Next Generation Sequencing

1. Introduction

“Next Generation Sequencing” (NGS; or “massively parallel sequencing”) refers to a group of new DNA sequencing technologies that can rapidly sequence DNA on the gigabase scale. These methods are now replacing Sanger sequencing, which was the dominant sequencing technology from the late 1970’s to the late 2000’s and was used for all of the initial genome sequencing projects (*H. influenzae*, *yeast*, *Drosophila*, *Arabidopsis*, human, etc.). Several different “next generation” sequencing platforms have been developed and commercialized, with more on the way.

2. “pre-next generation” – sequencing by hybridization

This approach can be used for SNP (single nucleotide polymorphism) detection or even whole genome re-sequencing. It requires that you already know most of the sequence and just want to identify rare differences among individuals. This can be used for model organisms that already have a high-quality reference genome sequence available.

The method is very similar to the Affymetrix microarray system that uses 25-base oligonucleotides with either a perfect match or a mismatch at the center (base 13). However, here genomic DNA is hybridized to an array that has 4 probes, each with a different nucleotide at the center. The center base can be “called” by determining which of the probes it hybridizes to.

For example, if we want to determine base N of the following part of a genome:

TGGTCATCGACT**N**GGTACCTGACTA

We hybridize it to an array with the following 4 probes:

1) ACCAGTAGCTGA**A**CCATGGACTGAT

2) ACCAGTAGCTGA**C**CCATGGACTGAT

3) ACCAGTAGCTGA**G**CCATGGACTGAT

4) ACCAGTAGCTGA**T**CCATGGACTGAT

if the strongest hybridization signal is to probe 3, then we assume base N is a C.

With enough probes on the array, it is possible to query almost every base in the genome and, thus, “re-sequence” the genome of an individual, finding all bases that differ from the reference genome.

Pro: fast, inexpensive, high throughput (can re-sequence many individuals)

Con: requires reference genome, is not good at base calling if there is more than one polymorphism (or insertion/deletion) in a 25 base region.

This approach has been used for a few whole genome re-sequencing projects, but it is mostly used for typing SNPs. Often hundreds of thousands of known SNPs are “typed” in one experiment.

3. Next Generation Methods

The first 3 NGS methods to gain widespread commercial use are described below. They follow an approach similar to Sanger sequencing, but do away with separation of fragments by size and "read" the sequence as the reaction occurs. They are not used to sequence a specific template, but instead simultaneously sequence entire libraries of DNA sequence fragments.

a) 454 (also known as pyrosequencing or Roche GS FLX)

This was the first next generation method to be commercially available and the first to be applied to large-scale sequencing projects, such as sequencing the genome of James Watson. Uses a "sequencing by synthesis" approach.

– DNA is broken into pieces of 500-1,000 bp, ligated to adaptors, and amplified on tiny beads by PCR (emulsion PCR)

– Beads (with DNA attached) are placed into tiny wells (one bead per well) on a PicoTiterPlate that has millions of wells. Each well is connected to an optical fiber.

– DNA is sequenced by adding polymerase and DNA bases containing pyrophosphate. The different bases (A,C,G,T) are added sequentially in a flow chamber. When a base complementary to the template is added, the pyrophosphate is released and a burst of light is produced. The light is detected and used to call the base. If the same base occurs multiple times in a row, the light signal will be proportionally stronger.

The original 454 read lengths were 100-150 bp, but they now have been improved to 700-800 bp. Paired end reads are also possible. One run produces about 1 million reads in 10 hours. In total, one machine can sequence >1 Gb per day.

b) Illumina (originally known as Solexa)

This method uses a "sequencing by synthesis" approach that is similar to traditional Sanger sequencing in that it uses fluorescently labeled terminators. It has already been used to sequence the entire genome of one African and one Asian human, plus the genome of a cancer patient.

– DNA is broken into small fragments and ligated to an adaptor.

– The fragments are attached to the surface of a flow cell and amplified.

– DNA is sequenced by adding polymerase and labeled *reversible* terminator nucleotides (each base with a different color). The incorporated base is determined by fluorescence. Then the fluorescent label is removed from the terminator and the 3' OH is unblocked. This allows a new base to be incorporated and the process repeats.

The original Solexa read lengths were 35 bp, they have now been increased to up to 150 bp (and in some cases 250 bp). Paired-end reads are also possible. One lane of the machine can give 180 million reads. The output is >1 Gb per day.

c) SOLiD (ABI)

This is a "sequencing by ligation" method. It does not use polymerase, but instead uses DNA ligase for sequencing. It is sold by Applied Biosystems (ABI), the leading company for automated sequencers using the Sanger method.

– DNA is broken into small fragments and ligated to an adaptor.

– The fragments are attached to beads and amplified by emulsion PCR. Beads are attached to the surface of a glass slide.

– DNA is sequenced by adding 8mer fluorescently labeled oligonucleotides. If an oligo is complementary to the template, it will be ligated and 2 of the bases can be called. The attached oligo is then cut to remove the label and the next set of labeled oligos are added. The process is repeated from different starting points (using different universal primers) so that each base is called twice (two-base encoding). This allows for more accurate base calls.

The original SOLiD read lengths were 25 bp, they were later increased to up to 50-100 bp. One run of the machine can give 85 million reads. The output is >1 Gb per day.

4. Applications

Whole Genome Sequencing

Since all of the above methods give rather short read lengths, they are often used for re-sequencing. In this case, it is not necessary to do a complete, independent genome assembly, but the sequence reads can be aligned to a reference genome sequence. For example, the sequence reads from a single person can be aligned to the reference human genome. However, all of the above methods have been modified to produce "paired reads" in which both ends of a DNA fragment of known length are sequenced. This makes it possible to do *de novo* assemblies of genomes.

Gene expression profiling

All of the above methods can also be used to sequence cDNA. Since even a short read length is enough to map a particular cDNA to the genome, these methods can be used as a way to measure gene expression. This is similar to SAGE, but does not require the isolation of SAGE tags. Millions of cDNA fragments are sequenced in a single run and each fragment is then mapped back to its corresponding gene in the genome. The more reads that match a gene, the higher the expression of that gene. Since this approach produces exact counts of transcript abundance, it is sometimes called "digital expression profiling". It is also known as "RNA-seq". As throughput increases and the cost decreases, this appraoch is replacing microarrays as the major technique used in transcriptomics. It is also possible to assemble *de novo* transcriptomes (non-model organisms) using sinlge- or paired-end reads

5. What's next? The next-next generation.

Although the 3 methods described above are the first to be commercialized on a large scale and applied to genomics, they do have some limitations:

a) Read lengths are relatively short. Usually 50-250 bp (but now up to 700 bp for 454 sequencing)

b) Base calling is relatively slow. A base is added to the template, then the base is interrogated (called), then the next base is added.

Methods are being developed that work on a single DNA molecule and call the bases directly as they are incorporated by the polymerase. These "real time" sequencing methods are much faster and can give longer read lengths.

6. Additional reading

Mardis, ER. (2008) The impact of next-generation sequencing technology on genetics. Trends in Genetics 24:133-141.

Sanger who? Sequencing the next generation.
http://www.sciencemag.org/products/lst_20090410.dtl