1.4 DNA sequencing: a fundamental tool for studying biology.

In which we take a detour to discuss DNA sequencing and genotyping.

Microscopes were invented in the early 1600s, and opened up a new unimagined world. Robert Hooke is credited with the discovery of plant cells in 1665; shortly after Antonie van Leeuwenhoek observed microbes including bacteria and protozoa, as well as human cells including spermatozoa, blood, and muscle cells. Despite important recent advances in microscopy, DNA and many of the processes that we focus on in genetics, are too small to see clearly by microscopy. In particular, DNA molecules are much too small to read the nucleotides directly.

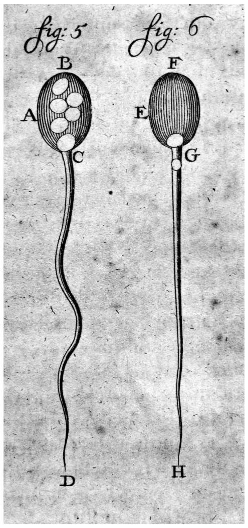

Figure 1.45: **van Leeuwenhoek's drawings of spermatozoa (1719)**. *The early microscopists revealed for the first time a world of cells, fine structures of tissues, and microbial life.* ***DNA sequencing is, similarly, now revealing new worlds.*** *Credit: Opera Omnia (1719). [Link] CC BY 4.*

However, starting from the 1970s, there has been a rise of new technologies that allow us to read the nucleotide sequences of DNA molecules: this is **DNA sequencing**. In modern biology, **DNA sequencing has become a truly transformative tool**, opening up new avenues of exploration that were unimagined prior to the sequencing era.

First, and probably most obvious, it's now relatively cheap and easy to sequence genomes. We now have genome sequences for thousands of different species. Hundreds of thousands of different humans have now been genome sequenced.

But beyond this, there's a dizzying array of different things that we can now measure using DNA sequencing technology [a]. For example if a patient has cancer, we can sequence the genome of the cancer cells, to understand what genetic changes enable uncontrolled growth (which may indicate the use of particular treatments). When a woman is pregnant, we can sequence her baby's DNA, using free-floating DNA fragments in the mother's bloodstream to predict genetic diseases before birth. We can use DNA sequencing to measure the microbial population that lives in everyone's guts (the microbiome), or in agricultural or wild soil samples. We can use sequencing to detect the presence of viruses in patients, or in the environment: on surfaces, in the air, or in wastewater. Sequencing has been an essential tool for tracking the evolution of the SARS-CoV-2 virus, and to identify outbreaks of novel strains.

In the lab, DNA sequencing has also transformed genomics research. We use DNA sequencing to measure many different aspects of how a genome functions in different cell types: for example which parts of the genome are bound by a particular protein; which parts of the genome are actively involved in regulating genes; which genes are being expressed, and how much; what cell types are present in a tissue sample; sequencing is being used to transform developmental biology. Almost any lab experiment involving genomes or cellular functions can now be set up to end with data collection by DNA sequencing.

In short, DNA sequencing is like a new microscope for the 21st century[75].

[a] *DNA sequencing is a fundamental technology that allows highly efficient detection, counting and sequencing of biological molecules. As such it is transforming a vast array of different applications in biological research and medicine, not just limited to the determination of genome sequences.*

A short history of sequencing. The first practical DNA sequencing was achieved in the 1970s. Two scientists (Fred Sanger and Walter Gilbert) won the Nobel prize in 1980 for developing techniques that could sequence up to around a hundred basepairs of DNA at a time. During the half-century since this work, sequencing has become millions of times faster and cheaper, now enabling rapid, affordable whole genome sequencing.

First generation sequencing. The technique introduced in 1977 by Fred Sanger – now known as **Sanger sequencing** – provided the first practical sequencing, and was the basis for nearly all sequencing projects until about 2005 [76]. Sanger sequencing is still used in lab-work for small-scale applications [77].

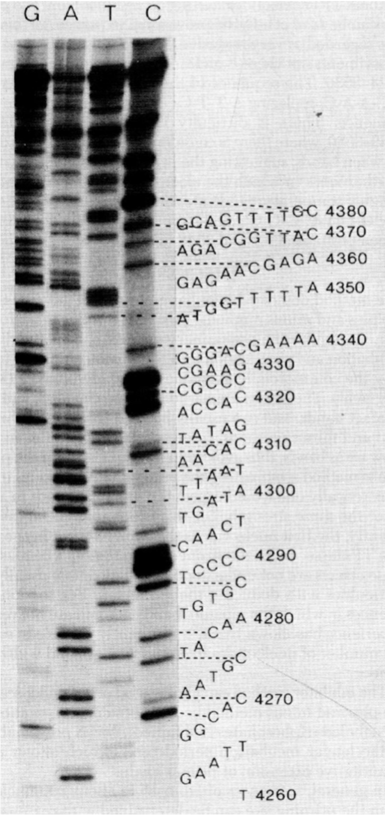

Figure 1.46: **Sequencing autoradiograph** *from the 1977 paper that introduced Sanger sequencing. The image displays a short sequence from the virus ϕX174. Each band corresponds to fragments of a specific nucleotide length (shortest at the bottom); each lane contains fragments that terminate in the nucleotide shown at top. Though blurry at points, this allowed the sequence to be read from the image, as shown at right.* *Credit: Figure 2 from Fred Sanger et al, 1977 [Link].*

In Sanger sequencing, DNA polymerase is used to copy a single-stranded DNA template. The reactions are run in a soup of ordinary DNA nucleotides and a small fraction of so-called dideoxy nucleotides, which block further extension of the sequence. In the original Sanger sequencing, template copying was performed in four distinct reactions, one for each termination nucleotide: with dideoxy-A, dideoxy-C, and so on. Ultimately, the dideoxy-A reaction would contain a collection of DNA fragments of different sizes, corresponding to all the fragment sizes that end in A. The fragments were run out on an electrophoretic gel that separates molecules by size. DNA fragments were labeled with radioactive atoms so that they could be detected on x-ray films. You can see an example in the image at the right.

In later iterations of Sanger sequencing, each dideoxy nucleotide was instead labeled with a different fluorescent dye. This means that sequencing could be performed in a single reaction, and the bands can be recognized by color using a scanning laser. This in turn enabled further miniaturization, as each reaction could be run through a thin capillary tube, instead of in a large slab gel.

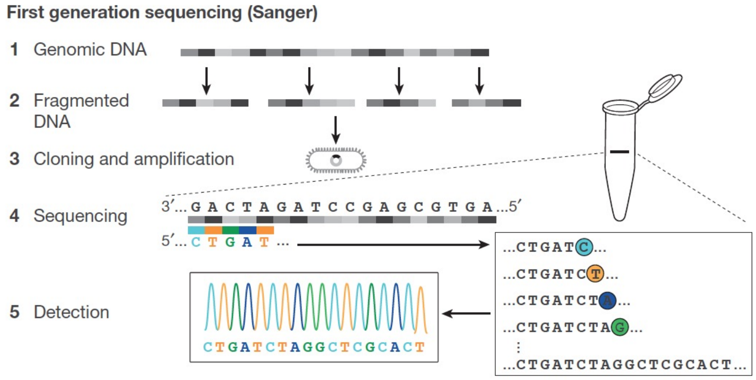

Figure 1.47: **Sanger Sequencing.** *In modern Sanger sequencing, DNA fragments of different sizes are labeled with fluorescent dyes according to the 3′ nucleotide on each fragment. The fragments are size-separated using gel electrophoresis, and the colors are recorded by a scanning laser as they migrate through the gel, thereby providing the DNA sequence.* *Credit: From Figure 1, Jay Shendure et al. 2017. [Link]Used with permission.*

These improvements to Sanger sequencing enabled the first sequencing of eukaryotic genomes in the 1990s, leading to the draft human genome in 2000 – completed at a cost of $2.7 Billion, calculated in 1991 dollars [78].

Second generation sequencing. But at these prices, genome-scale work was extremely expensive, and limited to “genome centers”, which were essentially dedicated sequencing factories.

The early 2000s saw a major paradigm shift, with a group of new sequencing technologies that achieved enormous advances over Sanger sequencing. These new methods – also called **next-generation** or **massively parallel sequencing** – were dramatically faster, and required smaller amounts of expensive reagents. These became commercially available by around 2006 and greatly reduced sequencing costs, enabling individual labs to perform genome-scale sequencing projects for the first time. Over the next few years one technology, owned by the company Illumina, gained a dominant position in the DNA sequencing market and currently enjoys a near-monopoly in high throughput sequencing [79].

Illumina's approach [80] starts by attaching billions of DNA fragments to a solid surface called a **flow cell**, which is similar to a microscope slide. These are used to create colonies of identical single-stranded DNA fragments. Sequencing proceeds using a **sequencing by synthesis** approach, in which fluorescent nucleotides are added to the complementary strand one-at-a-time. At each cycle of the experiment, one of the four possible nucleotides is added to each colony, depending on the sequence on the template strand, and the corresponding colors are recorded. The sequence of colors for each colony indicates the correct sequence.

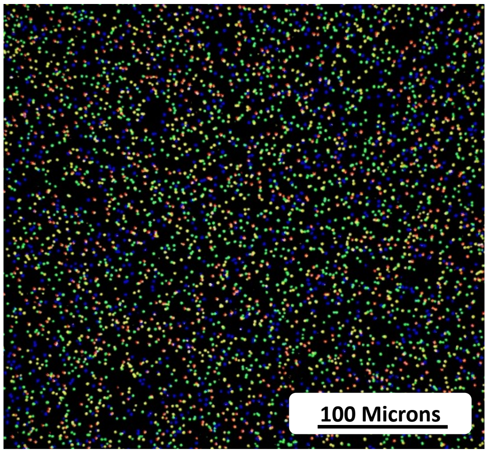

Figure 1.48: **Illumina flowcell.** *The image shows a tiny part of a flowcell. Each dot represents a DNA cluster, and the colors indicate the nucleotide added in the current cycle. Original image source unknown.*

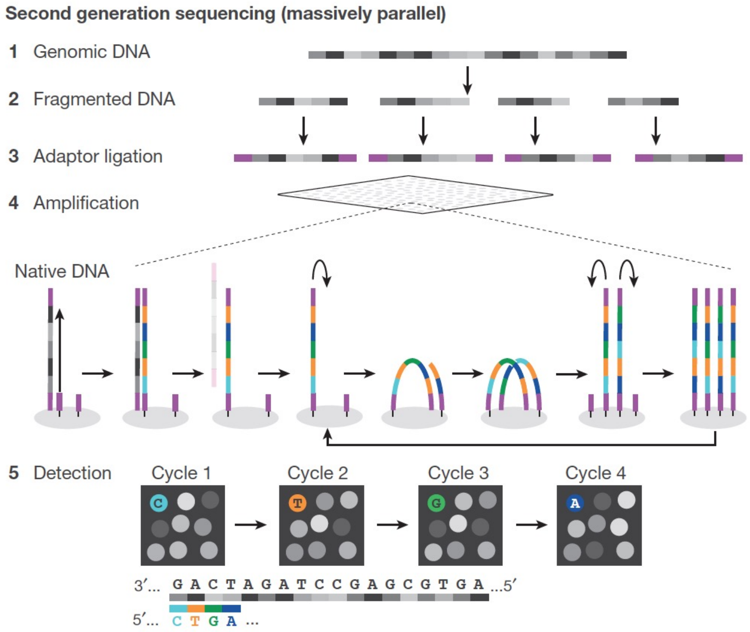

Figure 1.49: **Massively parallel short-read sequencing**, *e.g., Illumina. Colonies of identical single-stranded DNA fragments are attached to a solid surface. Sequencing occurs through DNA synthesis, as colored nucleotides are added one at a time and imaged. Credit: From Figure 1, Jay Shendure et al. 2017. [Link]Used with permission.*

Illumina sequencing is vastly more efficient than Sanger sequencing. In Sanger sequencing each reaction must run through a separate electrophoretic channel (a capillary, or lane on a gel); in contrast, Illumina sequencing is only limited by the number of DNA colonies that can be placed and imaged on a flow cell without overlapping.

However, an important limitation of Illumina sequencing is that the sequence reads are relatively short. The current main platform sequences 150 bp from both ends of a larger molecule (typically one might input

DNA molecules of perhaps 600 bp, and then sequence 150 bp from each end). Modern Sanger sequencing reads are slightly longer, reaching up to ∼800 bp. In sharp contrast, Third Generation methods, which we discuss next, are routinely tens of kilobases and can reach megabase lengths.

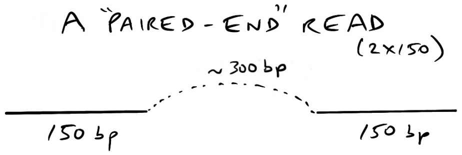

Figure 1.50: **Paired-end reads** *are a standard sequencing format on the popular Illumina platform. This involves adhering both ends of a larger DNA fragment to the flow cell, and sequencing from both ends. The precise read lengths and fragment sizes vary across applications.*

Third generation sequencing. The 2010s have seen the emergence of a third paradigm for sequencing, which for the first time involves direct sensing of individual DNA molecules. At the time of writing, these are lower throughput and with higher error rates than second-generation sequencing, but they can provide extremely long sequence reads of individual molecules, potentially even to megabase-length reads.

At present, the two leading commercial technologies are from Oxford Nanopore Technologies and Pacific Biosciences. Oxford Nanopore's approach measures electrical conductance of a DNA molecule as it passes through a biologically-derived membrane channel (a nanopore). The different nucleotides can be recognized as producing different electrical signatures. PacBio positions individual DNA polymerases inside a measurement well that can accommodate a single DNA molecule. The well detects light emitted by fluorescent nucleotides that are incorporated, one-at-a-time, into a single growing DNA strand.

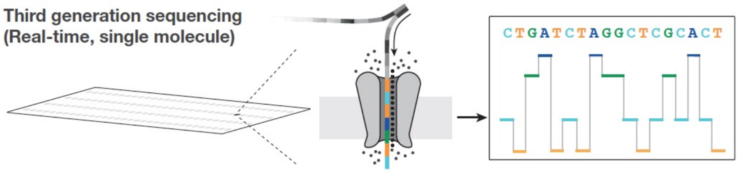

Figure 1.51: **Third generation single-molecule sequencing.** *Technologies including Oxford Nanopore and PacBio pass single molecules through molecular sensing devices. These can provide read lengths up to about 1 Mb.* Credit: *From Figure 1, Jay Shendure et al. 2017. [Link]Used with permission.*

At present, the long-read technologies are lower throughput and more error-prone than Illumina's short-read platforms; a 2020 paper estimated that the error rate per base pair for a single sequencing read can be as high as 10% versus around 0.1% for Illumina [81]. However, by sequencing to high depth it is possible to achieve comparable overall accuracy, albeit at higher cost per sample (about $5000 on the Nanopore platform for a clinical-grade genome in 2022 [82]). We can expect error rates and costs for 3rd-Gen sequencing to continue to drop.

Moreover, 3rd-Gen long reads enable some applications that are difficult, if not impossible, with short reads: these include sequencing of complex regions of the genome, and haplotype phasing. **3rd-Gen long reads were essential for the first truly complete human genome sequence, published in 2022** [83]. Furthermore, we can anticipate new advances in this space: for example, because these methods sequence individual molecules without amplification, it's possible to study DNA or RNA modifications such as methylation. Lastly, Oxford Nanopore sequencing is performed on portable USB devices that plug directly into a laptop – this makes it practical for field applications such as infectious disease surveillance in developing countries.

We can summarize the three main sequencing approaches as follows:

Sequencing Type	Read Length	Throughput	Error Rates	Single molecule
1st Gen (Sanger)	$\sim$800 bp	low	low	no
2nd Gen (e.g., Illumina)	$\sim 2\times 150$ bp	very high	low	no
3rd Gen (e.g., Nanopore, PacBio)	up to $\sim$1 Mb	medium	high	yes

Moore's Law and the dropping costs of DNA sequencing. During the last three decades, the increases in speed, and decreases in cost, of DNA sequencing have been absolutely gobsmacking [b]. It's interesting to compare the enormous improvements in the DNA sequencing industry to gains in another industry, computing, which has famously benefited from extreme miniaturization. **Moore's Law** is an observation from the computer industry that the number of transistors on a circuit chip doubled roughly once every two years; this remarkable rate of progress fueled the rise of computing.

DNA sequencing improved even more rapidly than Moore's Law from around 2007-2012, driven in large part by the transition to massively parallel sequencing. Subsequent stagnation in costs partly reflects that a single company currently controls the vast majority of the short-read sequencing market [84].

[b] *"Back in 1990, sequencing 1 million nucleotides cost the equivalent of 15 tons of gold (adjusted to 1990 price). At that time, this amount of material was equivalent to the output of all United States gold mines combined over two weeks. Fast-forwarding to the present, sequencing 1 million nucleotides is equivalent to the value of $\sim$30 g of aluminum. This is approximately the amount of material needed to wrap five breakfast sandwiches at a New York City food cart." –Yaniv Ehrlich (2015).*

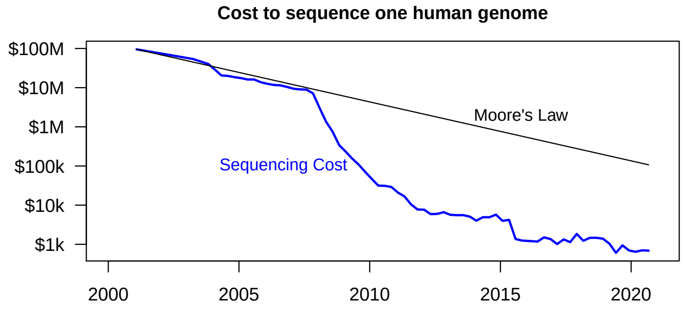

Figure 1.52: **The rapidly declining cost of DNA sequencing.** *The blue line shows the estimated cost to sequence one genome, from $95M in 2001 to $700 in 2020. The black line shows the Moore's Law prediction, projected forward from 2001.* *Credit: Redrawn from a figure and data by the US National Human Genome Research Institute (NHGRI) [Link].*

In the remainder of the chapter we will discuss the applications of sequencing in more detail, with an emphasis on "resequencing".

Sequencing applications in human genomics. We can classify most sequencing applications into three broad categories:

- **Genome resequencing and polymorphism discovery.** If we sequence your genome, and mine, what are all the places where we differ from each other? The analysis is usually performed by identifying differences from the Reference Genome. These applications are referred to as **resequencing** when analysis is based on comparison to a reference.

- **De novo genome sequencing and assembly.** How can we use sequencing to determine the genome of an unstudied species, or to determine the human genome in regions of high structural complexity and variability? Affordable 2nd- and 3rd-generation sequencing have now enabled genome assemblies for many thousands of different species, spread widely across the tree of life [85].

- **Sequencing as a molecular counting tool.** Most of this is outside our scope here, but since around 2005, there has been a huge shift toward using DNA sequencing as the readout for a huge array of molecular experiments: What are the expressed or regulatory regions of the human genome in any given cell type? What regions of the genome are amplified in a cancer cell? Which CRISPR guides lead to better cell survival in a functional screen?

In the remainder of this chapter we focus on the first of these goals, which is most essential for the topics of this book.

Genome resequencing and polymorphism discovery. Suppose we want to characterize the genetic variation in a sample of individuals; or perhaps we want to search for potential causal mutations in a child with severe developmental delays; perhaps we wish to find driver mutations in a cancer genome. How should we tackle these problems using current technologies?

Ideally you might imagine DNA sequencing providing a fully accurate end-to-end read-out of each chromosome. But no current technology can provide this directly. Instead, in practice we must balance a desire for high accuracy and completeness against considerations of cost and speed. At present (writing this in 2022), most resequencing projects are using Illumina short-read sequencing because Illumina reads are relatively cheap and accurate [c]. We discuss limitations below.

Resequencing with short reads. Remember that human chromosomes are 50-250 Mb long, but what we get are billions of short reads of ∼150bp. To make matters worse, there is no easy way to indicate where in the genome each read comes from. Indeed, most genome sequencing uses what is called **shotgun sequencing**, in which we break the genome into many small fragments, sequence them, and rely on our ability to make sense of the sequence reads when we have them. One further challenge is that all DNA sequencing comes with occasional errors (e.g., about 1 per thousand nucleotides per sequencing read on Illumina [86]) and the data analysis must be robust to this [d].

A common pipeline for sequencing human genomes is as follows:
• **Extract DNA** from a tissue sample, e.g., from blood cells. The initial sample contains millions of cells (each with its own copy of the genome), and so the DNA fragments that we sequence are a mixture of many different copies of the same genome;
• **Smash up the genome** into ∼600 bp fragments of DNA for shotgun sequencing;
• **Map reads** to a standard reference human genome (see below);
• **Infer genotype differences** from the reference (e.g., SNPs and structural variants)

[c] *In this section we focus on **whole genome sequencing (WGS)**. At the end of this chapter we describe two alternatives: **exome sequencing**, and **genotyping**.*

[d] *As we describe below, we usually have many sequence reads spanning every position in the genome. This allows us to correct errors so that the final error rate in a finished sequence is much lower than the raw rate per read.*

- **Interpretation of variation** – for example, identifying disease-associated variants.

The output from the sequencing machine consists of billions of short DNA sequence reads. Since we don't know in advance where each read comes from, the first step of analysis is to **map the reads** to a reference human genome. Conceptually you can think of read-mapping as being like taking a string of nucleotides and sliding it along the genome sequence until you find a location where it matches – very much like fitting a piece into a jigsaw puzzle. There are efficient computational algorithms to do this. The matching process must allow for modest levels of difference from the reference, as there may be SNPs, indels or sequencing errors.

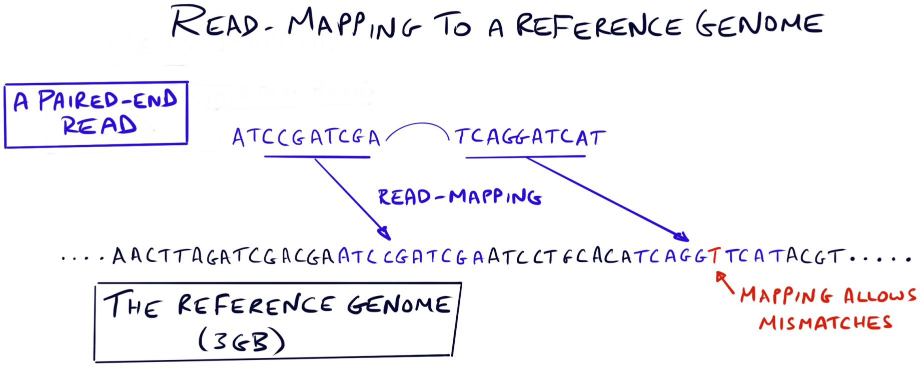
READ-MAPPING TO A REFERENCE GENOME
A PAIRED-END READ
ATCCGATCGA
TCAGGATCAT
READ-MAPPING
....AACTTAGATCGACGAATCCGATCGAATCCTGCACATCAGGTTCATACGT.....
THE REFERENCE GENOME (3GB)
MAPPING ALLOWS MISMATCHES

Figure 1.53: **Read mapping.** *Each paired-end read is compared to the reference genome to figure out where in the genome it came from – much like fitting a piece into a jigsaw puzzle. For a paired-end read we do not know the sequence in the internal part of the DNA fragment, but we do know the approximate size: for a correct match both sequence ends should fit into the reference genome with a gap of a few hundred base pairs between them. Low levels of mismatches (in red) are allowed as these may reflect SNPs or other types of variation.*

As described above, read mapping assumes that a read only matches a single location in the genome. But many sequences in the genome are repeated two or more times, so that it is ambiguous where a read comes from. This is especially problematic in the most complex regions of the genome, including centromeres, subtelomeric regions, ribosomal DNA clusters, and other locations where large blocks of DNA are repeated many times. Reads from transposable elements can also be difficult to map (although this depends on size – smaller elements such as the 300 bp Alu repeat, are generally mappable as long as any part of a paired-end read hangs off into unique sequence outside the element) [87]. Collectively, the most ambiguous regions are referred to as **unmappable**, and cover about 10% of the genome [88]. Third generation long-read sequencing is starting to resolve these regions that are inaccessible to short reads.

Genome coverage. A key experimental parameter for genome sequencing is referred to as **read depth** or **genome coverage**. These terms refer to the average sequencing depth in mappable regions of the genome.

For example, if we sequence a DNA library to a depth of "30X coverage", this means that an average (mappable) position in the genome is covered by 30 reads. 30X coverage is a commonly-used standard for high quality genomes: this relatively high read depth ensures that with high probability we have good coverage of both chromosomes at every position in the genome – as discussed next, this allows high-quality genotype calls

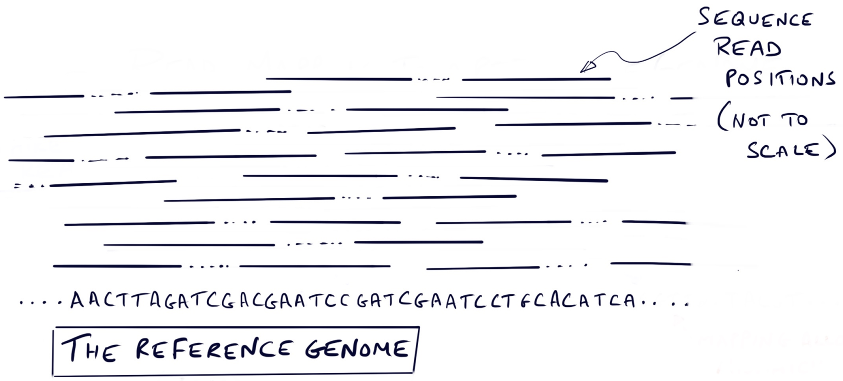
SEQUENCE READ POSITIONS (NOT TO SCALE)
....AACTTAGATCGACGAATCCGATCGAATCCTGCACATCA.....
THE REFERENCE GENOME

Figure 1.54: **Reads tiled across a genome.** *The number of reads (solid lines) that span any given position is the coverage. In practice, the reads are much longer relative to the reference genome than shown here.*

throughout the mappable genome. Since the human genome is about 3.1 GB this implies that we need ∼100 GB of sequence data per genome.

SNP calling. Our next goal is to identify SNPs and other types of variation. When we see a mismatch between the sequence read and the reference genome this might indicate one of several possibilities: a homozygous difference from the reference; a heterozygous difference; a sequencing error [e]. By getting deep sequence coverage of the genome we can distinguish these three scenarios:

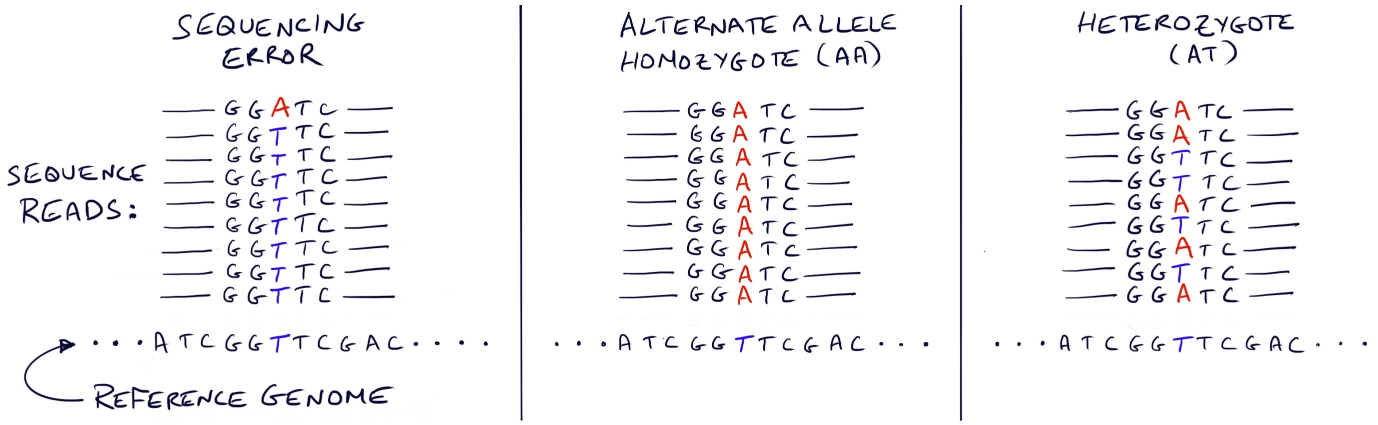

Figure 1.55: SNP calling from sequence data. *Sequencing errors occur at a rate of about 0.1% of nucleotides in current short-read data, but most of these mismatch the reference on only a single read. In contrast, homozyotes for the alternate allele differ from the reference on all reads (or nearly all reads, again remembering that there is a low error rate). Lastly, heterozygotes match the reference on about 50% of reads.*

An error rate of one per thousand nucleotides might sound pretty good, but it actually means that we get a lot more sequencing errors than actual SNPs. So for example with 30X coverage, we'll get a sequencing error roughly once every 30 bp, while true differences from the reference occur less frequently – around once per 500-1000 bp. This issue is particularly acute when we want to detect new mutations – as we'll see in Chapter 1.5, these are extremely rare in the genome, occurring about once per 100 million base pairs. This means that we need multiple supporting reads to confidently detect novel variants.

Lastly, one important point here is that while the read mapping tells us

[e] *Recall that an allele that matches the reference genome is known as the* ***reference allele****. The allele that differs from the reference is the* ***alternate allele****.*

where each read comes from in the genome, it cannot distinguish between the two homologous copies you have of each chromosome (the one from mum and the one from dad). This means that at a heterozygous site, we don't know which allele comes from which chromosome. This is another situation where long-read technologies can help out, by linking together heterozygous alleles that lie on the same chromosome.

Larger structural variants. So that gives you a sense of how SNP detection works. How can we detect larger structural variants like large deletions, using short reads?

Simplifying somewhat, we can think of two different kinds of information to detect events that are much larger than the scale of a single read-pair. First, most structural variants change the average depth of sequence reads: for example a heterozygous deletion cuts the average read depth to 50% of the genome average.

Second, structural variants may be recognized by the presence of read-pairs that span unexpectedly large distances along a chromosome, inconsistent with the DNA fragment sizes.

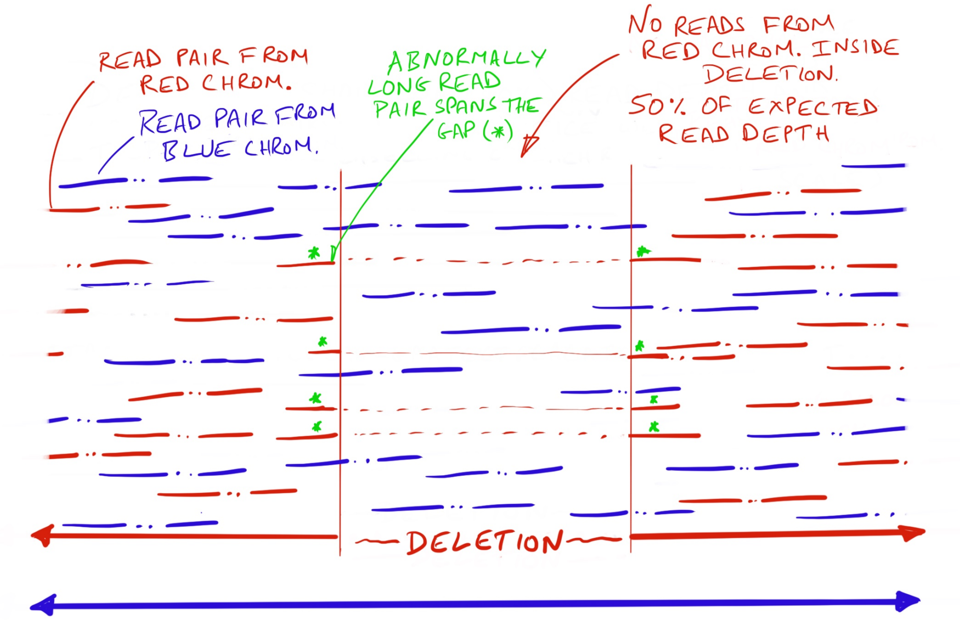
READ PAIR FROM RED CHROM.
READ PAIR FROM BLUE CHROM.
ABNORMALLY LONG READ PAIR SPANS THE GAP (*)
NO READS FROM RED CHROM. INSIDE DELETION.
50% OF EXPECTED READ DEPTH
*
DELETION

Figure 1.56: **Detection of a deletion from short-read data.** *Two homologous chromosomes (maternal and paternal copies) are shown at the bottom of the figure. The sequence read-pairs that come from each are indicated by red and blue lines above. Within the deletion the average coverage is about 50% of the average, because there are no red reads; we also see some read pairs that span across the deletion. When we map those back to the reference genome they seem extraordinarily long.* ***Note: In real-life analyses we usually do not know which homolog (the red and blue colors) each read comes from.***

However, in practice, many structural variants can be difficult, if not impossible to detect using short-read data. One important reason is that they are often associated with repetitive regions of the genome where read mapping is extremely difficult; we'll cover this further in Chapter 1.5. This is one area where the extremely long reads from 3rd-Gen sequencing are a real game changer compared to short reads. Long reads can span right across high complexity regions and make it far easier to detect the number and orientation of repeated elements and structural variants.

Low-budget approaches to studying genome variation. So far we have focused on **whole genome sequencing (WGS)** applications. But recall that only a small fraction of the genome is involved in coding for genes or controlling gene regulation. Given this, one might greatly reduce sequencing costs by sequencing only the functional regions. One approach to doing this is called **exome sequencing**. Exome (from the words *exons + genome*) sequencing uses a lab technique to preselect all the DNA fragments that span gene exons. Since exons only span about 1% of the genome, this greatly reduces the necessary amount of sequencing.

The single biggest advantage of exome sequencing is reduced cost compared to whole genome sequencing. The disadvantage is that obviously it misses all the functional variation outside exons. Later in the book we'll discuss how severe disease mutations tend to be concentrated in exons, but lots of other important types of variation are outside exons, and would be missed by exome sequencing. We may expect exome sequencing to become less relevant as sequencing costs continue to drop.

Genotyping. Last, I'll mention a completely different approach to measuring a limited subset of the genome. The term **genotyping** refers to a variety of different experimental methods that can **determine a person's genotype at a specific set of pre-selected SNP positions** (and nowhere else in the genome).

Current commercial genotyping platforms measure between 500,000 and 2 million SNPs. Genotyping provides less information than a full genome sequence–for example it cannot tell you if carry a rare mutation in a disease gene, as that mutation is unlikely to be included on the genotyping array.

While exome sequencing and genotyping are both used to get a cheaper look at a fraction of the genome, they have very different pros and cons. Exome sequencing provides complete DNA sequence for arguably the most important 1% of the genome. It is useful for identifying rare protein-coding mutations. In contrast, genotyping gives a truly genome-wide look at genetic variation but does not detect rare mutations.

However, genotyping is widely used because it can be applied to very large numbers of samples, and is accurate and relatively cheap (less than $100 per sample). If you have sent a DNA sample to a personal genetics company such as Ancestry or 23andMe, they probably did not sequence your genome, but instead used genotyping. Genotyping is also used in many large-scale research studies. At the time of writing (2022), tens of millions of people have been genotyped, either commercially or for academic research–far more than have been genome sequenced.

In summary, DNA sequencing technology has improved by more than 1 million-fold in the last 30 years. This has enabled cheap resequencing of human genomes for research and clinical applications; genome sequencing of thousands of diverse species; and widespread use of sequencing as a molecular counting tool for many applications. This continues to be a fast-moving area of innovation.

Notes and References.

[75] This phrasing is borrowed from Shendure et al (2017); that paper is a great source for history and technology of sequencing:

Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, et al. DNA sequencing at 40: past, present and future. Nature. 2017;550(7676):345-53. Another useful review is:

Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics. 2016;17(6):333-51

[76] [Link]

[77] Sanger sequencing is convenient for quick-turnaround applications in lab-work like checking that a plasmid has been constructed correctly, checking genome edits, or confirming that a PCR product contains the expected sequence.

[78] Cost of the Human Genome Project: [Link]

[79] One potential competitor is Beijing's BGI Genomics which has acquired and refined a technology called nanoball sequencing, originally from Complete Genomics.

[80] Background on Illumina technology, see eg [Link].

[81] Dohm JC, Peters P, Stralis-Pavese N, Himmelbauer H. Benchmarking of long-read correction methods. NAR Genomics and Bioinformatics. 2020;2(2):lqaa037. Note that PacBio's HiFi approach reads the same molecule multiple times, thereby lowering error rates to be competitive with Illumina.

[82] A 2022 paper considered the application of ultra-rapid genome sequencing in critical settings. They showed that it's possible to obtain extremely rapid (same-day) clinical-grade genome sequences on the Nanopore platform at a cost of about $5000 per sample.

Gorzynski JE, Goenka SD, Shafin K, Jensen TD, Fisk DG, Grove ME, et al. Ultrarapid nanopore genome sequencing in a critical care setting. New England Journal of Medicine. 2022;386(7):700-2

[83] 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

[84] Illumina has achieved near-monopoly status in the US in genome sequencing. In general monopolies lead to higher prices and lower rates of innovation in industries dominated by a single player: [Link].

[85] For one ambitious current effort in this direction see [Link].

[86] A 2018 paper estimated Illumina error rates at 0.24% per base pair

Pfeiffer F, Gröber C, Blank M, Händler K, Beyer M, Schultze JL, et al. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. Scientific Reports. 2018;8(1):1-14

[87] Teissandier A, Servant N, Barillot E, Bourc'his D. Tools and best practices for retrotransposon analysis using high-throughput sequencing data. Mobile DNA. 2019;10(1):1-12.

[88] Lee H, Schatz MC. Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. Bioinformatics. 2012;28(16):2097-105