Genomics Crash Course

Thanks to Brad Taylor and Eric Weitz for proofreading.

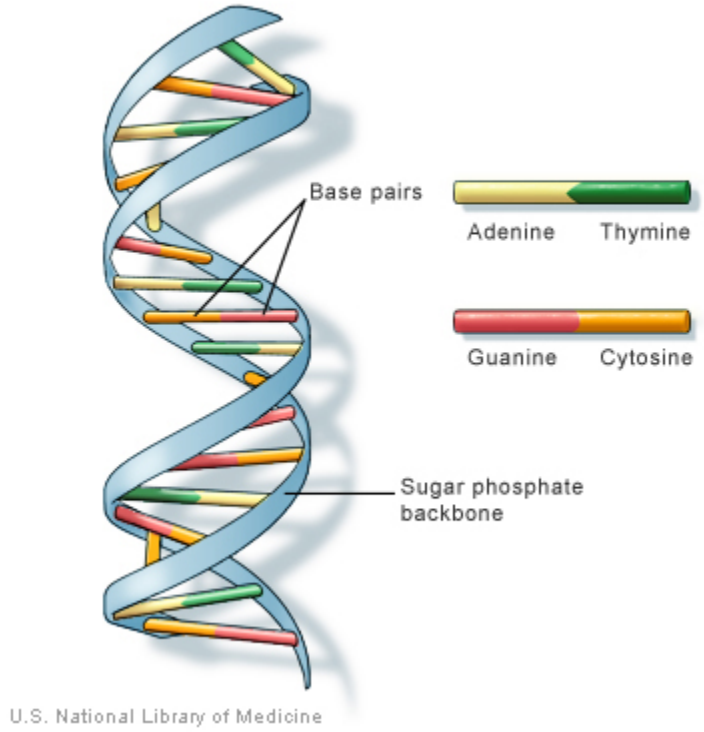
Who is this for?

This course is aimed at people with (roughly) a high school knowledge of biology, and a need to understand, at a general level, the kinds of things we do here at Broad. The original person we had in mind while writing this was a software engineer in DSP: you need to know *kinda* what’s happening in science-land, but depth and accuracy aren’t as important as breadth and knowing how to phrase the questions you’ll inevitably need to ask.

As a result this course will be prone to broad strokes, oversimplifications, and hand-waving. The goal is **not** to teach you everything you might want to know, but to give you a decent understanding of the material. Consider it a taste to get you started, and you should absolutely chase down people and ask more questions if your interest is piqued. There are some links at the end for those interested in diving deeper.

So with that caveat out of the way:

What is DNA? Why do we care?



DNA is a twisted ladder of chemicals (or a **double helix** if you want to be fancy about it). The two side rails are a repeating "backbone" structure to which the rungs attach; the rungs of the ladder are the interesting bit.

Each rung is two chemicals that have paired off. The chemicals are called amino acids, there are four of them, and they always[1] pair off as follows:

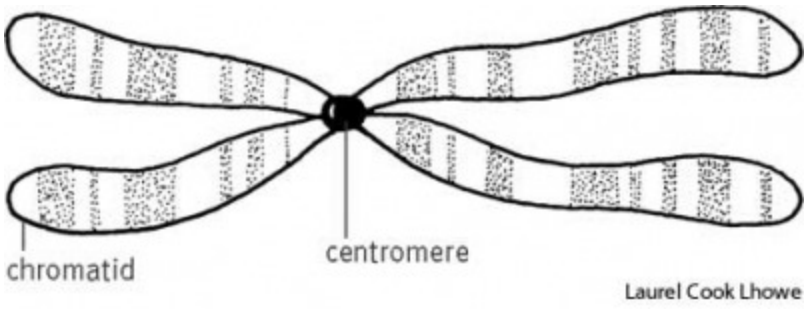
Adenine - **T**hymine
Guanine - **C**ytosine

Please now forget their names and refer to them by their initials: **A**, **C**, **G** and **T**. These are the **bases** (or **nucleotides**), and each rung is a **base pair**. The human **genome** - the entire ladder - has 3 billion base pairs. That's 3 billion rungs on the ladder.

[1] Not always, and you can get into trouble when this goes wrong. However, your cells' DNA repair mechanisms are surprisingly good at their jobs.

Chromosomes, genes, and alleles

However, it's not actually one long ladder. It's split up into 23 pairs of **chromosomes** (in humans; the number varies wildly between organisms).You have 46 chromosomes in total- 23 from your biological mom, 23 from your biological dad.



We'll get to this in more detail shortly, but for now, let's think about DNA as a store of information.Various sections of the DNA tell you things about the person. One region codes for hair colour, another for eye colour, another for the tendency to fall asleep reading long explainer documents. These regions are called **genes**; chromosomes contain many genes.

So: each chromosome is made of DNA, and the interesting bits of a chromosome are called genes. Maybe bio-mom gave you "blonde hair" in the "hair colour" gene. To be clear: the "hair colour" section - the *region itself* - is the gene. The values it can take: blonde, brown, black, blue - are called **alleles**, and they can vary.z

Because you actually have two copies of each chromosome,, you can be carrying the alleles for both brown and blonde hair - one allele from each of your biological parents.. Or maybe the two alleles you carry are identical - brown hair for both. The number of different alleles you can carry for the same gene is called **ploidy**. A single chromosome in a sperm or egg is haploid, human cells are diploid, and other organisms - and often cancer - can be triploid or even more.

So, that's what DNA looks like. But what does it do?

DNA is the instructions for building proteins

Proteins do most of the business inside your body: hemoglobin carries oxygen in your blood, keratin is your hair, glucagon tells your liver to release more sugar (aka energy) into your blood.

Structurally, they are long sequences of smaller chemicals called **amino acids**, strung together like beads on a string. The long string of proteins then folds up according to the chemical attraction forces between the amino acids, and the resulting shape is what allows the proteins to

do whatever they do in your body. For example, hemoglobin folds into a shape that grabs on to oxygen atoms when there's lots of oxygen around, and releases them when there's less of it around. Congratulations! You now have something that can move oxygen from your lungs to your muscles.

How your body builds proteins: the "central dogma" of biology

DNA hangs out in the cell nucleus. When it's time to build a protein, the process goes something like this:

Step 1: Transcription

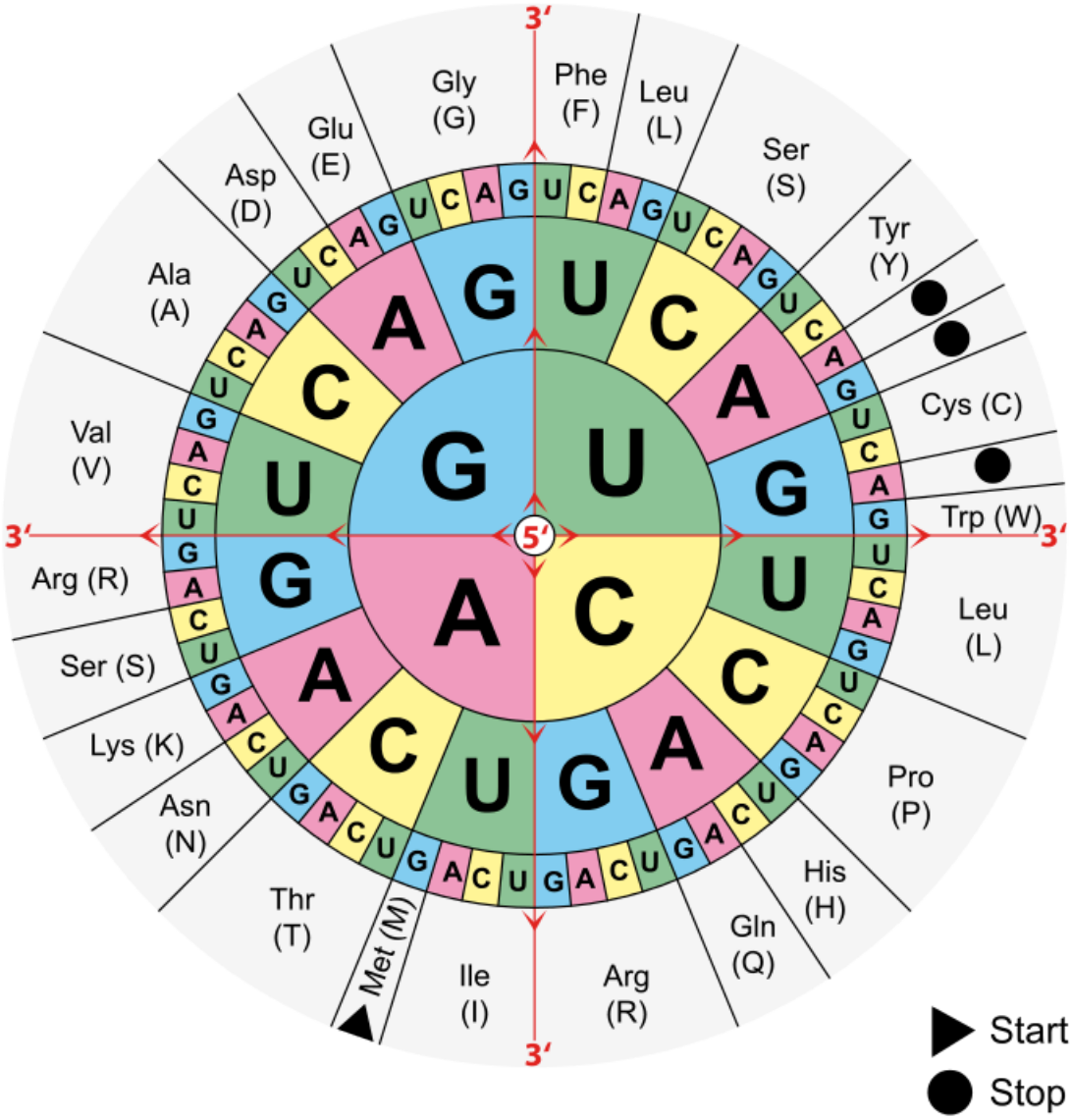
A robot called **RNA polymerase** runs down the DNA, cutting the ladder in half "vertically" down the middle. As it does so, it looks at each base, and glues together the matching bases to form another half-ladder. Once it's done, it ejects the half-ladder it created into the cell, and stitches the DNA back up again.

This half-ladder that got created in the process is called RNA. The only difference between RNA and "half a strand of DNA" is that **T**hymine is replaced by a similar chemical **U**racil in the same places. So instead of **ACGT** it's **ACGU**, but the principle is exactly the same.

Step 2: Translation

Once the RNA is floating around in the cell, a different robot called a **ribosome** is responsible for actually building the protein. It starts at one end of the RNA and reads three bases at a time. Each set of three bases is called a **codon**, and it's an instruction for the ribosome: "start gluing amino acids together", "amino acid X is the next one in the sequence", or "the protein is done now".

It is literally a lookup table:



So if the ribosome reads AGU, it glues on an amino acid called serine, and if it reads UAG, it stops, and the protein is complete. The sequence AUG both codes for the amino acid methionine but also means “start making a protein” if the ribosome is yet to be activated. So anything between the “start” codon and the “stop” codon potentially codes for a protein and thus does something. This part of your genome is called the **exome**.

Putting it all together

So: for each chromosome, you have two genes -- two distinct pieces of DNA -- that code for the same functionality in your body. So you might carry both the “blonde hair” allele and thus create blonde hair-protein, plus the “brown hair” allele in the same place, and *also* create brown

hair-protein. What happens when you have different alleles varies between genes: sometimes one allele is **dominant**, and will prevent the other (**recessive**) one from being transcribed and having proteins made from it. Other times they'll each happily do their own thing.

The vast majority of the variation in your DNA (relative to some imaginary "human average" DNA sequence) comes from your parents. There's also the fact that cell division isn't completely accurate, and sometimes cells will make a mistake when they duplicate their DNA; also, environmental factors, including smoking and radiation, can physically knock out bases in your DNA and replace them with different ones. The DNA repair mechanisms in your body are extremely good at fixing errors in both of these cases, but some do slip through.

The important thing to remember is this: **what happens in your DNA directly affects how your body works**. If some DNA in a cell has a mutation -- regardless of whether you got it from your parents or from a radioactive spider -- that DNA will tell the ribosome to do something different, potentially affecting the resulting protein massively. Consider a mutation UAC → UAG: instead of putting a tyrosine next on the protein chain, the ribosome will *stop producing the protein altogether*, leaving it truncated.

As another example, consider hemoglobin. It carries oxygen does this by being the right shape to accept four oxygen atoms, and it has the right chemical properties to attract and release those atoms based on the ratios of carbon dioxide and oxygen around it. Sickle cell anemia is a single-base change in the code for the hemoglobin protein, from GAG → GTG, replacing in the amino acid glutamic acid with valine. This affects how the protein folds, and the resulting hemoglobin proteins now clump together, forming long rods inside the red blood cells, instead of staying isolated and smooth. These long rods affect the shape of the red blood cells that carry them, making them stiffer, less able to move around the body more easily, and more likely to interrupt blood flow. All because of a single base change.

One final thing to keep you up at night: what happens if you add or remove a single base? Then when the ribosome tries to read bases in groups of threes, the groups are all off by one, and now the ribosome is off following a completely different set of instructions. This is known as a **frameshift error**, and it's predictably bad news if it survives -- though it's highly likely that your cell repair mechanisms will notice a mistake of this magnitude and trigger **apoptosis**, aka cell death.

So that's why DNA is scientifically interesting: because it tells your body how to make the things it uses to stay alive.

Sequencing, alignment, and variant calling

Now we know what DNA is and why we're interested in it, let's talk about how we, as scientists, go from "DNA is inside you" to "I can look at this on my computer". The process of going from goo in a tube to a computer file is called **sequencing**.

Sequencing the entire genome is referred to as **Whole Genome Sequencing**, or **WGS**, and it's expensive. One way to reduce the cost, if it's appropriate for the research, is to only sequence the exome -- the parts of your DNA that code for proteins. This is **Whole Exome Sequencing**, or **WES**. Before chopping up the DNA into fragments, the DNA is washed with enzymes which bind to certain sites on the genome and chop them out, leaving only the exome behind. If you're only interested in certain areas of the genome, you can go a step further and do **targeted exome sequencing**, chopping out all but a very small section of the genome. As you progress along this spectrum you gain more coverage for less money, and you save disk space in the process.

The sequencing process

The ability to sequence DNA was invented in 1977 by Frederick Sanger. The original technology was relatively laborious, but the same basic process was used for the human genome project in the late '90s. Innovation in the mid 2000's made sequencing much faster and cheaper, leading to a data revolution in genomics.

Here we'll (very roughly, with much handwaving) outline what we call "next generation" sequencing, which started around 2007. One thing you should keep in mind going in: sequencers are imprecise machines. They make mistakes, and incorrectly identify a base around 1 in 1000 times. This means that over a single human genome of 3 billion bases, there are 3 million errors!

Prior to sequencing

The sequencing process starts with obtaining the genetic material. There are a number of ways you can do this: blood draw, cheek swab, biopsy, etc. Then you extract the DNA from the sample -- you might have done this with strawberries as a science project.

Because the accuracy of a sequencing machine goes down the more bases a sequencer reads from the same strand, the next step in the process is to chop up the DNA into small fragments. The resulting fragments of DNA end up being around 300 base pairs long.

Depending on the experiment, the sequencing protocol may now call for a PCR step. PCR stands for **polymerase chain reaction**: this process makes many copies of each fragment of

DNA. This helps offset the errors the sequencer makes: it's extremely unlikely the sequencer will make the same mistake at the same bases for multiple copies of the same fragment, so at the end you can collect them together and cross-compare for more accurate results. This is less common as a pre-sequencing step nowadays, as it introduces biases that affect downstream analysis, and you can get a similar affect by sequencing more sample instead of replicating the fragments.

PCR or no, one last thing happens to the fragments before sequencing: they get **adapters** attached to each end. There are two kinds of adapter molecules, and each one attaches to one end of a DNA fragment. The adapters have specific chemistry that matches the internals of the sequencer, so the sequencer can preferentially "attract" a particular end of the fragment in order to make it stand up.

Inside the sequencer

A **sequencer** is a machine that takes fragments of DNA and reads the bases off them. Imagine a square grid of hundreds of tiny, DNA-fragment sized test tubes (called a **well plate**). Each DNA fragment slides into one well, suspended vertically. The sequencer then attaches an enzyme at to the adapter at the top end of the fragment, and washes over four chemicals (called **buffer solution**), one for each base.

The chemicals in the buffer solution have two components: one that binds to the next base in the fragment, and the other is a molecule that will fluoresce with a colour corresponding to the base when exposed to a certain light source. This fluorescent molecule also prevents another chemical binding to the next base in the template.

So the sequencer washes over the buffer solution for A, which will attach to all the fragments whose next base is A. Then it washes over the solutions for C, G, and T. Once this has happened, each fragment will have one matching chemical attached to it. The light source is applied to make the fluorescent molecules light up, and the sequencer takes a picture. By looking at the colours of the lights in the picture, you know the next base in the sequence for all of the fragments. The sequencer repeats this process a number of times, building up a sequence of bases for each fragment, known as a **read**.

After 100 or so bases are read off the fragment, it's time for the sequencer to do its next trick: flipping the fragment over. The adapters at each end of the fragment are designed to have a very strong covalent bond with a chemical the sequencer can generate, so the sequencer does so and attracts the other end of the fragment to the bottom of the well, bending the fragment into an upside down U shape. The first adapter, and now the fragment is standing on its other end. The sequencer now repeats the process to get a second read from the other end of the fragment -- the **mate** of the first read.

Here's a great video by Illumina walking through the process.

Sidebar: Ways the sequencer can make mistakes

- The fluorescence for one well might bleed over into an adjacent occupied well in the picture, making two pictures light for the same well plate.
- The fluorescence for one well might bleed over into an adjacent empty well in all the pictures, creating an identical entire read called an **optical duplicate**.

What a sequencer produces

The output of a sequencer is typically a FASTQ file, which records all the reads the sequencer produced. Each FASTQ record contains:

- The name of the record, which is used for matching reads with their mates
- The sequence of bases the sequencer read in that read
- A quality score for each base in the sequence

The quality score is the sequencer saying how confident it was when it made that read. Downstream tools use it when trying to determine whether a surprising base is a real mutation present in the original DNA, or just the sequencer making a mistake.

Alignment

So, to recap, we've taken some DNA, chopped it up into fragment of ~300 bases, and then read around 100 bases from each end, working towards the middle. The two reads combined are called a **mate pair**, and the gap of unread bases between them is called the **insert size**.

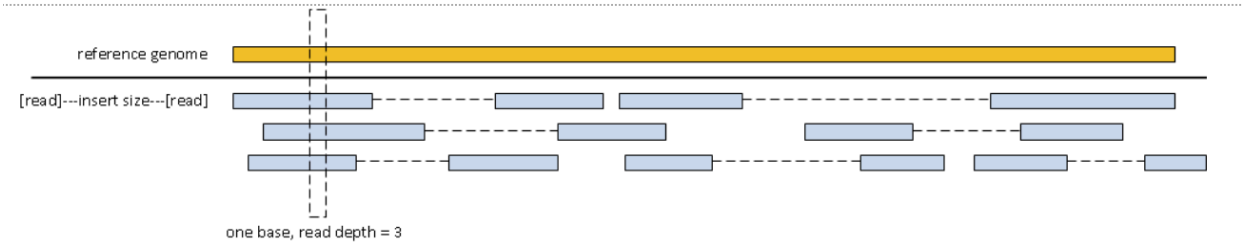
You'll notice from the numbers that the vast majority of the time, the two reads constituting any given mate pair don't cover the entire fragment. All we know is that they are from opposite ends of a fragment and they go together - we don't even know which end is which, and we *certainly* have no idea where the fragment used to live in the original piece of DNA we chopped up into little pieces. **Alignment** helps us answer all those questions.

A prerequisite for alignment is having a reference genome. This is a fully sequenced genome, with all the bases in order; the standard human reference genome is a "mosaic" of the DNA of 13 anonymous volunteers who all lived in Buffalo, NY. As such it is both not an individual person's genome and also doesn't contain genetic data for all the possible alleles any gene might have.

Nonetheless, it is the responsibility of a computer program called an **aligner** to take the FASTQ file and determine from each read where in the human genome it's located. It does this by attempting to find the best positional match between the read and the reference, accounting for

potential errors in sequencing (this is where quality scores come in handy) and real genetic variation between people. If the match is good enough, the read is given a location.

The aligner fails to match around 1% of reads; these are put into a pile labelled "I don't know where this goes".



Since multiple cells are sequenced at once, and their DNA is fragmented at different locations, the reads can be "stacked" on top of each other once they've been aligned to the reference. The number of reads at a particular base is known as the **read depth**, which is usually averaged across all bases to give an indication of the overall **coverage**. Since the sequencer makes mistakes, more coverage is better.

The aligner can also detect **duplicate reads**, which may come from the PCR process, or be artifacts of sequencing (e.g. the optical duplicates mentioned before).

The most commonly used aligner at Broad is one of the BWA (using the Burrows-Wheeler alignment algorithm) family, often BWA-mem which contains optimisations for high-memory machines. Other aligners you might hear about are Bowtie, and STAR for RNA-seq projects (more on that later).

Sidebar: before alignment, we had assembly

The process of alignment relies on a reference genome. But how do you put together a reference genome when you don't have a reference genome to align to? The answer is by doing **assembly**, and it sucks. The algorithm is so simple, and it will make you cry:

1. Calculate pairwise alignments of all fragments.
2. Choose two fragments with the largest overlap.
3. Merge chosen fragments.
4. Repeat step 2 and 3 until only one fragment is left.

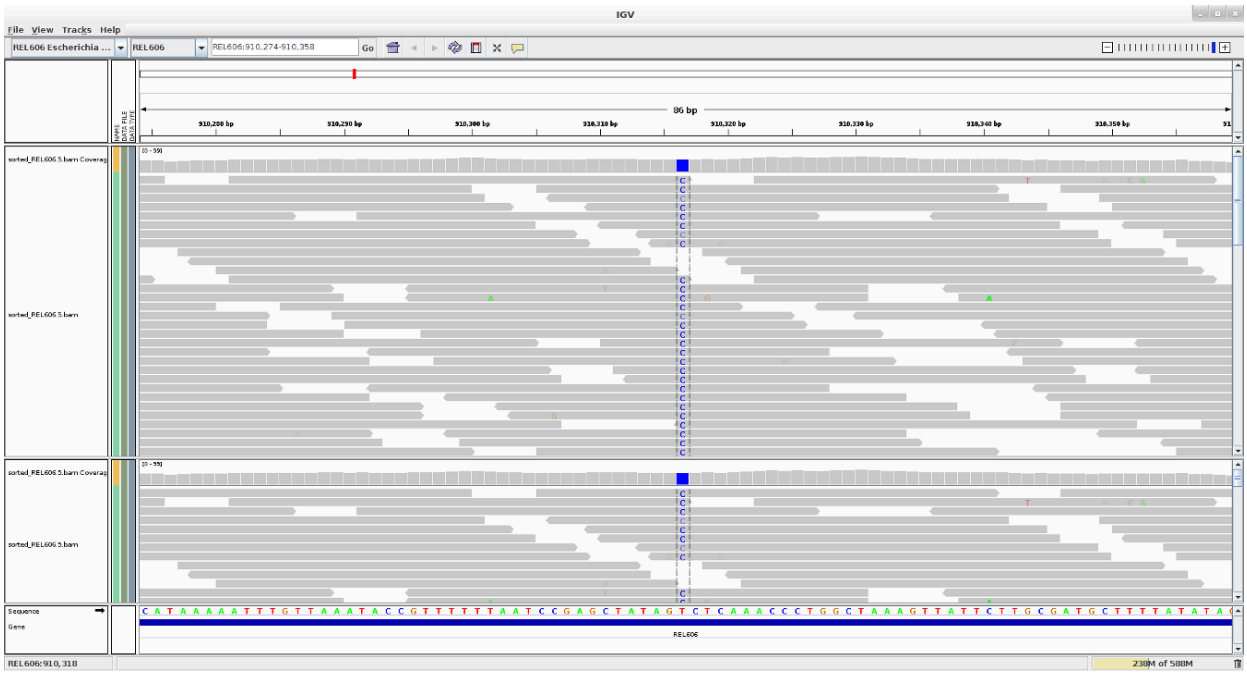
Yup. It's $O(n^2)$ over the entire genome.

Aligners produce BAM files

In most cases, an aligner produces a .bam file. This stands for Binary Alignment Map. A single bam record contains, amongst other things:

- The name of the read
- The start position of the read, aligned on the reference genome
- A quality score produced by the aligner, indicating how confident the aligner is that the read does in fact go here on the genome
- The sequence of bases in the read
- A pointer to the read's mate pair in the bam file

IGV is the standard visualiser for bam files. It looks like this:



At the very top of the screen you see a map of where the visualiser is focused on the human genome. Underneath it are two bam files. Each horizontal grey bar is one read. You can see how the sequencing process generated multiple reads that all "stack up", providing multiple independent measurements of the base at this location.

Right at the bottom is the sequence of bases in the reference genome. The base highlighted in the sample seems to be different to the reference genome (a C rather than a T). Because it's consistent across all of the reads in the sample it's most likely a real variant rather than a sequencing error.

Variant calling

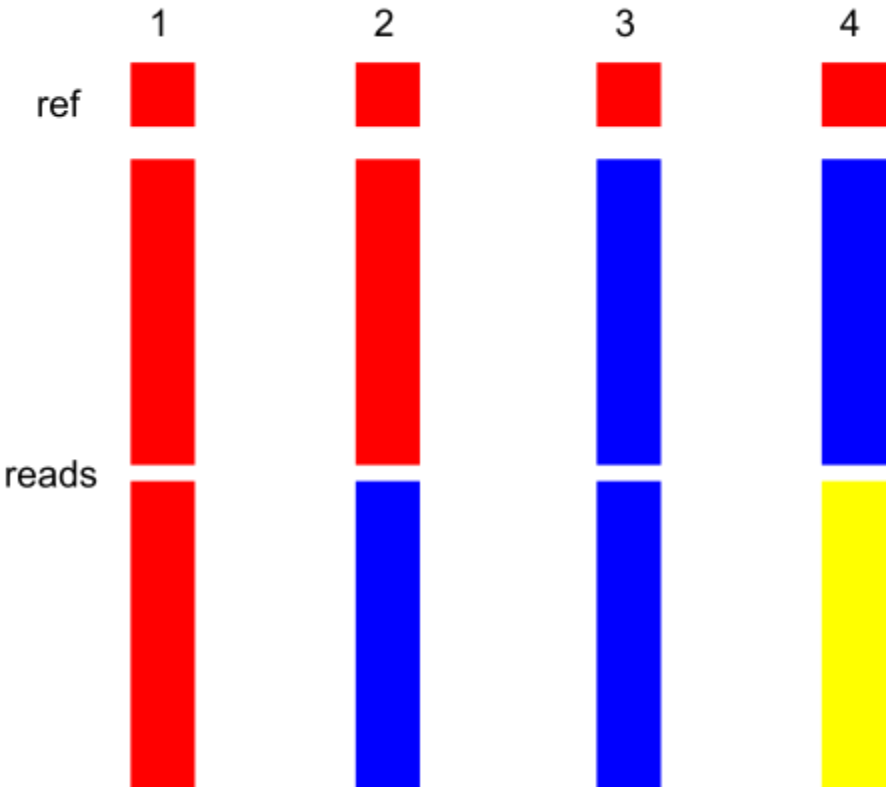
By now we've got pretty far: we've taken a sample from a person and turned in into a file that's a map of their genome. Now we're getting to answering a scientific question: what's interesting about this sample? We might want to compare it to the reference genome, to see if the person has any hereditary mutations that might predispose them to certain diseases. But sequencers

make mistakes. Given some difference between our sample and (say) the reference genome, is that difference real -- i.e. the difference is truly present in their DNA -- or is it a sequencer error?

This is **variant calling** -- deciding, for every base in a sample, if it's the same or different to whatever you're comparing it to, and how.

Recall that human DNA has two chromosomes, each of which might hold a different allele in any given location on the genome. For a moment, let's imagine a world where we're using a sequencer that never makes a mistake. Given that you have the same amount of genomic information in your DNA from each allele, then compared to the reference at any given location, there are only four possibilities:

1. Both of your alleles are the same as the reference genome. 100% of the reads intersecting the vertical line in the IGV screenshot will match the reference genome's base. A variant caller would call this location **homozygous ref**: both alleles are the same as each other (homo), and are equal to the ref<u>ref</u>erence.
2. One of your alleles is the same as the reference genome, but the other is different. So 50% of the reads will match the reference genome, and 50% of the reads will differ, but they'll all differ the same way. This is **heterozygous ref**: the alleles are different (hetero) to each other, and one of them matches the ref.
3. Neither of your alleles is the same as the reference genome, but they do match each other. 0% of the reads will match the reference genome, but they'll all agree with each other. This is **homozygous alt**: the alleles are the same as each other, but are an <u>alt</u>ernate base to the ref.
4. Neither of your alleles is the same as the reference genome, and they don't match each other either. So 50% of your reads will be one base, and 50% will be a different one. This is **heterozygous alt**, as you've probably guessed :)



Given that the percent of the reads matching the reference can *only* ever be 0%, 50%, or 100%, you now know that any deviation from these ratios *must* be because of a sequencing error. Simply looking at the ratios you do have and picking the closest option is really all you need to do; with even a modest number of reads covering each location, the chance that the sequencer will be wrong for enough reads at that exact location to throw off the variant call is extremely low.

This is pretty much the variant calling algorithm that population geneticists use. Cancer geneticists have a much more difficult time of it, as cancer mutates quickly, and you're no longer guaranteed that your sample will only contain at most two different bases at any location, nor are you guaranteed the 0/50/100% ratio. We'll go into this in more detail in the cancer section.

The output of variant calling is usually a **VCF** (Variant Call File). Since most of your sample will exactly match the reference, it saves space by only writing out the locations where there are variants. You can variant call multiple samples at the same time, and get a single VCF containing the variants for all of them. Each VCF record contains:

- The location (chromosome and position) of the variant
- Space for an identifier, if this is a known variant in one of the many variant databases
- Which base the reference genome has at this position
- What the samples that got variant called look like at this position

- The variant caller's confidence that this variant is real

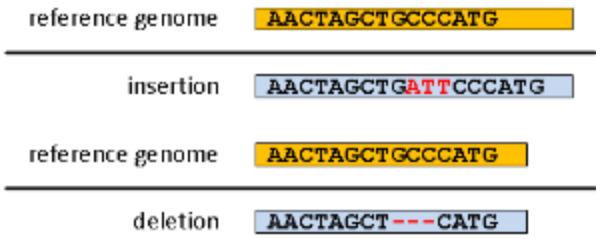
Types of variant

Changing only a single base, as described above, is not the only kind of variant that we might see, but it is one of them.

Single Nucleotide Polymorphisms (SNPs)

A **SNP** (you will sometimes see **SNV**, for Single Nucleotide **Variant**) is a substitution of one base for another: for instance, a C changing to an A. Because bases make up codons and codons form instructions, a change in base may mean the ribosome does something it shouldn't when it hits the site of the SNP - for example, using a different amino acid to build the protein or stopping early.

Indels

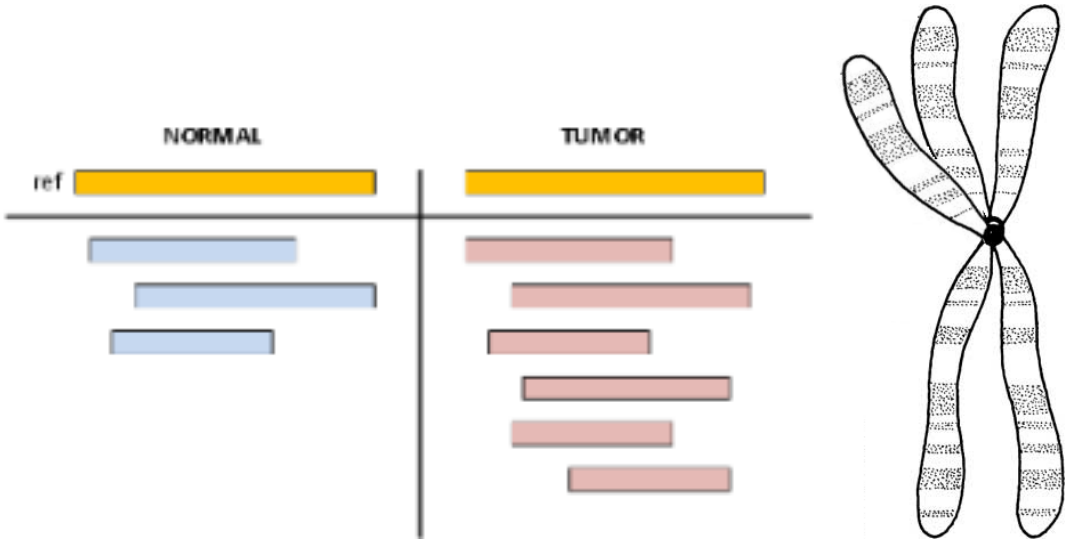


A portmanteau of **insertions** and **deletions**, these cover the case where some bases are added or removed relative to the reference, rather than just "flipped". These may cause frameshift errors, or if the indel is a multiple of three, it'll "only" change the structure of the resulting protein -- potentially in a way that causes it to fold and thus function differently.

Rearrangements

When two (or more) parts of the genome sequence swap places. In reality, when a rearrangement happens, significant chunks of the genome get swapped - not just a few bases as shown in the picture.

Copy number variations



In places, the genome contains sequences which repeat. An aligner can't place a repeat fragment in its correct place amongst all the repeats, so they just get stacked underneath the first one. This means that repeat sequences show up as regions with unusually increased coverage. Copy number variations can therefore be detected by comparing the difference in coverage between the normal and tumor.

Copy number variation is a symptom of a number of possible underlying mechanisms. As discussed, it's possible for a repeat to get over- (or under-) repeated; but it's also possible for the chromosome to be mutated such that it gains or loses an arm. An extra arm will align to the reference in the same place and show up as increased depth, just like a repeat.

Cancer

How is cancer formed?

The cells in your body grow, die, and get replaced. Cells grow by doubling up all the stuff inside them and then dividing into two. "All the stuff" includes the DNA too. DNA replication is easy: a robot runs down the ladder, cutting all the rungs in half. Because each base only binds to one other, you only need one half of the ladder to regenerate the other half. So the two halves go their separate ways and the corresponding other halves of the rungs are attached by more robots using new bases your body has made for precisely this purpose.

However, biological robots aren't perfect - they make a variety of mistakes when they're copying DNA. We'll get to the types of mistakes they make later, but for now we'll focus on what happens

when a mistake is made. Your cells have two mechanisms to deal with badly copied DNA: first, they'll try to detect that a mistake has been made and fix it. If it works, hooray! No more mistake. If it doesn't, the cell will try to hit a self-destruct button, and kill itself. This is called **apoptosis**. Both of these happen all the time, and they're very effective at making sure that all the cells in your body have identical copies of your DNA.

But it doesn't *always* work. Sometimes a DNA replication error won't get fixed, and then the self-destruct button won't go off. This isn't the end of the world; very often the error will be in a region of the genome that doesn't do anything. Or the error won't actually make a difference to the final protein being created (this is known as a **silent mutation**). Or the opposite - the error will be so critical that the cell will die anyway because it's just so broken.

Unfortunately, not all errors are harmless or immediately fatal to the cell. What if the error is in the part of the gene that determines when it's time for the cell to divide - and it decides that the answer to that question is *always, right now*? Or what if the error is in the part of the gene that determines when it's time for the cell to die, and it decides the answer is *never*?

Then you've got a cell that goes wild and wants to replicate as fast as possible, as soon as possible. And it refuses to die.

That's cancer.

What happens next?

So we have a ball of cells that are replicating way too fast, crowding out the other, well-behaved cells around them. Let's start a list of our problems:

- They're going to compete with the well-behaved cells for space and resources. The well-behaved cells are going to have fewer nutrients, and may starve.
- Other replication errors in the cancer cells may cause them to do their job less well than the well-behaved cells. So not only have you got well-behaved cells starving to death, their replacements aren't even picking up the slack from their dead compatriots.

Unfortunately, this is only the beginning. Faster cell replication means that there is more opportunity for *more* DNA replication errors to happen. And they're less likely to be fixed, because the original cancer cells have had their fixing-ability compromised - that's how they ended up being cancer cells in the first place. So:

- Our cancer cells are now also picking up new mutations.

Remember how we said that the DNA repair mechanisms ensure that all your cells have the same DNA? They're all clones; this is called **clonality**. In cancer, your normal, healthy cell has spawned the initial "Patient Zero" cancer cell, which has compromised DNA repair and is now spawning mutants of its own. These mutants may be *further* mutated, and you're left with a big ugly tree of multiply mutated cancer cells. This "family tree" of cancer cells is called **subclonality** - the cancer cells are different *even to each other*.

When Cancer Goes Bad

Our initial big ball of cancer is known as the **primary** tumor. It's also **benign** - it's stuck growing locally. At this point judicious application of a scalpel can usually remove the tumor in its entirety.

The strict definition of cancer doesn't include benign tumors - only malignant ones. A **malignant** cancer is one that has found a means of spreading elsewhere in the body - often through the circulatory or lymph systems. Imagine a few cancer cells detaching from the primary tumor and then travelling through the bloodstream to make a home in a new location. This second, new tumor is a **metastatic** tumor, a **met** for short. It has the same cell type as the its primary - it's just set up a second franchise elsewhere in the body. So it's perfectly possible (and unfortunately common) to have lung cancer cells turning up in bones or the brain once the cancer has gone malignant.

Addendum: cancer sequencing particulars

Most of the analysis in CGA is done on **tumor / normal pairs**. Two chunks of flesh (or blood) are taken from the cancer patient - one is from a cancerous tumor, the other from normal tissue unaffected by cancer. These form a pair: case (the tumor) and control (the normal). The normal tissue is helpful because it provides a benchmark to compare the tumor sample against.

Unfortunately, the tumor is somewhat more complicated a case. When taking a tumor biopsy, you cannot be sure that the only cells you have taken are cancerous - there could be some non-cancer normal cells floating around in there too. Even in a hypothetical case where you were able to guarantee that 100% of the cells in the biopsy were tumor, subclonality would still mean that not all of the cells would have the same DNA. The constitution of a sample in terms of normal, tumor, and subclonal tumor cells is called the sample's **purity**.

A sample's purity can in theory be vastly improved by isolating a couple of individual cells from the sample and creating a **cell line** from them - essentially, setting them apart and waiting for them to divide. Whether you get cells with "Patient Zero" DNA or instead the further mutated cells is entirely down to luck, though - and meanwhile you don't have a representative sample of the entire cancer's subclonal "family tree". The choice of whether to grow one or more cell lines from a sample therefore depends on the study being done.