**Fact sheet**

**Genomics - Decoding DNA**

**Background on DNA function and Genomics [6-10]**

*Learn more about diseases*

Changes in our genes, often through mutations, can cause various diseases. Sometimes only one gene is involved, sometimes multiple mutated genes interact to cause a disease.

*Be able to identify people*

Each person has a unique genetic sequence, which can be used to identify people and their relatives, as this information is passed onto the next generation.

*Learn how genes control our bodies*

The order of bases in the genetic code determines the sequence of amino acids in a protein. This then determines the structure and therefore function of the proteins encoded by your genes. Depending on your genes, different proteins may be produced which affects how our bodies look and work, and how we change as we age.

*To know if we are at risk of disease*

People may go through the process of genetic counselling to learn more about their individual risk of disease. This is often when there is a family history of genetic disorders.

*To tailor treatments to individual patients*

We are now able to tailor treatment plans to individual patients by sequencing their genome first. This gives us information about their genes, which we can use to recommend which treatments they might respond better too, and which treatments might not work.

The genome is the complete set of genes or genetic material present in a cell or organism. Genomics describes the study of all of a person's genes (i.e. the entire genome), whereas Genetics is the study of genes and how 'traits' are passed down from one generation to the next​. Only about 1% of the genome codes for proteins. We used to think the remaining 99% was ‘junk’ DNA but we now know it plays a role in the regulation of gene expression, although there are still areas with as yet unknown functions!

Sequencing can tell us the exact sequence of nucleotide bases in a sample of DNA, telling us lots of information about an organism. Sanger sequencing, invented in 1977 by Frederick Sanger, was the first method developed to determine the exact DNA code of an organism.

Basics of Sanger sequencing:

1. Sample is divided into four sequencing reactions containing:
	1. Normal nucleotides
	2. dideoxynucleotides (ddNTPs) which have a similar structure to nucleotides, but they cannot form polymers, they are also radioactively labelled. Each sequencing reaction will be for one of the four ddNTPs (ddATP, ddGTP, ddCTP, ddTTP)
	3. DNA sample annealed with primers
	4. DNA polymerase
2. During the sequencing reactions when DNA polymerase inserts a ddNTPs (which occurs by chance) the sequence terminates as the ddNTPs cannot attach another nucleotide
3. Complementary DNA chains of different lengths will be produced
4. After the reaction, complementary DNA is removed from template DNA by heat denaturing.
5. The single stranded complementary DNA is then separated according to its size using gel electrophoresis.
6. X-ray film is exposed to the gel, the radioactively labelled ddNTPs will produce a dark band and the sequence can be read by comparing all four sequencing reactions.

Sanger sequencing is highly accurate, however it is very time consuming! It took 13 years to finish The Human Genome Project, completed in 2003, and many collaborators working together.

Thanks to Next Generation Sequencing (NGS) we are able to sequence a human genome in around day. NGS uses Polymerase Chain Reaction (PCR) to amplify DNA which is much quicker and more cost effective than Sanger sequencing, where a DNA sample is cloned into a plasmid then amplified in bacteria. During NGS the sample genome is broken into smaller fragments. Shorter fragments are quicker to sequence but need to be pieced together again using bioinformatics. NGS is able to analyse millions of DNA fragments in parallel, making it a high-throughput technology.

Since the invention of Sanger sequencing new sequencing technologies have been developed. Many have improved Sanger sequencing to make it more efficient or cost effective, however many are now based on very different chemistry. New high-throughput sequencing methods include:

* Sequencing by synthesis (Illumina)
* Dye sequencing (Illumina)
* Pyrosequencing (454 Life Sciences)
* Single-molecule real-time sequencing (Pacific Biosciences)
* Nanopore technology sequencing
* Sequencing by ligation (SOLiD sequencing)
* Combinatorial probe anchor synthesis (cPAS- BGI/MGI)
* Ion semiconductor (Ion Torrent sequencing)

Basics steps of modern sequencing:

1. Prepare sample and extract DNA. The sample needs to be of high quality and free from contaminations, otherwise the contaminated DNA will be sequenced instead.
2. Break genome into smaller fragments. This makes sequencing much quicker as the reactions can be shorter and many can happen at once.
3. Amplify fragments to create millions of sequence reads. This improves the likelihood of the fragment being sequenced and reduces the error rate.
4. Sequence all of the reads you have created using a sequencing machine. The machine will create a complimentary strand to your sequence reads, measuring which nucleotides are used in the process.
5. Align sequence reads onto a reference genome. All the information gathered by the sequencer is fed into bioinformatics software, which will reconstruct the genome using a reference genome. A reference genome is an accepted genome sequence which is used as a standard to compare other sequences of the same species to.

**Background on sequencing technologies at Babraham Institute [slides 11-12]**

NextSeq and MiSeq are machines which will take a library (a DNA sample that has been prepared) and sequence the DNA present. Both NextSeq and MiSeq machines use sequencing by synthesis (SBS) chemistry. SBS is similar to Sanger sequencing in that they are creating a new chain of DNA using a pool of labelled nucleotides. As each fluorescently labelled dNTP is added to the single stranded DNA fragment an image is taken revealing which base has been incorporated.

NextSeq 500 System is often used for Whole-Genome Sequencing​, this looks at all the protein-coding and non-coding parts of the DNA. It can generate 75-300 bp reads​, the length of the fragments that are sequenced. This can take anywhere between 12-30 hrs​ depending on the specific experiment.

Generating 20-120GB of data​

MiSeq System​ is used for a more targeted approach in DNA sequencing. It looks at specific regions in the DNA, generating 50-600 bp reads​. This can take between 4-56 hrs​ generating 0.5-13.5GB of data​.

10x Genomics Chromium controller can label single cells individually, so you are able to tell them apart during analysis. This is useful if you want to run multiple samples in a single experiment, saving time and money.

The Hamilton GNS star is an automated liquid handling robot making it easier to process large amounts of samples. This is useful for high throughput experiments.

**Activity: Decoding DNA [slide 13]**

OBJECT OF THE GAME

The object of DNA mastermind is to guess a secret code consisting of a series of 4 DNA bases. Each guest results in feedback narrowing down the possibilities of the code. The winner is the player who solves his opponent's secret code with fewer guesses.

SETUP

Decide which player is to be the Code Maker and which to be the Code Breaker. Position the board between the 2 players so that the 4 spaces for the secret DNA code face the Code Maker. Cut out and glue the shield together, use this to cover the secret DNA code.

GAME PLAY

1. One player, known as the Code Maker, secretly writes the DNA code in the 4 spaces, which are then covered by with the shield to conceal them from their opponent's sight. The Code Maker can use any combination of the 4 bases they choose (A, T, C, G). They can also use 2 or more of the same base if they wish.
2. The other player, known as the Code Breaker, sits opposite the Code Maker and writes DNA bases in the 1st row of the code guesses (closest to them). The Code Breaker is attempting to duplicate the exact bases and positions of the secret DNA code.
3. The Code Maker responds by writing ‘X’ or ‘O’ in the quadrant on the 1st row as follows:
	1. An ‘X’ to indicate a guess of the right base and in the right position (without indication of which base it corresponds to).
	2. An ‘O’ to indicate a guess of the right base but in the wrong position.
	3. Leave blank, to indicate a wrong base that does not appear in the secret code.

.

There is nothing about the order of the ‘X’s or ‘O’s to indicate which particular bases are referred to, one ‘X’ or ‘O’ corresponds to one base.

1. Based on the ‘X’s or ‘O’s the Code Breaker writes another set of DNA bases in the 2nd row and the Code Maker writes ‘X’ or ‘O’ the quadrant in the 2nd row as appropriate.
2. The Code Breaker keeps writing rows of DNA bases and getting feedback from the Code Maker until they guess the code exactly. At this point the Code Maker reveals their secret DNA code.
3. If all 10 guesses are used and the code has not been broken, the game is over and the Code Maker is awarded 11 points (10 points + 1 bonus point). The players then switch roles.
4. Provided they have given correct information, the Code Maker gets 1 point for each guess played by the Code Breaker. The players then switch roles for the next game.
5. Play continues for a series of games, the number being determined in advance. The winner is the player with the most points in the series.