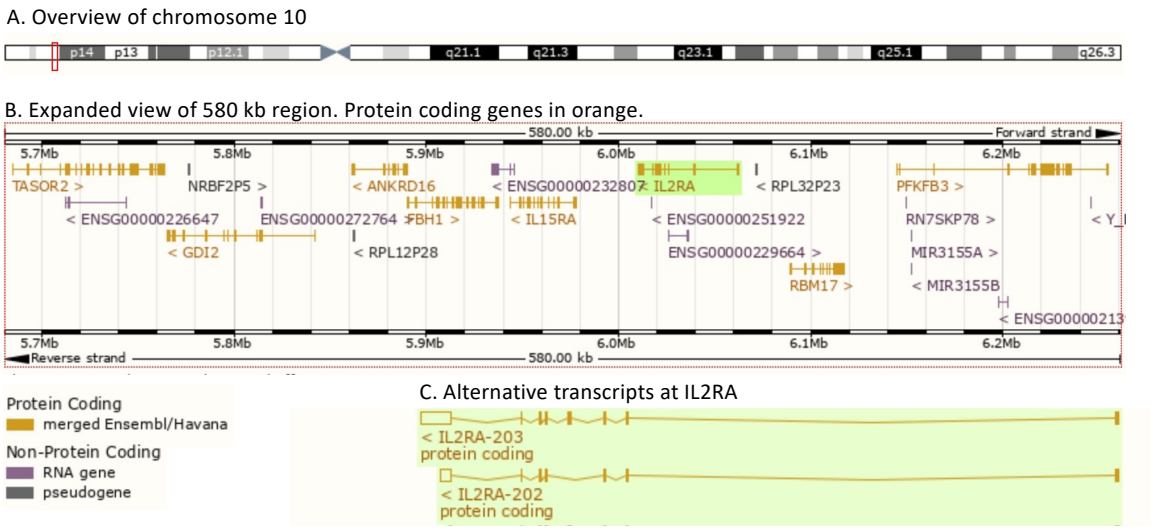


Figure 1.15: Genome browser view of the genome. *Screenshots from the Ensembl Genome Browser show a gene-dense region around IL2RA (an important immune gene).* **A.** *The IL2RA region is marked by the red box at the left end of the chromosome.* **B.** *Coding genes are marked in orange. '>' or '<' mark the direction of transcription of each gene (i.e., whether it is coded on the forward or reverse DNA strand).* **C.** *Expanded view of two possible transcripts at IL2RA. Coding exons and UTRs are marked by filled/open boxes respectively.*

Source: Ensembl browser [Region][Transcripts]

So if only about 1% of the genome is protein coding, and a small fraction of the rest (~10%) encodes regulatory information, then what is all the rest of the genome doing?

Remarkably, most of the genome doesn't have any clear function.

Indeed about two thirds of the genome is made up by **repetitive DNA**: short sequences of hundreds to thousands of base pairs that are repeated many times in the genome. A few of these repetitive elements are involved in gene regulation, but most don't do anything useful for you. In fact they are sometimes referred to derisively as "junk DNA". (These elements really need to hire a better PR team.)

The single most common repetitive element is a 300 base pair sequence called an Alu element, which occurs about 1 million times in the genome! In other words, about 10% of the storage space of the human genome is given up to recording Alu elements – this about 10 times as much space as we devote to storing all genes.

Alu is a type of **transposable element (TE)**. TEs are DNA elements that can copy themselves and reinsert elsewhere in the genome. They are usually considered **selfish DNA**, meaning that they proliferate due to their ability to replicate, while having little or no value to the host genome – in short, they are genome parasites. In the case of Alu, it first infected the genome of our ancestors about 65 million years ago, and has been wildly successful in spreading itself around the genome since then. In some cases Alus and other repetitive elements have evolved new functions, but most are essentially inert elements. These must be copied every time a cell divides, but most do not contribute to genome function.

It is outside our scope but there is fascinating work on mechanisms that have evolved to prevent transposable elements from spreading in the genome, and the ways that TEs evolve to evade those mechanisms [26]. At the same time, there is great work on TEs that have been "domesticated" by host genomes to serve as protein domains or regulatory elements [27].

Genome sizes and TEs. The genome sizes of different organisms vary enormously, as you can see in the table below. Notice the switch from measuring genomes in megabases (Mb) to gigabases (Gb) partway through the table. There's about a 10,000-fold difference in genome size between E coli and Axolotol, even though the numbers of genes varies by less than a factor of 10.

Organism	Genome Size	Number of genes
E. coli (bacterium)	5 Mb	4,000
S. cerevisiae (yeast)	12 Mb	6,000
C. elegans (nematode)	101 Mb	20,000
A. thaliana (flowering plant)	135 Mb	27,000
D. melanogaster (fly)	175 Mb	15,000
human	**3.1 Gb**	**20,000**
Picea abies (spruce tree)	20 Gb	28,000
Axolotl (salamander)	32 Gb	23,000

Table 1.1: Haploid genome sizes for representative organisms. *Notice the enormous range of genome size (by a factor of $\sim 10^4$), while gene numbers vary by less than a factor of* 10*. The largest genomes are cluttered with repetitive DNA. Gene numbers are approximate and are for protein-coding genes.*

Naively, one might perhaps have expected that the genome-sizes of or-

ganisms would reflect their complexity. While it's true that single-celled organisms generally have smaller genomes and fewer genes than multi-celled organisms, there is no very clear pattern of genome-sizes beyond that. (It may not be entirely clear how to measure organismal complexity but, for many of us, it might hurt our feelings to be told that axolotls are ten-fold more complex than we are.) But actually a major determinant of genome-size is how active TEs have been in each evolutionary lineage.

The inheritance of genomes. So far we've been talking about genomes as a device for storing biological information. The next crucial feature is that genomes can be copied to make new cells and new individuals.

In animals, there are two main forms of copying: **mitosis**, in which the genome of one cell is essentially copied to produce two identical genomes; and **meiosis** in which a diploid genome (two of each chromosome) is reduced to haploid (one of each chromosome) to create gametes prior to **fertilization**.

Genome copying: mitosis. Your body started from a single fertilized egg cell. Now that you're reading this, your body contains some 40 trillion cells, each with nearly-identical copies of those original 46 chromosomes. For organisms to increase their number of cells – and to grow in size – the cells need to go through **cell division**.

In cell division, a "parent" cell divides into two "daughter" cells. The parent cell copies each of its 46 chromosomes; then as the cell splits into two, each daughter cell inherits 46 chromosomes to match the genome of the parent. This process of first copying, and then correctly distributing the chromosomes into the daughter cells, is called **mitosis** (pronounced "my-toe-sis") [28].

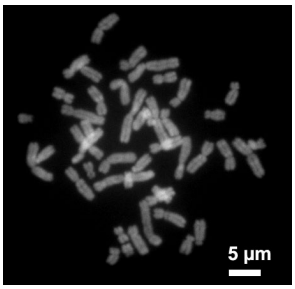


Figure 1.16: **Human chromosomes**, *condensed during mitosis.* *Credit: Steffen Dietzel. CC BY-SA 3.0 [Link]*

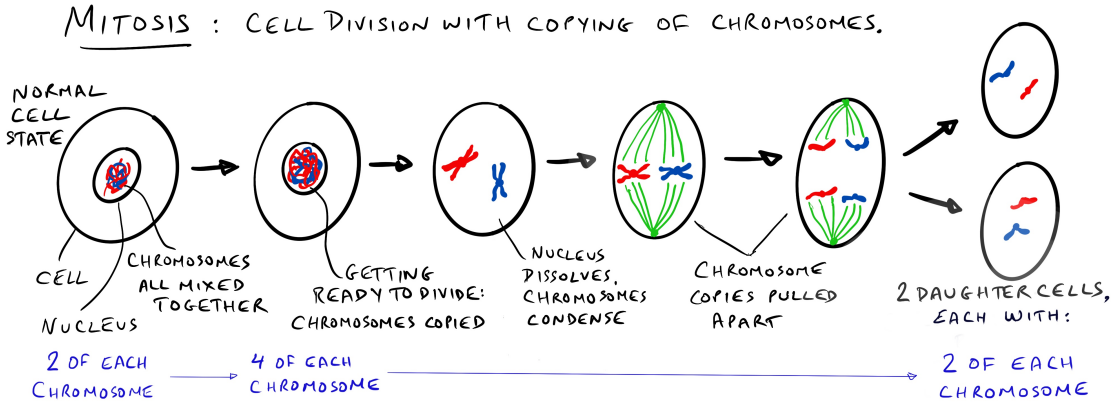


Figure 1.17: Mitosis. *For simplicity, we just show one chromosome; red and blue indicate the two versions of that chromosome carried by each cell (e.g., the chromosome that came from mum in red, and from dad in blue). The x-shaped structures in the middle of the plot show that both the red and the blue versions have been made into pairs of identical copies; one red and blue copy is distributed to each daughter cell.*

Genome reduction and shuffling: Meiosis. In contrast, we need a very different type of cell division to make **gametes** (i.e., sperm and eggs).

While ordinary cells carry 2×23 chromosomes, the gametes only carry 1×23. This is so that when sperm and egg fuse, the fertilized egg have pairs of each chromosome like a regular cell: i.e., 2×23. The process of halving the chromosome numbers to make gametes is called **meiosis** (pronounced "my-oh-sis").

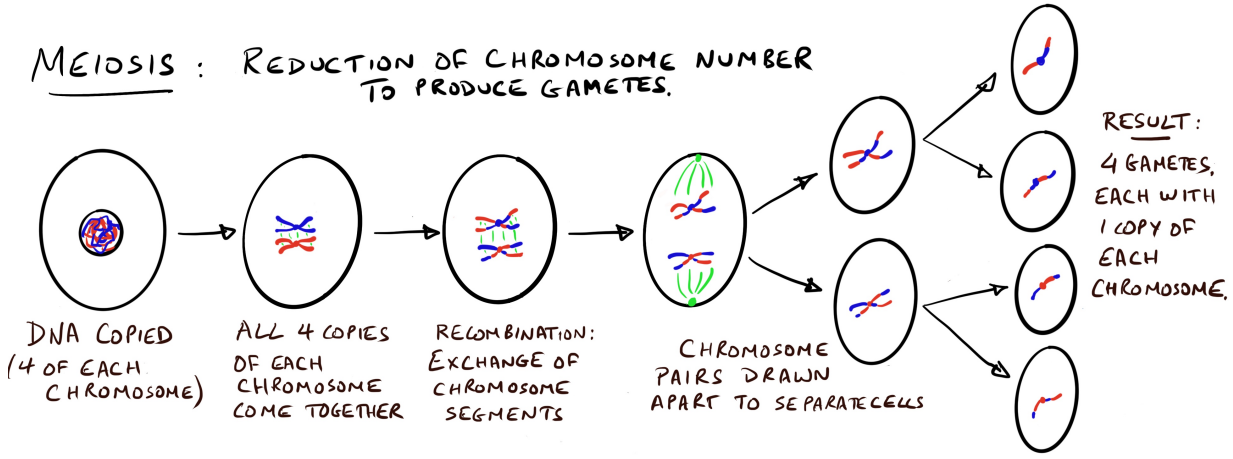


Figure 1.18: Meiosis. *Meiosis starts with DNA copying to make four copies of each chromosome. Next, these come together to exchange pieces: recombination. Then, two stages of cell division result in four gametes, each with one copy of each chromosome. As above, we show one chromosome: red is the version of that chromosome inherited from mum, and blue the version from dad. In females, only one of the four resulting cells develops into an egg.*

Like mitosis, meiosis begins by doubling the amount of DNA in the cells, so that there are 4 copies of every chromosome. It then goes through two rounds of cell division to result in four gametes, each with 1 copy of each chromosome.

Meiosis includes **a crucial process called recombination**, or **crossing-over** [29], which shuffles segments of chromosomes between the maternal and paternal copies. In the figure you can see that the red and blue chromosomes–originally red came from mum and blue from dad–have been shuffled to result in new combinations in each of the 4 gametes.

Meiosis assigns a random 50% of the genome into each gamete. Meiosis is a fundamentally random process that produces a different outcome every time. This is in sharp contrast to mitosis, which is fully predictable: i.e., mitosis produces highly accurate copies of the parent cell every time.

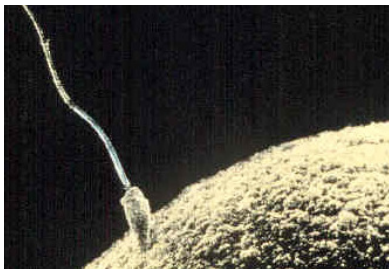


Figure 1.19: **Sperm and egg fusing.** *Unknown author, Public Domain [Link].*

There are two stages of randomness in meiosis: **first, recombination produces a random shuffling of chromosomes**. Later in the book we'll come back to the importance of recombination. **Second, the recombined chromosomes are assigned randomly to gametes**. This combination of two levels of randomness means that every sperm or egg that you produce across your lifetime carries a random, and different, 50% of your genome.

Fertilization. Meiosis is used to create sperm in males, and eggs in females. Each of the sperm and eggs now has a total of 23 chromosomes. Fertilization occurs when a single sperm cell inserts its 23 chromosomes

into the egg to create a fertilized egg that is back to the normal chromosome number of 2×23.

Mutation. The ultimate source of all genetic variation. We'll discuss the types of mutations, mechanisms, and abundance of mutations. Genome replication is extraordinarily accurate, and a typical child carries only about 70 new single nucleotide mutations genomewide. This works out to an average human mutation rate of about 1.2×10^{-8} per base pair per generation. Mutation rates are higher in males than in females, such that a typical child inherits about 3/4 of their new mutations from their dad.

Major data resources for human genetics. In the last part of this chapter we turn our attention to a short description of some of the key resources that are widely used in human genetics. Some combination of these resources were used in virtually all of the modern studies described in this book.

Figure 1.20: **A production line for automated sample preparation**, *built at the Whitehead Institute for use by the HGP. Equivalent work could now be performed on a single benchtop.* *Credit: International Human Genome Sequencing Consortium (2001) [Link]. Used with permission.*

A standard paradigm for research is that if I am interested in a specific research question X, I might collect data relating to X, but I will analyze my new data in the context of other existing data sets. For example, a project that collects any kind of human sequencing data will usually map the sequence reads onto the human Reference Genome, will probably rely on standard gene annotations, and will likely also make use of other more-specialized data sets.

In addition to these large-scale public data sets, researchers also benefit from an enormous number of smaller data sets that analyze specific samples or questions. It's been a huge boon to science that during the last two decades, there has been a strong shift toward making data freely available without preconditions [30]. This is part of a larger movement toward *open science*, which emphasizes the value of making all the results and tools of research publicly available as far as possible [31]. It's now widely recognized that anyone who publishes research in a scientific journal has a responsibility to make the underlying data publicly available [32].

The Human Reference Genome. The central data set that underlies everything practically everything else is the human Reference Genome. For example, in virtually any project that involves sequencing, the first step of data analysis is usually to map reads to the Reference Genome. This genome sequence was the main product of the **Human Genome Project (HGP)**, a huge, $3 Billion international effort that ran from 1990 to 2003, including teams from the US, Britain, Japan, France, Germany, and China.

By the mid-1980s techniques for mapping and sequencing DNA had reached a point where a number of leading scientists started to argue for a "moonshot" type of project to sequence the human genome. Early on, this audacious goal was highly controversial: critics said that it would be purely technical and scientifically uninteresting; that it would divert money from more-focused research; that it is wasteful to sequence the

99% of the genome that is noncoding; that we wouldn't know how to interpret the finished sequence anyway; and that the project would contribute to misconceptions of genetic determinism [33].

Despite the controversy, the genome project was greenlighted by the US Congress in 1990 [g]. During the early years it developed physical and genetic maps of the chromosomes and sequenced several much smaller genomes including the worm *C. elegans* and the fly *D. melanogaster*. During this time frame the costs of sequencing also dropped steadily due to technical advances. But in 1998 a privately funded company named Celera announced a plan to beat the public project to completion using a different strategy, thus spurring the Human Genome Project into a much faster timeline [35].

[g] *This came the year after one of the all-time great presidential malapropisms, when George HW Bush lauded the "**Human Gnome Initiative**" at a White House ceremony on recombinant DNA* [34].

In the end, the public project and Celera battled to a negotiated draw, announcing simultaneous completion of draft genomes in the year 2000. The completion was a major international news event, and the announcement was made by US President Bill Clinton and British Prime Minister Tony Blair in a White House ceremony [Link]. The genome was announced complete three years later [36]. Although the project was controversial at the time, modern biology could not exist in anything like its current form without genomes.

Given that everyone's genome is unique, you might wonder what exactly is in the Reference Genome. In a quirk of history, the sequence is based on a mixture of anonymous donors who were recruited by a newspaper ad in Buffalo, New York, in 1997. At any given position, the reference genome reflects the sequence of a single donor; thus, at many positions, the Reference Genome carries rare, and sometimes even deleterious, alleles [37]. About 70% of the Reference Genome comes from just one of the Buffalo donors, denoted RP-11. Analysis of sequences from RP-11 shows that he had mixed African and European ancestry in roughly equal proportions. Most of the rest of the Reference Genome is derived from ten other donors of East Asian or European ancestry [38].

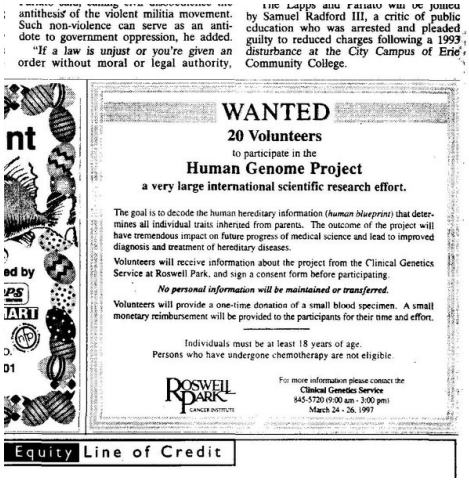


Figure 1.21: **Ad in the Buffalo News, 1997.** *The donors for the Human Genome Project were recruited using this ad, placed in the Buffalo (NY) newspaper on 3/23/1997, by Pieter de Jong of the Roswell Park Cancer Institute.*

Since the end of the Human Genome Project, a group called the Genome Reference Consortium has continued to update the Reference, by fixing assembly errors and providing alternate builds in certain regions with high levels of structural variation. You can download the genome, or browse specific regions using genome browsers at UCSC [Link] or Ensembl [Link]. Genomes for hundreds of other species can also be accessed through the same websites.

While the genome was announced complete in 2003, "complete" was used as a term of art that didn't actually mean *complete*. At that time, existing techniques were unable to span the most repetitive regions of the genome, including the centromeric and telomeric repeat regions, huge arrays of ribosomal DNA genes, and recent segmental duplications. The 2003 genome only covered 2.8 Gb (out of about 3.1 Gb) and included an estimated 341 gaps. The essential problem was that these regions of the genome contain large blocks of highly repetitive DNA that could not be

assembled: imagine trying to assemble a jigsaw puzzle with huge repetitive blocks and a mixture of many pieces randomly sampled from each block. During the next 15 years, even while genome sequencing became massively cheaper (by a factor of 10^5-fold), the sequence reads didn't get longer and most of these regions remained untouchable.

However, by the late 2010s, advances in ultra-long read sequencing using technologies developed by companies called PacBio and Oxford NanoPore enabled extraordinarily long reads that can bridge right across these repetitive elements. Using a mixture of these technologies, the **Telomere-to-Telomere Consortium** announced the first fully assembled human genome in 2021 [39]. We can expect that their work will usher in a new generation of genome sequencing in humans and many other species.

Functional annotation. Of course knowing the genome sequence is only a first step toward understanding the information encoded in it. Since the 1990s there has been a huge amount of work to identify the genes and regulatory elements and to understand their functions. Annotations showing the locations of genes and exons, and their splicing patterns, have been developed by two major projects: **RefSeq** and **GENCODE**. As we discussed above, a more challenging problem is to interpret the regulatory information encoded in genomes. The main approach to this uses a variety of experimental assays; much of this work has been performed and analyzed by the **ENCODE** Consortium. Gene expression profiles for different tissues and cell types are available from **GTEx** and **Human Cell Atlas**, respectively. Information about gene functions can be obtained from many sources, including comprehensive databases from **UniProt**, and **GeneCards**.

Human genetic variation. The Reference Genome only provides a single DNA sequence, and thus doesn't tell us anything about genetic variation across individuals. Thus, as the Human Genome Project was ending, it was recognized that the Genome would be much more powerful if we also had a good catalog of which sites in the genome are variable.

To address this need, from 2002 to 2010, the **International HapMap Consortium** created cell lines from around 100 individuals each from 11 global populations intended to represent some of the world's largest groups. Each individual was genotyped at up to about 3 million known single nucleotide variants across the genome. This landmark work created the first genomewide maps of genetic variation, and paved the way for a huge range of studies.

Subsequently, from 2008 to 2015, the **1000 Genomes Project** performed genome sequencing of a total of 3,200 individuals from 26 human populations. All of the data are freely available for browsing or bulk download [Link]:

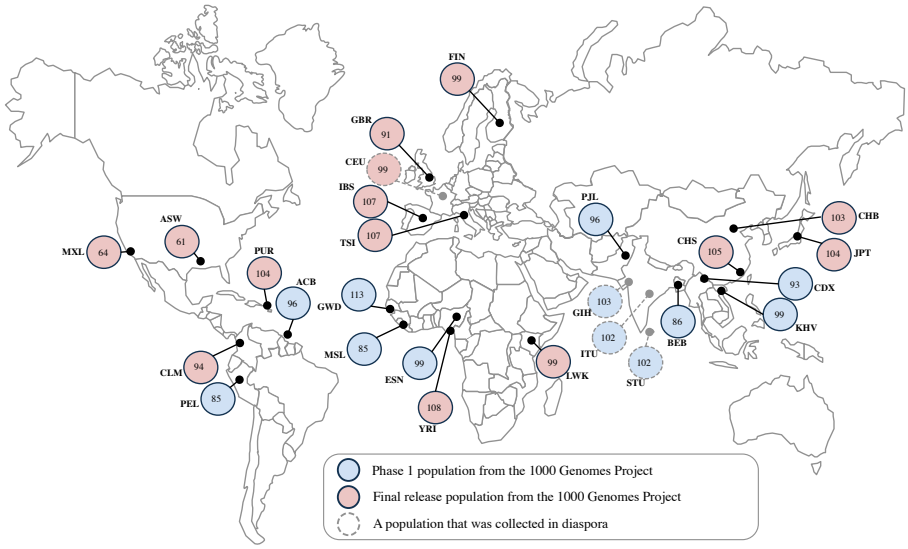


Figure 1.22: **The 1000 Genomes Project** *provides an essential reference set of human genomes from 26 human populations. The blue samples are mainly from HapMap. Some populations (dotted lines) were collected at a different location than their recent ancestral origins, such as CEU (west-Europeans in Utah), and GIH (Gujarati Indians in Houston, Texas).* *Credit: Modified from Fig. 1 of Taras Oleksyk et al (2015) in GigaScience. CC BY 4.*

Four of these populations are often used as example populations for data analysis and figures, so it's worth remembering their sample codes: **YRI** is a group of Yoruba individuals (a west-African ethnic group) sampled in Ibadan, Nigeria; **CEU** is a group of west-European descent individuals sampled in Utah; **CHB** and **JPT** are Chinese Han and Japanese individuals sampled from Beijing and Tokyo, respectively.

While the 1000 Genomes is an essential resource for many purposes, it has poor coverage of some human populations, especially smaller indigenous groups. For example, some groups including southern Africans, Papuans, Pacific Islanders, and Native Americans, are poorly covered by the 1000 Genomes Project. In contrast, the **Human Genome Diversity Panel (HGDP)** and a later extension, the **Simons Genome Diversity Panel (SGDP)**, provide much broader sampling of indigenous populations, albeit with fewer individuals [40]. These panels have helped to reveal wonderful insights into human history that would not have been possible with 1000 Genomes alone; we'll return to these especially in Part 3 of the book.

Figure 1.23: **The Simons Genome Diversity Project** *(shown here) consists of 300 genomes from 142 human populations. The* ***HGDP*** *consists of 1050 genomes from 52 populations.* *Credit: Image courtesy of Simons Foundation [Link].*

Lastly, another important study design involves cataloging genetic variation by sequencing extremely large samples, such as **gnomAD** and **TOPMed** [41].

The genotype-phenotype relationship. Our final category of data sets aim to measure the effects of genotype on phenotype. The most influential of these is an extraordinary dataset collected by the **UK Biobank**, which has collected genome-wide genotypes, and a huge array of phenotypic measures on about 500,000 British residents. Enrollment began in 2006, targeting an age range of 40–69, and continuing to track those individuals through middle and old age. Any qualified researcher can go through an application process to get access to the de-identified data. Due to the relative ease of data access, and the richness of information available, the UK Biobank has had a huge impactful on our understanding of human genetics. It's not a large exaggeration to say that the UK

Biobank has managed to get all of the world's human geneticists studying the British population.

Other very large cohorts have replicated aspects of the UK Biobank, including Biobank Japan, the China Kadoorie Biobank, FinnGen, the Estonian Biobank, the Million Veteran Program and All of Us (the latter are both in the US). These other cohorts either have less data at present, or are less accessible to outside researchers than UK Biobank. There are also disease-specific projects, such as the Psychiatric Genetic Consortium, that aggregate case-control data for focused study of particular diseases. One important concern about current cohorts is that, in aggregate, individuals of recent European descent are over-represented across these studies. This challenges human geneticists to ensure that the future benefits of genetic research can be shared equitably among people from all ancestries [42].

In this first chapter we have given an overview of some important background that will be helpful before we dive more deeply into the main areas of human genetics. We next turn to a more focused description of human genetic variation.

Notes and References.

[2]Although the Human Genome Project was declared complete in 2003, about 10% of the genome was unsequenceable at that time. The first truly complete human genome sequence was reported in 2021 and published the following year:

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022;376(6588):44-53

[3]As with some other complicated topics, for the sake of brevity we will generally simplify important points relating to sex, gender and familial relationships, except when the complexities are specifically relevant. For example it's convenient to refer to XX and XY individuals as female and male respectively. We do so despite the fact that (i) biological sex is not entirely binary – some individuals have physical characteristics of both sexes due to mutations in sex-determination genes, unusual karyotypes such as XXY, or other causes not all of which are currently understood; (ii) biological sex does not necessarily correspond to gender; gender is actually *more* relevant than biological sex for many aspects of our lived experiences – even though it is generally less connected to the core topics of this book; (iii) familial relationships do not always imply genetic relationships – for example in the case of parents of adopted children.

[4]During our lives, our bodies produce about a light-year of DNA: [Link].

[5]For more on DNA storage systems see eg:

Erlich Y, Zielinski D. DNA Fountain enables a robust and efficient storage architecture. Science. 2017;355(6328):950-4

Kim J, Bae JH, Baym M, Zhang DY. Metastable hybridization-based DNA information storage to allow rapid and permanent erasure. Nature Communications. 2020;11(1):1-8

[6]Improbable Research's video of Eric Lander's 24 second and 7 word descriptions of the human genome: [Link]

[7]In 2021 the AlphaFold team reported huge progress on computational prediction of protein folding, thereby helping to transform this field:

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583-9

[8] An important set of exceptions to the standard genetic code is found in the mitochondrial genome. The mitochondrion is thought to have evolved from an endosymbiotic prokaryote, and it still retains a very small genome of its own. This genome is so small that rare minor changes in the genetic code have been tolerated by natural selection. Specifically, the genetic code in vertebrate mitochondria differs from the conventional code at four triplets: AGA and AGG are stop codons instead of arginine; TGA codes tryptophan instead of stop; and ATA codes methionine instead of isoleucine.

[9]There are various categories of genes in which the RNA itself is functional. For example, in females one copy of the X chromosome is inactive in each cell; this is achieved in part by transcribing an RNA called Xist off one of the two X chromosomes. The Xist transcript coats that X chromosome and prevents transcription from most other genes. Xist is an example of what is known as a long noncoding RNA (lncRNA). In addition to lncRNAs, other functional RNA genes categories include microRNAs, transfer RNAs, ribosomal RNAs, and piRNAs.

[10]Another important exception to the Central Dogma is that some viruses use RNA as their genetic material, and then use an enzyme called *reverse transcriptase* to make a DNA copy for replication. Reverse transcriptase is also used in the lab to make DNA copies of RNA when we want to sequence RNA.

[11]The fact that the introns are so very long is probably not functionally important in most cases, and instead reflects a tendency for genomes to accumulate noncoding junk, as we will discuss below.

[12]There is some uncertainty about exactly how much alternative splicing is functionally important. One approach that is often used to evaluate functional importance of biological features is whether a feature is maintained (conserved) over evolutionary time, or whether it evolves rapidly, suggesting malleability and (usually) lower functional importance. Curiously, alternative splicing patterns (specifically, exon skipping events) are not very conserved across species – and are less conserved than overall expression levels. However interpretation of this is not entirely clear:

Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, et al. The evolutionary landscape of alternative splicing in vertebrate species. Science. 2012;338(6114):1587-93

Merkin J, Russell C, Chen P, Burge CB. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. Science. 2012;338(6114):1593-9

Li YI, Knowles DA, Humphrey J, Barbeira AN, Dickinson SP, Im HK, et al. Annotation-free quantification of RNA splicing using LeafCutter. Nature Genetics. 2018;50(1):151-8

[13]There's been quite a bit of interesting work on the sequence controls of splicing; these include both high-throughput experimental approaches as well as machine learning methods to learn highly complex rules from genome sequence data or experiments. See for example:

Rosenberg AB, Patwardhan RP, Shendure J, Seelig G. Learning the sequence determinants of alternative splicing from millions of random sequences. Cell. 2015;163(3):698-711

Jaganathan K, Panagiotopoulou SK, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting splicing from primary sequence with deep learning. Cell. 2019;176(3):535-48

Zeng T, Li YI. Predicting RNA splicing from DNA sequence using Pangolin. Genome Biology. 2022;23(1):1-18

[14]For example Down Syndrome occurs in individuals who have an extra copy of Chromosome 21. Chromosome-level changes in copy number change the expression levels of the genes on that chromosome relative to the genes on other chromosomes. It's interesting to note that cells can often tolerate duplication of the entire genome better than duplication of a single chromosome, as whole-genome duplication maintains the relative proportions of genes. Somewhat similarly, many monogenic diseases are due to defects in the core transcriptional machinery, leading to broad transcriptional dysregulation rather than disruption of specific biological pathways; see Table 6.2 of

Calof AL, Santos R, Groves L, Oliver C, Lander AD. Cornelia de Lange syndrome: Insights into neural development from clinical studies and animal models. In: Neurodevelopmental Disorders. Elsevier; 2020. p. 129-57

For example, Cornelia de Lange Syndrome is due to mutations that disrupt the cohesin complex; these cause minor disruptions of many genes leading to diverse developmental disorders.

[15]Muniz L, Nicolas E, Trouche D. RNA polymerase II speed: a key player in controlling and adapting transcriptome composition. The EMBO journal. 2021;40(15):e105740.

[16]Technically, the direct copy of DNA is called a pre-mRNA. This must be spliced to produce the mature mRNA. Most splicing occurs at the same time as transcription.

[17]Expression (i.e., mRNA levels) of any given gene depend on the rate of transcription in the relevant cell type (defined as the number of new mRNAs synthesized per unit time), and the mRNA decay rate. For most genes, control of gene expression acts mainly on synthesis.

[18]An exception is that several proteins called General Transcription Factors are components of the Pre-Initiation Complex and lack DNA binding domains.

[19]Most of the genome is bound by nucleosomes, and TF binding requires nucleosome removal. This can be much more stable if multiple TFs can bind within the same nucleosome-free region.

[20]There's a large, growing body of work using machine learning approaches to predict enhancer regulatory activity, e.g.,

Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. Nature Methods. 2015;12(10):931-4

Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Research. 2016;26(7):990-9

Avsec Ž, Weilert M, Shrikumar A, Krueger S, Alexandari A, Dalal K, et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. Nature Genetics. 2021;53(3):354-66

[21]One famous example of long-range looping occurs at the FTO locus:

Sobreira DR, Joslin AC, Zhang Q, Williamson I, Hansen GT, Farris KM, et al. Extensive pleiotropism and allelic heterogeneity mediate metabolic effects of IRX3 and IRX5. Science. 2021;372(6546):1085-91

[22]For empirical work on predicting enhancer-promoter interactions see e.g.,

Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, et al. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. Nature genetics. 2019;51(12):1664-9

[23]This number is a bit rough because we still don't have a complete accounting of functional regulatory sequences in all cell types. But around 10% of the genome shows signals of evolutionary conservation. This provides an estimate of what fraction of the genome is functional – in the sense that changes in the DNA sequence have consequences for organismal fitness.

Ward LD, Kellis M. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. Science. 2012;337(6102):1675-8

Rands CM, Meader S, Ponting CP, Lunter G. 8.2% of the human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. PLoS Genetics. 2014;10(7):e1004525

[24]For statistics about gene sizes see

Piovesan A, Caracausi M, Antonaros F, Pelleri MC, Vitale L. GeneBase 1.1: a tool to summarize data from NCBI gene datasets and its application to an update of human gene statistics. Database (Oxford). 2016;2016.

[25]Many of these regions are transcribed but not translated; as noted above, these are referred to as long noncoding RNA (lncRNA) genes. Some lncRNA genes play essential roles, but most show limited evolutionary conservation and only

a tiny fraction are currently associated with putative functions, suggesting that most lncRNAs are likely nonfunctional:
Ransohoff JD, Wei Y, Khavari PA. The functions and unique features of long intergenic non-coding RNA. Nature Reviews Molecular cell biology. 2018;19(3):143-57
Ponting CP, Haerty W. Genome-wide analysis of human long noncoding RNAs: a provocative review. Annual Review of Genomics and Human Genetics. 2022;23

[26]See for example L1 silencing mechanisms:
Liu N, Lee CH, Swigut T, Grow E, Gu B, Bassik MC, et al. Selective silencing of euchromatic L1s revealed by genome-wide screens for L1 regulators. Nature. 2018;553(7687):228-32.

[27]For examples in which TEs have been co-opted by their host genomes see
Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science. 2016;351(6277):1083-7
Bartonicek N, Rouet R, Warren J, Loetsch C, Rodriguez GS, Walters S, et al. The retroelement Lx9 puts a brake on the immune response to virus infection. Nature. 2022:1-9

[28]Mitosis and meiosis are complicated and deeply studied processes, and it's impossible to do them justice here. We'll touch on a few of those complexities later in the book as they become relevant.

[29]To be more precise, meiotic recombination events can be resolved either with crossover or non-crossover events. Non-crossovers involve copying of a small region (average 30–40bp in mice) from one chromosome to the other.
Li R, Bitoun E, Altemose N, Davies RW, Davies B, Myers SR. A high-resolution map of non-crossover events reveals impacts of genetic diversity on mammalian meiotic recombination. Nature Communications. 2019;10(1):3900
While non-crossovers are very common they are difficult to detect in data. However the term "recombination" is often used in human genetics synonymously with crossovers.

[30]Some of the major resources, such as the human genome and 1000 Genomes Project data sets are freely downloadable. Other data sets such as the UK Biobank contain personal information about research subjects, albeit anonymized, and can only be used by qualified researchers who agree to certain conditions for appropriate use of the data. However in all these cases, researchers have a large amount of flexibility in how they use the data for their own analyses.

[31]Open science: [Link].

[32]For example, the prominent journal Nature writes on their website: "It is a condition of publication that authors deposit their data in an appropriate repository, and agree to make the data publicly available without restriction, excepting reasonable controls related to human privacy or biosafety." [Link], accessed 10/01/2021.

[33]Roberts (2001) wrote "Sydney Brenner of the MRC facetiously suggested that project leaders parcel out the job to prisoners as punishment–the more heinous the crime, the bigger the chromosome they would have to decipher."
Lewontin R. The dream of the human genome: doubts about the Human Genome Project. The New York review of books. 1992;39(10):31-40
Roberts L. The Human Genome. Controversial from the start. Science. 2001;291:1182-8

[34]This was in a White House ceremony in 1989 to award the National Medal of Honor to Stan Cohen and Herbert Boyer who developed recombinant DNA technology; as recalled by Carol Ezzell in Scientific American, July 2000 [Link].

[35]There was a great deal of acrimony between the two groups, not least because Celera's build incorporated data that the Human Genome Project was releasing into the public domain on a daily basis (in part to prevent attempts to patent genes). Some of the back-and-forth can be found here: HGP critique
Waterston RH, Lander ES, Sulston JE. On the sequencing of the human genome. Proceedings of the National Academy of Sciences. 2002;99(6):3712-6;
Celera reply: Myers EW, Sutton GG, Smith HO, Adams MD, Venter JC. On the sequencing and assembly of the human genome. Proceedings of the National Academy of Sciences. 2002;99(7):4145-6

[36]Flagship papers on the Human Genome Sequence:
International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921
Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science. 2001;291(5507):1304-51
International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004;431(7011):931-45

[37]There have been occasional calls to change the reference to remove rare alleles, but such large changes to the reference genome would create all kinds of compatibility issues and in this case the medicine may be worse than the disease.
Ballouz S, Dobin A, Gillis JA. Is it time to change the reference genome? Genome Biology. 2019;20(1):1-9

[38][Link] and p 146 of the supplementary information of
Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the Neandertal genome. Science. 2010;328(5979):710-22

[39]Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022;376(6588):44-53

[40]The HGDP was started at Stanford in the early 1990s by two of my mentors, Luca Cavalli-Sforza and Marc Feldman. This project pioneered the concept of collecting cell lines from diverse human populations as permanent resources for studies of genetic diversity, a concept later adopted by HapMap and 1000 Genomes. The HGDP was used for limited genotyping in the 1990s, genomewide genotyping in the 2000s and, ultimately, whole genome sequencing in the 2010s.
Rosenberg NA. Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Annals of human genetics. 2006;70(6):841-7
Hellenthal G, Busby GB, Band G, Wilson JF, Capelli C, Falush D, et al. A genetic atlas of human admixture history. science. 2014;343(6172):747-51
Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons genome diversity project: 300 genomes from 142 diverse populations. Nature. 2016;538(7624):201-6

[41]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
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

[42]Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature. 2016;538(7624):161-4