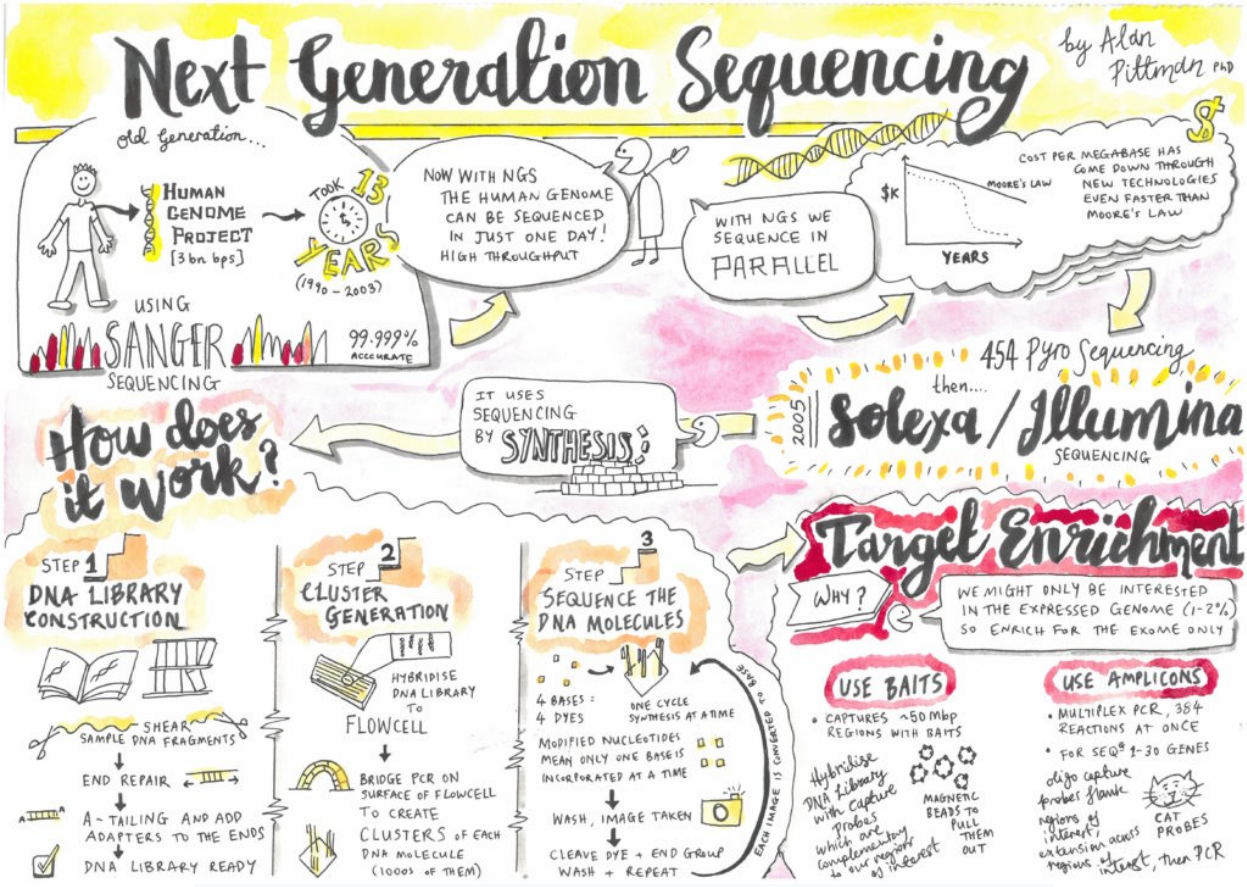
bioscentric

Notes on NGS

By bioscentric | Biology



Next Generation Sequencing *by Jenny Cham, on flickr (Creative Commons license)*

I sometimes have a chance to speak to audiences about how developments in genomics and industrial biotech are relevant to the food, nutrition, personal care and beauty industries. The conversations generally cover bio-based raw materials, new active ingredients, and what the implications of personalized genetic data might be for the food and personal care regimens of the future. However, people often ask questions about the nuts and bolts technology pipeline that allows high-throughput biology to actually get done. With this in mind, I thought we could explore one of the foundational technologies for most of today's disruptive biotech and genomics work – Next Generation Sequencing (NGS).

Reading the Book of Life

Roughly speaking, the book of life (aka the "genome" of an organism) is written in the four types of nucleotides that comprise DNA: "A" for adenine, "T" for thymine, "G" for guanine, "C" for cytosine. For us humans, that book is entirely composed of

combinations of A, T, C, G written out to a length of approximately 3 billion letters. If you were to print it (and people have), it would fill about 130 volumes and take about 90 years to read. Fortunately, Next Generation Sequencers are much faster readers than people and are currently able to sequence a human genome in about three days for around $1500.

Getting to this current point of efficiently sequencing a genome has been a monumental undertaking. Researchers finished the first human genome sequence only 15 years ago, a process that took over 13 years and 2.8 billion dollars to complete. The trick that makes NGS so much faster than previous sequencing methods is that it does not read DNA like we would read a book (i.e. word-by-word, line-by-line). Instead, NGS takes multiple copies of the book, shreds them, and reads random sentences in parallel. Researchers then reassemble the book from this shredded mess by matching overlapping copies of sentences together. For example, we could piece together "The quick brown fox jumped over the lazy dog" as follows.

The quick brown fox jumped over the lazy dog.

The quick brown fox

brown fox jumped

fox jumped over the

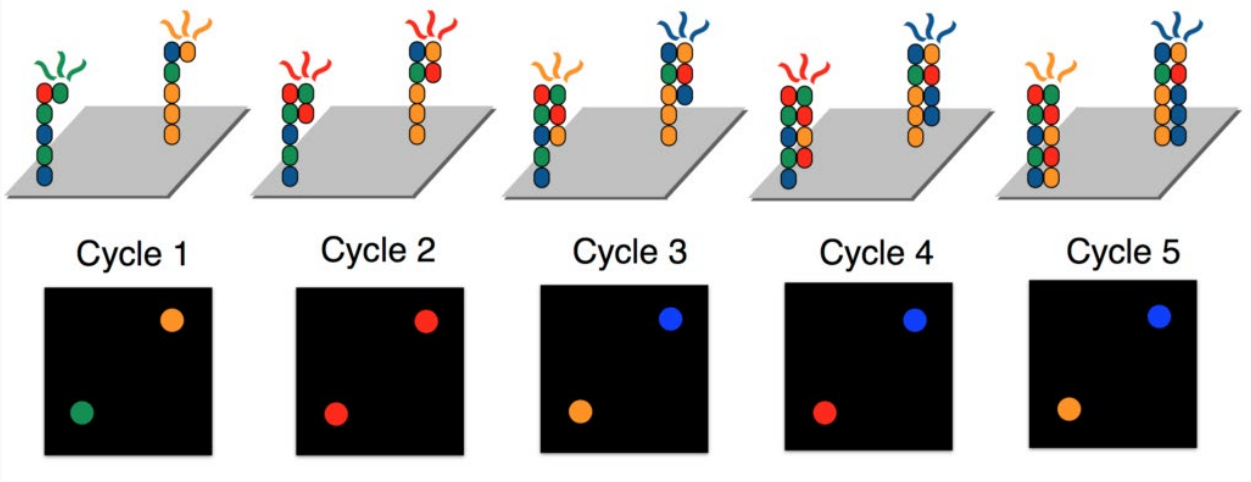
over the lazy dog.

In reality this is a much harder problem as the genome does not have punctuation, capital letters, and a regular grammar. To complicate matters further, DNA sequencing is imperfect and will introduce errors into the fragments. We can also see why it is so difficult to sequence the first genome of any organism – it's like piecing together a novel without knowing the story!

NGS Nuts and Bolt

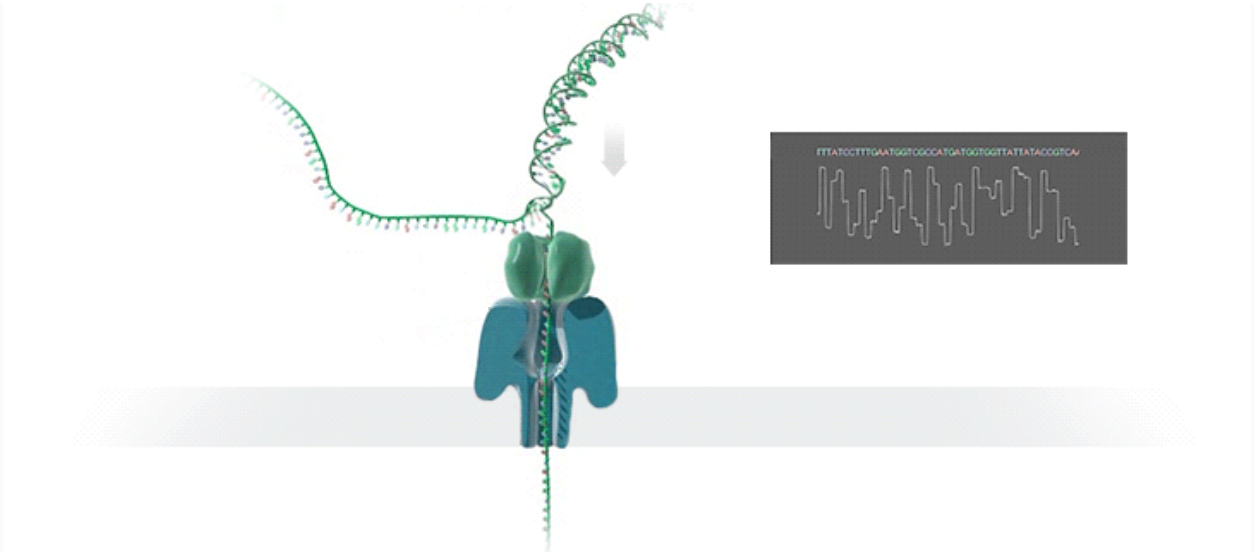
So how does NGS actually work? Every sequencing platform employs a trick to convert the chemical structure of the four bases of DNA (A, T, G, C) into a signal that can be processed digitally. In the case of the most popular sequencers used today (from a company called Illumina), this means associating a color with each base, as in the diagram below. First, DNA is unzipped into single strands and attached to a flow cell, which is a fancy piece of glass that DNA can bind to and chemical mixtures can flow

through. Next, DNA bases with fluorescent dyes attached to them are pumped through the flow cell and bind to their complementary base pair (A-T, G-C, and vice versa) on the exposed single strands of DNA. The sequencer then shines a laser on the flow cell that makes the newly added bases fluoresce a specific color: A – Blue, T – Green, G – Yellow, and C – Red. A camera takes a picture of the flow cell, which now looks like a grid of colored dots, and the cycle repeats again. Turning back to our book analogy, each dot corresponds to a different sentence fragment, and by registering the sequence of colors we can read that sentence letter by letter.



(Ben Langmede)

Instead of using colors, another technology developed by a company called Oxford Nanopore, uses sequencers that associate each base with changes in an electrical signal (see below). As the name would suggest, nanopore sequencing involves tiny holes or pores. These holes are actually a protein positioned on a synthetic membrane. When a voltage is applied across this membrane, ions flow through the nanopore, creating an electrical signal that can be measured. As something is pulled through the pore, it disrupts the flow of ions and the voltage changes. Amazingly, each base of DNA affects the flow of ions in a unique enough way that we can identify them by their characteristic change in voltage. The nanopore is an amazing example of a biology based “machine” comprised of both cellular and solid state parts, where the output goes from biological information to digital.



(Oxford Nanopore)

The different approaches to NGS trace out a product landscape which is roughly segregated along three principal axes – throughput (how much can you read), read length (how long are the sentence fragments), and accuracy (how good is the spelling). Illumina is by far the dominant player in the field here, with the highest throughput, best accuracy, but short read lengths. Oxford Nanopore sequencers have much longer read lengths than Illumina (>100x in some cases), but have lower accuracy and throughput. Consequently, nanopore sequencers are primarily still used in academic settings, but with their accuracy and throughput increasing they could start to threaten Illumina's hegemony.

Going Forward

Hopefully this post gives you some idea of what NGS hardware is and how it works. In the coming weeks I'll get into the costs associated with NGS, why you should be thinking that NGS is like AWS, and some practical examples of how NGS is leveraged. Here are some additional resources if you want to learn more about NGS or nanopore sequencing.