YOUSEQ

A BEGINNER'S GUIDE TO NEXT GENERATION SEQUENCING

youseq.com

Next Generation Sequencing

Let's keep things simple. The world of Next Generation Sequencing (NGS) can seem complex and intimidating. It need not be. Let's start by reminding ourselves what its useful for and why we use it.

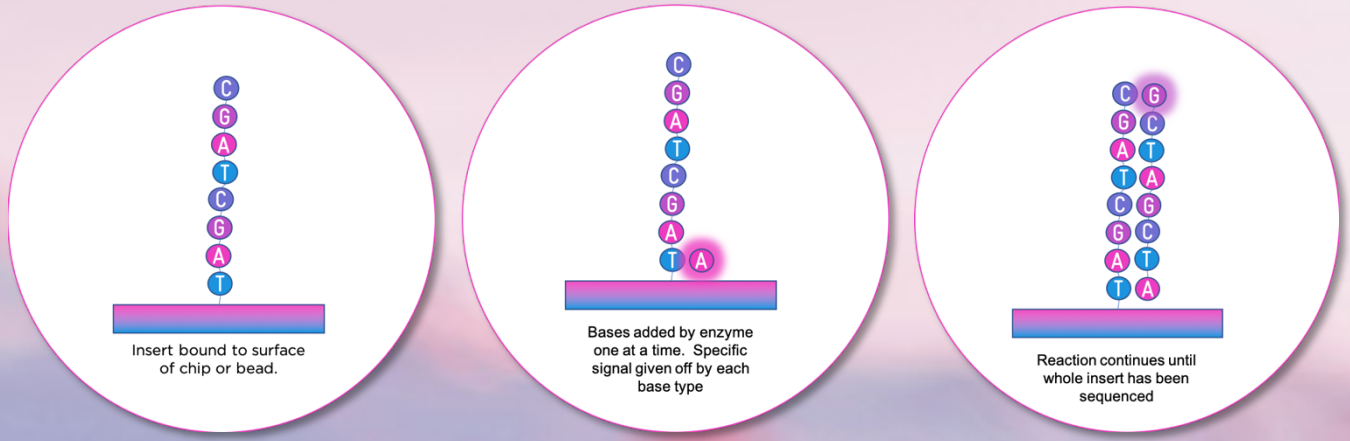
All of life is coded in it's DNA. A remarkably simple code of four molecules that act as a blue print to define the proteins that we and all of the organisms we share our planet with are made of.
Reading this code is one of the most astonishing achievements that the human species has ever and will ever accomplish. Reading this code helps us to understand how we are made, how we are all related, how errors or mutations in our DNA cause disease and how we may respond best to medicines. It holds the promise to revolutionise healthcare and has already begun to do so.
The first human genome "read" was competed in 2001. It took 10 years and the best part of $2.7bn. It was achieved by DNA sequencing. A method by which the sequence of the DNA is read painstakingly in small fragments and then reassembled to create a complete sequence. Sanger Sequencing as it is known, was the method used to achieve the first publication of the first human genome.
Next Generation Sequencing is a phrase used to describe a range of technologies that speed up and reduce the cost of DNA sequencing vs the traditional Sanger sequencing.

What are all these different Next Generation Sequencing Technologies?

Well there are quite a few of them. They are sometimes known by their scientific name or by a brand name given by the company that invented or owns them.

The most widely used of these technologies are owned by Illumina Inc and ThermoFisher Inc respectively and it is these two technologies that most of the scientific community refer to as Next Generation Sequencing or NGS.

Both Illumina's SBS and Thermofisher's IonTorrent work in similar ways. The final detection method differs but they both centre around "Sequencing by Synthesis".



n.b. This is a deliberate simplification. Actual details of reaction mechanism vary according to the specific instrument you are using. By they all center around these principles of "sequencing by synthesis" . You can learn much more about the details of NGS chemistries in the YouSeq learning zone: https://youseq.com/learning

Essentially you take your DNA sample and fragment it in to small pieces. The two strands are separated, and a controlled chemical reaction occurs to add the complementary DNA bases back, one at a time, to the single strand to create a double strand. A chemical or fluorescent signal is given off as each base is added. These signals effectively "read out loud" the sequence of the DNA therefore.

These chemistries are “short read” technologies. By this we mean that only short fragments (typically <600bp) of DNA can be read. So to sequence a larger piece of DNA you read lots of short fragments and then the sequences are “Stitched back together” by computer afterwards to create the complete sequence.

Is NGS accurate? Can I trust the sequence?

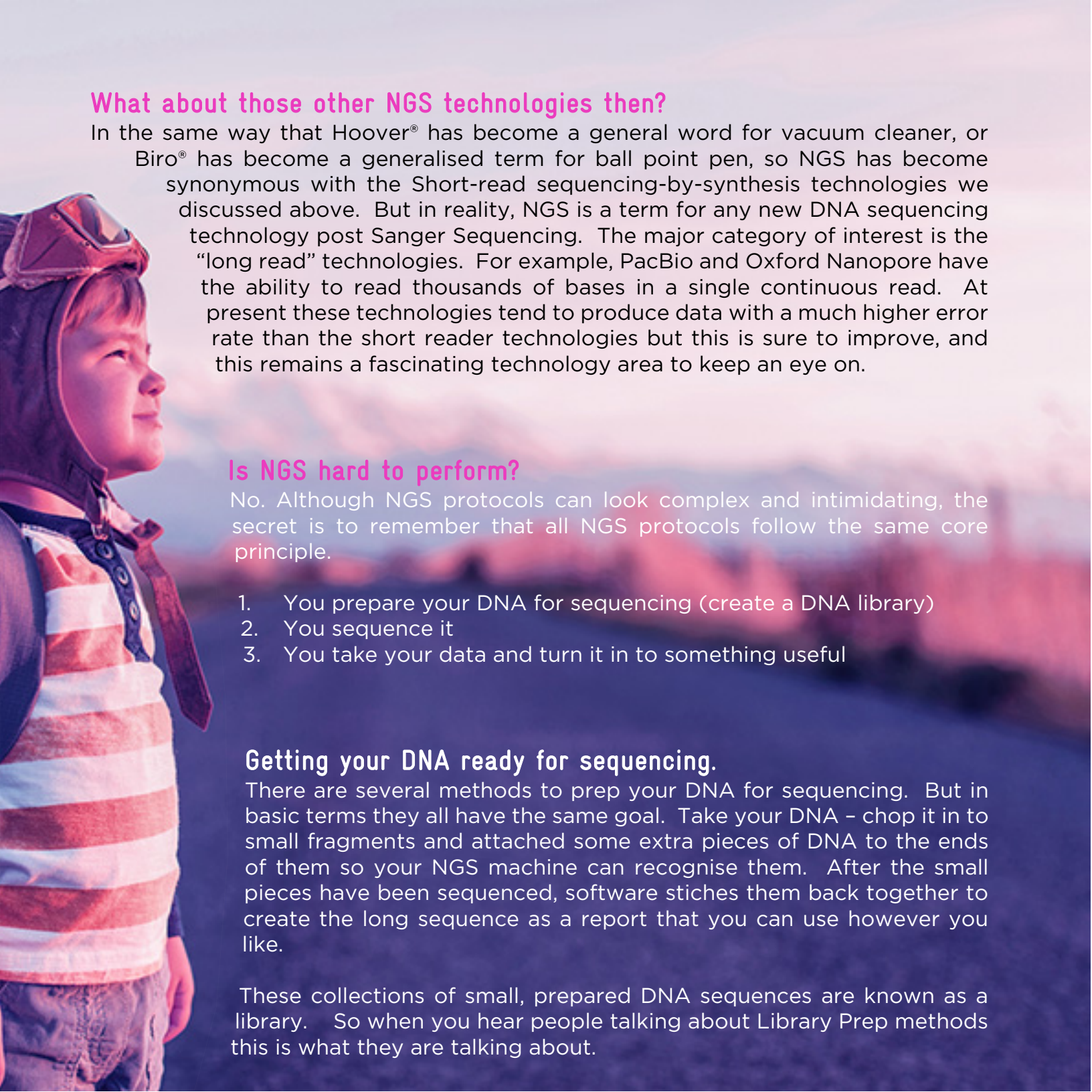
It is pretty accurate and specialised enzymes with very high accuracy rates are used in the reactions. However, to ensure accuracy these technologies repeat the sequencing of each fragment many, many times over to make sure the same result is achieved each time. This happens simultaneously in the NGS instrument with many parallel reactions occurring at the same time. The results from each parallel reaction are compared by computer and statistics are used to eliminate any errors or spurious results.

This repeated sequencing of the same region is often referred to as a “sequencing depth.” The more times you sequence the same region the “deeper” the read and the more certain you can be of sequence being correct.

Depending on the question you are trying to answer defines the amount of depth (or in affect accuracy) that you require. For example, if you’re hunting for a rare cancer-related mutation in amongst thousands of strands of healthy DNA you may need a read depth of 100,000 or more to find the mutation. Whilst if you are sequencing to gather general sequence data of an individual then a depth of 50 may well be sufficient.

What are the advantages of NGS vs Sanger sequencing?

The “massively parallel” nature of NGS that we discussed above is the big advantage of the technology. This means millions of sequencing reactions can occur at the same time on the same instrument. This has dramatically increased the speed and reduced the cost of DNA sequencing vs the older Sanger technology. So where as it took 10 years and £100m to sequence the first human genome, NGS makes it possible today to sequence a whole human genome in a day for less than £1000



What about those other NGS technologies then?

In the same way that Hoover® has become a general word for vacuum cleaner, or Biro® has become a generalised term for ball point pen, so NGS has become synonymous with the Short-read sequencing-by-synthesis technologies we discussed above. But in reality, NGS is a term for any new DNA sequencing technology post Sanger Sequencing. The major category of interest is the "long read" technologies. For example, PacBio and Oxford Nanopore have the ability to read thousands of bases in a single continuous read. At present these technologies tend to produce data with a much higher error rate than the short reader technologies but this is sure to improve, and this remains a fascinating technology area to keep an eye on.

Is NGS hard to perform?

No. Although NGS protocols can look complex and intimidating, the secret is to remember that all NGS protocols follow the same core principle.

1. You prepare your DNA for sequencing (create a DNA library)
2. You sequence it
3. You take your data and turn it in to something useful

Getting your DNA ready for sequencing.

There are several methods to prep your DNA for sequencing. But in basic terms they all have the same goal. Take your DNA – chop it in to small fragments and attached some extra pieces of DNA to the ends of them so your NGS machine can recognise them. After the small pieces have been sequenced, software stiches them back together to create the long sequence as a report that you can use however you like.

These collections of small, prepared DNA sequences are known as a library. So when you hear people talking about Library Prep methods this is what they are talking about.

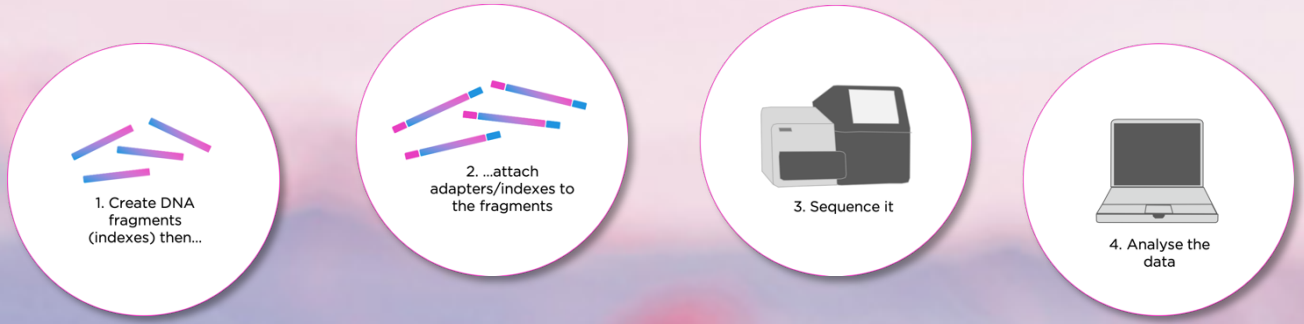
Library prep methods

There are several different library prep methods. You need to choose the right one depending on what you are trying to achieve

No matter which method you choose, they all end up with your DNA in small chunks (e.g. 150 base pairs) with some extra pieces of DNA added to the ends of them so your NGS instrument can recognise them. These additions are known as adapters and indexes.

What are adapters and indexes?

Once you've made your DNA inserts (the small chunks) you attach adapters and indexes to the end. Adapters are small pieces of DNA that allow the insert to bind to the a chip or bead in your sequencing instrument. The index is a small sequence of DNA that is unique – like a barcode. This gets sequenced and allows you to identify which index comes from which sample.



Lots of companies have subtle variations of the following library preparation protocols. But the vast majority centre around these three main principles:

1. If you want to sequence a small region of DNA: e.g. Some diseases are diagnosed by looking for a 1, 2 or maybe a small handful of mutations in a certain gene. In these instances its normal to use PCR to amplify the region of interest in approx. 150bp lengths. Then use another round of PCR to attach the adapters. You then perform a simple clean up protocol to get rid of impurities and the DNA (your library) is ready to be loaded in to a sequencing instrument to be sequenced.

   This PCR based method ideal for sequences of less than 10,000 bp

2. If you want to sequence a medium sized region of DNA: e.g. you may want to sequence a chunk of somebody's genome (10,000-500,000 bases). In this instance you use PCR to create longer pieces of DNA (amplicons). These can be around 2-10,000 bases long.
   Then, as a next step, you use enzymes (or other mechanical methods) to chop up these long pieces in to small fragments (e.g. 150 base pairs). Then you use an enzyme called a ligase to attach your adapters. Then, after a simple clean up protocol to get rid of impurities, the DNA (your library) is ready to be loaded in to a sequencing instrument to be sequenced.

3. If you want to sequence a very large region of DNA: e.g. you many want to sequence a whole genome or a whole exome (the portion of the genome that encodes for proteins). These can be huge -up to 3000 mega bases (3 billion bases!).
   To prepare your DNA for this you use a method called hybridisation capture. You use an enzyme or mechanical methods to break the DNA in to small chunks. Then you incubate the DNA with DNA probes that are designed to bind (hybridise) all over the genome. These probes are "Baits". Next you add magnetic beads to the mix. The baits bind to the beads allowing you to wash away everything else. Finally the DNA fragments are released from the beads and ligase enzyme is used to attach adapters to them. Your DNA (your library) is ready to be loaded in to your sequencing instrument.

What about the actual Sequencing reaction – that must be difficult perform?

The companies that build NGS sequencing instruments have all done a very nice job of taking all the complicated chemistry and putting it in a neat box that does the clever, complex part for you. Your job is to create a good quality DNA library, load it on to the sequencer the box will do the magic for you!

The data processing. Now that really does look scary…

Well yes, it can look intimidating. And yes, it can involve using powerful computers to crunch large amounts of data through complex algorithms. But all of these bioinformatic exercises centre around the same ideas:

- Your sequencer produces a raw a data file called a BCL file with all of the data for all of your samples
- The next step is normally a demultiplexing step. The BCL file data is split out so that the data for each of your samples exists as a separate file.
- Then the data for each sample is converted in to its own FASTQ file. A text-based format that contains the sequence info as well as info about the quality of the data.
- Your data will need to go through various filters to cleanse it and remove anything you don't want
- Next the data is aligned against a reference sequence of DNA. Simply comparing the data to another known sequence to establish which bases from which genes you have sequenced (BAM or SAM file)
- You may then compare your data to another database of data to identify any interesting variants in your data. (VCF file)

The best software solutions do all of this quickly and automatically at the click of a button and produce a neat report with your data annotated and easy to understand

You can learn much more about NGS with our simple, quick videos on the YouSeq learning zone www.youseq.com/learning

Next Generation sequencing
Glossary / Keywords / Buzzwords

- **Adapter** – A small section of DNA that is attached to an index to bind it to the chip/bead in an NGS sequencer.
- **Amplicon** – An insert for sequencing that is made by PCR.
- **Cluster**. A cluster of the same DNA insert on the surface of a chip in an NGS sequencing machine.
- **Coverage**. The amount (normally expressed as a %) of the DNA target that has been successfully sequenced.
- **Depth**. The number of times a nucleotide is sequenced. The deeper the read depth, the higher degrees of confidence in the base calls.
- **Index**- A small section of DNA that is attached to an index to identify it when it is mixed with other samples. The index itself is sequenced thus it effectively acts as a "barcode" to identify the molecule.
- **Insert**. A small piece of DNA of an appropriate size to be sequenced.
- **Library**. A collection of DNA fragments prepared for sequencing.
- **Mate Pair Read (ME).** A sequencing chemistry similar to paired-end (PE) but both reads come from a single strand of DNA.
- **Overhangs**. An extra section of DNA on the end of a PCR primer that acts as a template to bind adapters to.
- **Paired-End Read (PE).** A sequencing chemistry which reads from both ends of a DNA insert.
- **PhiX** - PhiX is a ready-made DNA library. It's useful as a positive control to check all is working well with your sequencing run. It also adds "complexity" which is useful. DNA sequencers don't like to sequence multiple identical libraries at the same time so adding PhiX mixes things up a bit and helps your sequencer perform better.
- **Read length**. The length of the DNA insert that is sequenced in a sequencing reaction.
- **Reference genome**. A complete set of genome data that is used as a refence for post sequencing analysis. Once your DNA sample of interest has been sequenced it is common practice to align it against a reference genome to work out which region has been sequenced.
- **Single-End Read (SE).** A sequencing chemistry which reads from only one end of the DNA insert.

- **Targeted sequencing.** A protocol where a relatively small region of targets is sequenced.
- **Whole genome sequencing (WGS).** A protocol where the entire genome of the species of interest is sequenced.
- **Whole Exome sequencing (EGS).** A protocol where the exome – or protein coding component of the genome is sequenced.

For any more information about NGS please don't hesitate to contact our friendly team of experts at http://www.youseq.com/