Sequencing Whole Genomes

Hierarchical Shotgun Sequencing v. Shotgun Sequencing

How do you sequence a whole genome?

There are two general strategies for sequencing a complete genome. The method preferred by the Human Genome Project is the **hierarchical shotgun sequencing** method. In this approach, genomic DNA is cut into pieces of about 150 Mb and inserted into BAC vectors, transformed into *E. coli* where they are replicated and stored. The BAC inserts are isolated and mapped to determine the order of each cloned 150 Mb fragment. This is referred to as the **Golden Tiling Path**. Each BAC fragment in the Golden Path is fragmented randomly into smaller pieces and each piece is cloned into a plasmid and sequenced on both strands. These sequences are aligned so that identical sequences are overlapping. These contiguous pieces are then assembled into finished sequence once each strand has been sequenced about 4 times to produce 8X coverage of high quality data.

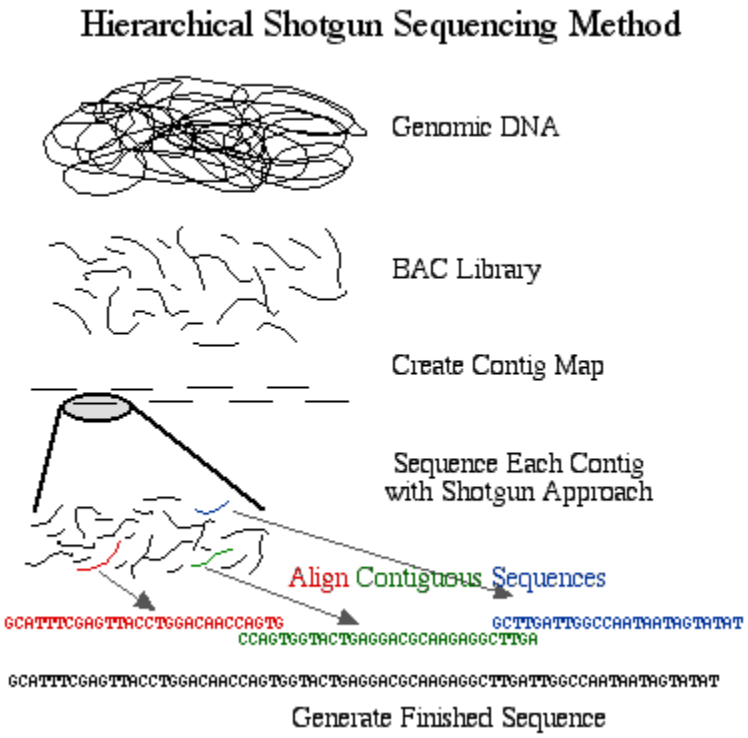


Figure 1. Schematic diagram of sequencing strategy used by the publicly funded Human Genome Project. The DNA was cut into 150 Mb fragments and arranged into overlapping contiguous fragments. These contigs were cut into smaller pieces and sequenced completely.

The method developed and preferred by Celera is simply called **shotgun sequencing**. This approach was developed and perfected on prokaryotic genomes which are smaller in size and contain less repetitive DNA. Shotgun sequencing randomly shears genomic DNA into small pieces which are cloned into plasmids and sequenced on both strands, thus eliminating the BAC step from the HGP's approach. Once the sequences are obtained, they are aligned and assembled into finished sequence.

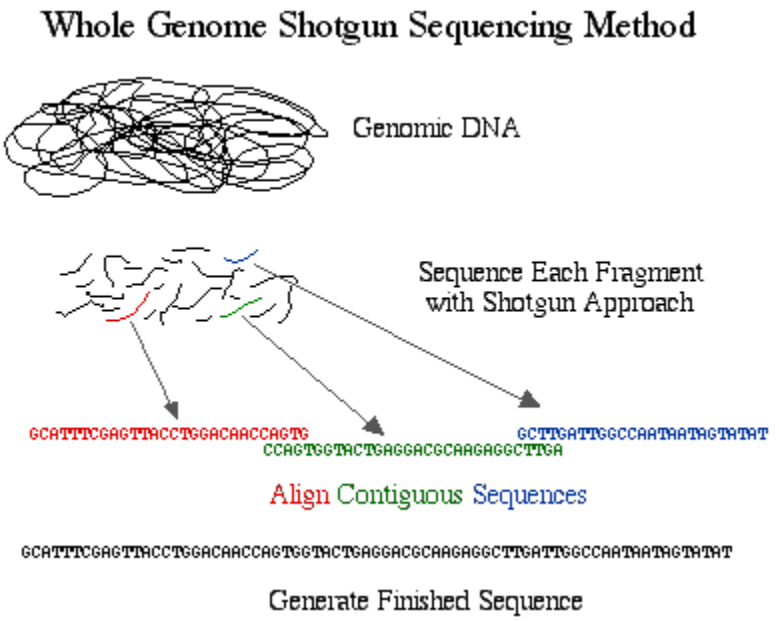


Figure 2. Schematic diagram of sequencing strategy used by Celera. The DNA was cut into small pieces and sequenced completely. These fragments were organized into contigs based on overlapping sequences.

The advantage to the hierarchical approach is sequencers are less likely to make mistakes when assembling the shotgun fragments into contigs as long as full chromosomes. The reason is that the chromosomal location for each BAC is known, and there are fewer random pieces to assemble. The disadvantage to this method is time and expense. The shotgun method is faster and less expensive, but it is more prone to errors due to incorrect assembly of finished sequence. For example, if a 500 kb portion of a chromosome is duplicated and each duplication is cut into 2kb fragments, then it would be difficult to determine where a particular 2 kb piece should be located in the finished sequence since it occurs twice. You might think, "who cares since they're duplicates?" But duplications seldom retain their original sequences; they tend to drift over time. So a small region may be retained while other parts may mutate. This might create overlapping sequences for small pieces that are located several hundred kb apart on the chromosome.

Which method is better? It depends on the size and complexity of the genome. With the human genome, each group believes its approach to be superior to the other. We only have draft sequences and each has gaps and unfinished regions so it is not possible to say for sure. It is worth mentioning that Celera had access to the HGP data but the HGP did not have access to the Celera data. Furthermore, since the Celera data is not freely available, most investigators will use the HGP sequence for further research. Therefore, we may never know which method "won".

SEQUENCING THE GENOME

By Bijal P. Trivedi

June 2, 2000

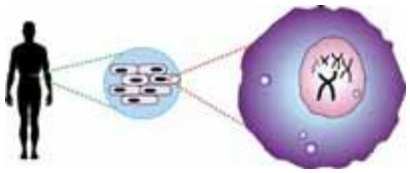
There are essentially two ways to sequence a genome. The BAC-to-BAC method, the first to be employed in human genome studies, is slow but sure. The BAC-to-BAC approach, also referred to as the map-based method, evolved from procedures developed by a number of researchers during the late 1980s and 90s and that continues to develop and change.*

The other technique, known as whole genome shotgun sequencing, brings speed into the picture, enabling researchers to do the job in months to a year. The shotgun method was developed by GNN president J. Craig Venter in 1996 when he was at the Institute for Genomic Research (TIGR).*

Now that the human genome sequence is nearing completion, the next phase—understanding the meaning and function of genes—can begin.

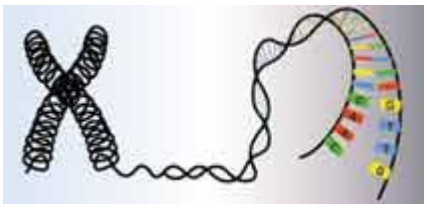
A primer on the two approaches to sequencing follows.

* References to the scientific literature are at the end of this primer.

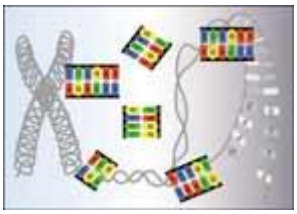


View larger

The human body has about 100 trillion cells. Inside each of those cells is the nucleus that contains the genome—46 human chromosomes—which govern human development.



Each chromosome is one long string of DNA that is tightly coiled in a compact bundle. Chromosomes are comprised of millions of copies of the four letters of the genetic code—the DNA bases A, C, G, T—that are arranged into genes and non-coding sections. Finding the order, or sequence, of these four letters is the goal of genomics. The entire human genome is made up of about 3.5 billion bases.



To read the DNA, the chromosomes are cut into tiny pieces, each of which is read individually. When all the segments have been read they are assembled in the correct order.

Two approaches have been used to sequence the genome. They differ in the methods they use to cut up the DNA, assemble it in the correct order, and whether they map the chromosomes before decoding the sequence. First there was the BAC to BAC approach. A second, newer method is called whole genome shotgun sequencing.

BAC to BAC Sequencing

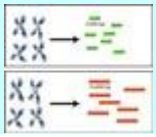
The BAC to BAC approach first creates a crude physical map of the whole genome before sequencing the DNA. Constructing a map requires cutting the chromosomes into large pieces and figuring out the order of these big chunks of DNA before taking a closer look and sequencing all the fragments.

Whole Genome Shotgun Sequencing

The shotgun sequencing method goes straight to the job of decoding, bypassing the need for a physical map. Therefore, it is much faster.

Several copies of the genome are randomly cut into pieces that are about 150,000 base pairs (bp) long.

Multiple copies of the genome are randomly shredded into pieces that are 2,000 base pairs (bp) long by squeezing the DNA through a pressurized syringe. This is done a second time to generate pieces that are 10,000 bp long.

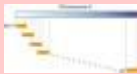
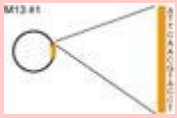


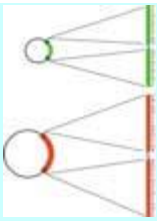
Each of these 150,000 bp fragments is inserted into a BAC-a bacterial artificial chromosome. A BAC is a man made piece of DNA that can replicate inside a bacterial cell. The whole collection of BACs containing the entire human genome is called a BAC library, because each BAC is like a book in a library that can be accessed and copied.

Each 2,000 and 10,000 bp fragment is inserted into a plasmid, which is a piece of DNA that can replicate in bacteria. The two collections of plasmids containing 2,000 and 10,000 bp chunks of human DNA are known as plasmid libraries.

These pieces are fingerprinted to give each piece a unique identification tag that determines the order of the fragments. Fingerprinting involves cutting each BAC fragment with a single enzyme and finding common sequence landmarks in overlapping fragments that determine the location of each BAC along the chromosome.

This step not needed in shotgun sequencing

Then overlapping BACs with markers every 100,000 bp form a map of each chromosome. 	
Each BAC is then broken randomly into 1,500 bp pieces and placed in another artificial piece of DNA called M13. This collection is known as an M13 library. 	This step not needed in shotgun sequencing
All the M13 libraries are sequenced. 500 bp from one end of the fragment are sequenced generating millions of sequences. 	Both the 2,000 and the 10,000 bp plasmid libraries are sequenced. 500 bp from each end of each fragment are decoded generating millions of sequences. Sequencing both ends of each insert is critical for the assembling the entire chromosome.



These sequences are fed into a computer program called PHRAP that looks for common sequences that join two fragments together.

Computer algorithms assemble the millions of sequenced fragments into a continuous stretch resembling each chromosome.

(All images created by Mary S. Gibbs (GNN))