DNA Sequencing and Assembly

CS 262 Lecture Notes, Winter 2016
February 2nd, 2016
Scribe: Mark Berger

Abstract

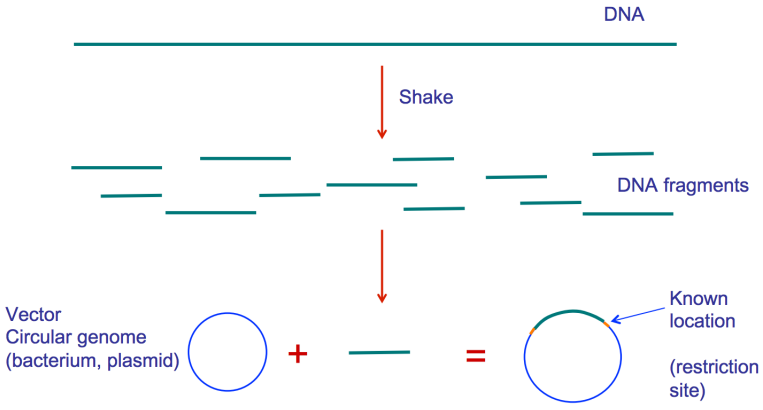
In this lecture, we survey a variety of different sequencing technologies, including their respective advantages and disadvantages. Then, we begin to explore the topic of genome assembly by discussing the Human Genome Project, and how genetic repeats make assembly a difficult computational task.

1 Sequencing Technologies

Currently, we live in a time where genome sequencing can be done for $1,000 a person, an amazing feat, considering that the first human genome was sequencing in 2000 for over two billion dollars. This is now possible due to a variety of next generation sequencing technologies, which we will explore today. Additionally, we will discuss technologies, which while more expensive than $1,000 per person, provide much higher quality data, and overcome some shortcomings of the cheaper technologies.

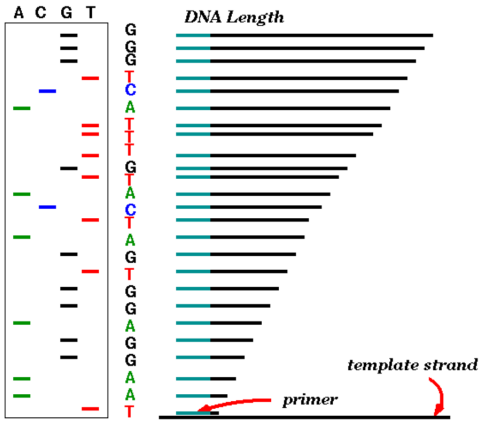
1.1 Sanger Vectors

Before the arrival of next generation sequencing technologies, scientists used Sanger Vectors to sequence DNA. First, DNA is fragmented into small pieces, and then inserted into a Bacterial Artificial Chromosomes (BAC). These artificial chromosomes allow the DNA to replicate, thereby producing enough DNA for gel electrophoresis.



Once the DNA has been replicated to a sufficient amount, it is ready to be sequenced. Starting at the primer, the DNA chain is grown, with a supply of dideoxynucleosides. These dideoxynucleosides serve as modified As, Cs, Gs, and Ts, which end the reaction when they attach to the DNA. Since these modified nucleotides attach to the DNA fragment randomly, the reaction is stopped at all possible points in the sequence. Then, using gel electrophoresis, the products are separated to infer

the sequence. Bigger fragments move less in the gel, and this allows us to infer the position of the base pair at the end of the fragment.



At its peak, this process could produce genetic reads of roughly 900 base pairs. However, there are a variety of caveats associated with this method. First, the length of the fragments follow a geometric distribution. Therefore, as the read length becomes longer, there are fewer and fewer fragments which produce information about the base pairs at the end of the read. Secondly, as the read gets longer, it is harder to infer which base pair is at position N, and which base pair is at position $N+1$.

1.2 Quality Scores

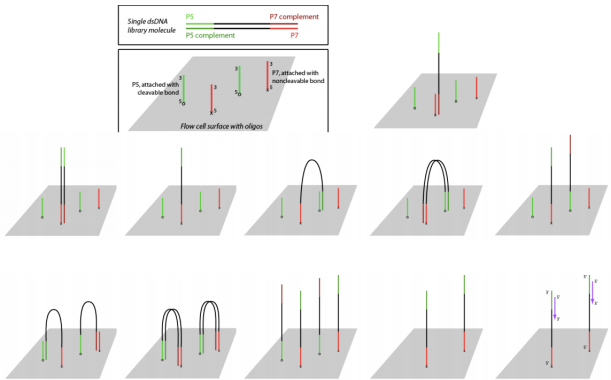
In order to automate the reading of the gel electrophoresis, a technology called Florescent Sanger Sequencing Trace was developed. This technology led to the introduction of genomic quality scores. They are defined as follows:

$$Quality(bp) = -10 \cdot log_{10} P(error)$$

This equation gives us the following quality scores for a variety of different accuracies: $90\% \rightarrow 10$, $99\% \rightarrow 20$, $99.9\% \rightarrow 30$, and $99.99\% \rightarrow 40$. 40 is considered to be the gold standard of read qualities.

1.3 Illumnia Clustering

Illumina short read technology is currently the most popular sequencing technology. Reads are sequenced in clusters, which are generated using a process called bridge amplification. First, DNA is sheared and tagged with adapters. Then, with a series of cycles, the DNA fragment is complemented to produce an additional strand. The DNA is then folded over to another receptor, and the receptors on opposite strands are detached, to form two strands.



This process is then repeated to form large clusters of DNA strands simultaneously. Finally, in order to read the bases, a process called sequencing by synthesis is used. Through a variety of cycles, the complement base pair is added to the strands of DNA, and the signal from that complement base pair is recorded. Then, the process is repeated to sequence the next base pair, and is repeated until the entire strand has been sequenced. Overall, the read error is less than one percent with this technology.

While this approach is the standard due to its affordable price point, the technology has a variety of disadvantages. First, this technology is limited to reads of roughly 150 bp. Additionally, this technology has a large sequence bias, so regardless of the coverage, some regions of the genome will always be poorly sequenced.

1.4 PacBio SMRT Technology

While Illumina reads are rather short, PacBio SMRT technology is used to generate reads with a mean size of 10 - 15 kilobases, with a maximum size of 30 kilobases. This is accomplished by reading a single DNA molecule at a time, but at the cost of having a very high error rate. These long reads, along with a high error rate, make processing and assembling the reads a different, much more difficult, computational task [1].

However, long reads provide a variety of advantages compared to short reads. Structural variants, such as an insertion, or a deletion, are difficult to detect if they are longer than the size of the read. Therefore, longer reads allow us to detect many more structural variants. Additionally, longer reads are better for haplotype phasing, which we will discuss in a future lecture. Finally, the errors associated with this long read technology are more stochastic, and less bias, than the popular Illumina technology. Therefore, long reads enable us to sequence certain areas of the genome with much higher accuracy.

While PacBio sequencing is not very popular in human sequencing, it is popular in agriculture, where the high quality of long reads is more suitable for the field's specific needs.

1.5 Oxford Nanopore

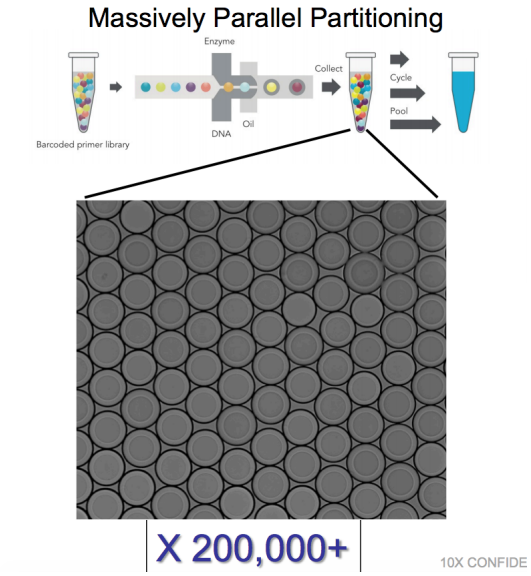
Similar to PacBio, Oxford Nanopore technology is able to read a single base pair at a time using its proprietary nanopore technology. While the error rate is currently even higher than PacBio, this technology is interesting because it is a portable device that attaches to a computer via USB. Additionally, the device will be free after agreeing to purchase $75,000 worth of reactants.

1.6 Moleculo

Moleculo is another long read technology that was developed here at Stanford by Steven Quake's group, and then purchased by Illumina in 2013. First, DNA is sheared to roughly lengths of roughly 10kb. Then, the fragments are diluted, and distributed into 384 wells. The fragments in each well are then amplified with PCR, cut into short fragments, and tagged with the individual well's unique barcode. Finally, these short fragments are pooled together, and sequenced. This is advantageous because the barcodes allow for greater accuracy during assembly, as well as haplotype phasing.

1.7 10X Genomics

10X Genomics is a similar technology to Moleculo, only instead of having 384 wells, there are 200k to one million water bubbles with beads placed in oil. The additional barcodes further enhance the advantages of the Moleculo technology.



2 Assembly

Assembly is the process of taking the set of raw reads from sequencing, and placing them together to form a fully assembled genome. There are two types of assembly: de novo assembly, and resequencing. De novo assembly is when we are sequencing a species for the first time, and therefore, we have nothing to reference where each individual read should go. Resequencing is when we are sequencing an individual, and we want to see how that individual differs from a reference genome.

2.1 Shotgun Sequencing

In 1990, the Human Genome Project began in earnest, and labs across the world collaborated to sequence the genome using a large number of E. Coli clones. Each clone would produce 5 - 20 kb, and this was thought to be the best approach to sequence the first human genome. While it was possible to sequence the genome this way, the process was incredibly expensive and slow, with an estimated budget of $3 billion and an end date of 2005.

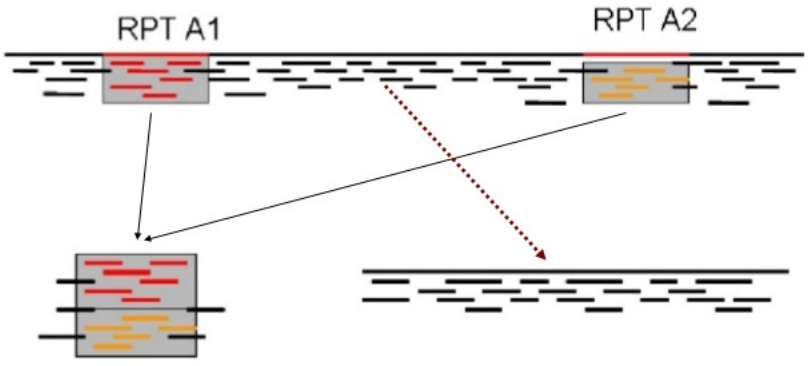
In 1997, James Weber and Jim Myers published a paper titled, "Human whole-genome shotgun sequencing" [2]. They argued, using computational simulations, that instead of using long reads, short reads would contain enough information to sequence the first human genome, not only cheaper, but faster than the Human Genome Project. This idea was so controversial that Phillip Green wrote an enormous review, stating why shotgun sequencing was not feasible [3]. Both works were eventually published side by side in Genome Research.

Craig Ventur believed Weber and Myers were correct, and along with Jim Myers, he founded the company Celera Corporations. The goal of Celera was to sequence the human genome using the shotgun sequencing approach. This spurred the Human Genome Project, as the government was worried that the human genome would become property of Celera, instead of an open scientific resource. Eventually, both projects finished at the same time, and a draft of the first human genome was completed in early 2000. Therefore, it is computer scientists, not geneticists, that we must thank for the rapid advancements of genomics.

2.2 Assemblies Difficulties

At a high level, we want to assembly the genome in the following way: first, we produce a lot of reads to create high redundancy, and then, we overlap and extend these reads to form an entire genome. While this would work fairly well for completely random genome, and sufficiently long reads, this does not work in practice. On average, half of the human genomes consists of repeats, which are patterns of sequence which appear multiple times in the genome.

Since repeats can span a large number of base pairs, short reads have a difficult time correctly assembling the genome.



The repeated element is collapsed into a single contig

As you can see in the above figure [4], with a large repeat, it is difficult to know where a repeat begins and ends relative to the rest of the genome. Therefore, identical repeats can be accidentally merged into a single continuous piece of sequence.

In order to address this problem, we can either cluster the reads, or link the reads. We will discuss these approaches next time.

References

[1] Chakraborty, Mahul, et al. ”A practical guide to de novo genome assembly using long reads.” bioRxiv (2015): 029306.

[2] Weber, James L., and Eugene W. Myers. ”Human whole-genome shotgun sequencing.” Genome Research 7.5 (1997): 401-409.

[3] Green, Philip. ”Against a whole-genome shotgun.” Genome Research 7.5 (1997): 410-417.

[4] Figure by Dr. Torsten Seeman from his talk ”De novo Genome Assembly”.