

CS262 Computational Genomics (Lecturer: Serafim Batzoglou)

Lecture 5: DNA Sequencing

The purpose of DNA sequencing is to figure out the nucleotide sequence of a given piece of DNA. Because of its wide implications on disease and biological discovery in general, there have been both government-funded and private efforts (Celera Genomics) to push forward the sequencing of the human genome. Using the shotgun approach and scores of sequencing machines, Celera Genomics completed a draft of the human genetic code in 2000.

1. Polymorphism

Polymorphism rate = number of nucleotide changes between the DNA sequences of two different members of the same species

The DNA sequence between any two people is highly conserved -- the polymorphism rate is approximately 1/2000. This number is estimated to be significantly higher for other species.

For example, there is a small chordate (hint of spine present), called *Ciona savignyi*, that swims around the sea. It attaches itself to a rock, and then waits with an open mouth to absorb food/nutrients. Any two members of this organism have 5% - 10 % difference in parts of their genome. In contrast, there are also species that have higher genomic conservation than humans. An example is tigers, which have a very low polymorphism rate.

The first human whose DNA was sequenced is Craig Venter, the CEO of Celera Genomics. However, because the DNA content between any two human beings is > 99.95 % similar, whichever human we use for human genome sequencing will be representative of the human species as a whole.

2. Population Migration

Evolution Theory 1 (predominant): **Out of Africa Replacement Hypothesis**

Population of humans originated from Africa and replaced other human populations (ie Neanderthals) about 40,000 years ago. Mitochondrial DNA (inherited from mother only) and Y chromosome genetic evidence supports this.

Evolution Theory 2: **Multi-regional Hypothesis**

Many species of humans evolved independently in many different regions. There are some racially motivated, controversial reasons behind the support for this theory and disbelief for the Out of Africa theory.

By tracing mtDNA and using molecular clock (how many mutations per generation), we can conclude that Eve (150,000 years ago) is the “grandmother” of all human beings. Adam (70,000 years ago) is the “grandfather” of all human beings. Adam lived much more recently than Eve because there was higher variance in the number of children for men than for women. Higher variance means a higher chance that a few men’s genes got propagated to the entire population.

Why are we all so similar?

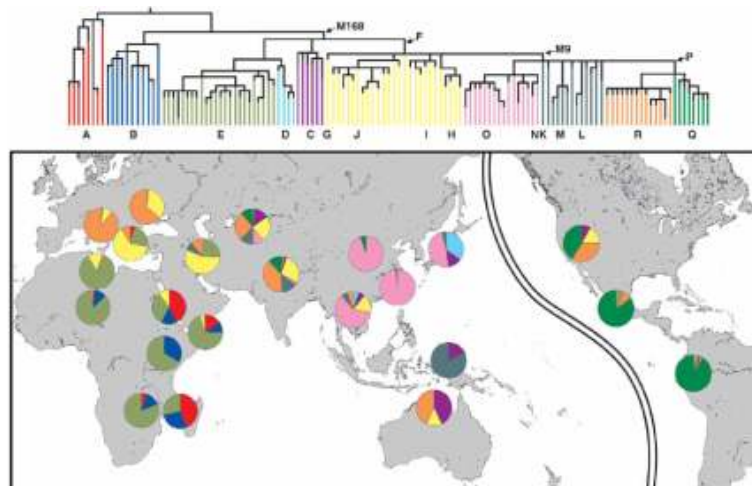
Before humans migrated throughout the world, there was a small population in Africa that interbred over many generations to reduce genetic variation (Bottleneck Effect). Today, we can create lab animals with a specific genetic profile by interbreeding heterozygous (in a particular gene) animals. In every generation, there is a non-trivial chance of losing one copy of the gene → genetic variation is reduced in each round of interbreeding.

It appears that humans migrated from Africa → Middle East → Europe & Australia → US.

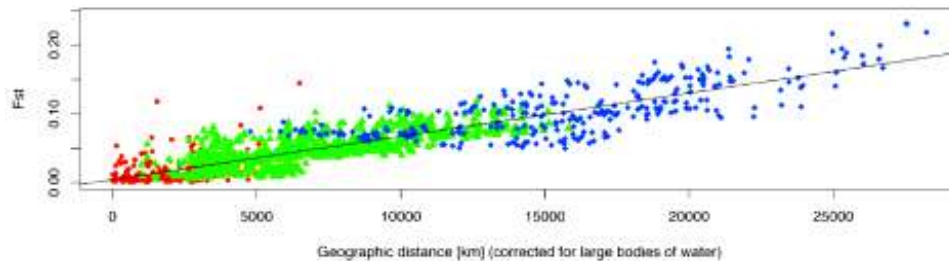


Evidence that origin= Africa:

- 1) Phylogenetic tree constructed by looking at different variations of key nt’s in the Y chromosome, throughout the world. Can see the geographical location of every population migration.



- 2) Analysis correlating geographic distance with genetic similarity. Best correlation between geographic distance and genetic similarity if the origin is assumed to be in Africa (in particular, the Sub Sahara Africa). If you assume origin is in South Africa, you also get a decent fit between geographic distance and genetic similarity if you assume origin is in South America. This is because humans populated South Africa last, so there is some correlation if you examined everything in reverse.



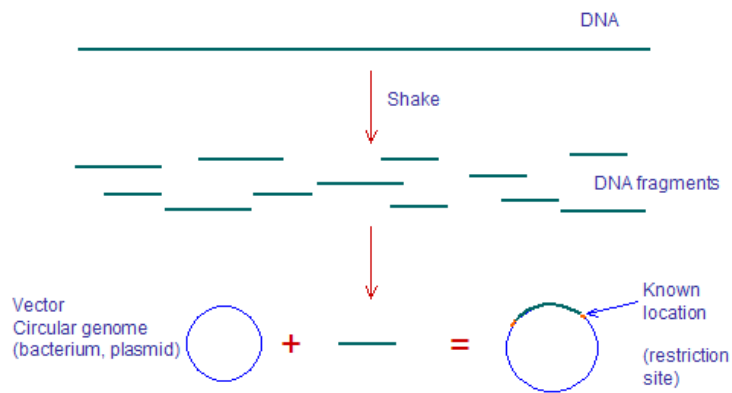
3. DNA Sequencing

3.1 Concepts

Gel Electrophoresis is the main technology used to perform DNA sequencing today. The protocol was invented by Fred Sanger, who received a Nobel Prize for this work. DNA sequencing was, initially, very expensive and laborious, but now the process has become much more efficient and cheaper.

In 1990, the cost of sequencing was 30 dollars per nucleotide. In 2003, the cost was 1 cent per nucleotide. The trend of sequencing costs over the years shows that the cost is decreasing at an exponential rate. However, gel electrophoresis is likely to be replaced by other methods by the end of the decade.

The gel electrophoresis machine can sequence a read of 500-1000 nucleotides of DNA fragment ends. DNA must be excised into fragments of random length using restriction enzymes. The RE digested fragment can then be inserted into a vector (cut with the same set of restriction enzymes) to form a clone. The clones can be replicated by the cellular machinery to give rise to many copies, thus forming a library of clones.



A vector is basically a circular piece of DNA that can be replicated. Examples of a vector are: plasmid and BAC (bacterial artificial chromosome). The vector contains restriction sites that allows for the insertion of a DNA fragment in a site-specific manner. After the fragment is inserted, it can get replicated as part of the vector by the cellular machinery. Also, we know exactly where in the vector the fragment has been incorporated (Restriction sites flank this fragment).

Many types of vectors can be used, depending on the size of the insert. The length of the actual fragment insert can be controlled to +/- 20%.

<u>VECTOR</u>	<u>Size of insert</u>
Plasmid	2,000-10,000 Can control the size
Cosmid	40,000
BAC (Bacterial Artificial Chromosome)	70,000-300,000
YAC (Yeast Artificial Chromosome)	> 300,000 Not used much recently

3.2 Experimental Procedure

- 1) Start with a DNA primer (short ssDNA) that anneals (complementary to) the restriction site on the vector
- 2) Grow DNA chain starting from primer
- 3) The nucleosides that are incorporated include normal a, c, g, t as well as modified a, c, g, and t. The modified nucleosides serve as reaction terminating points once they get incorporated into the growing chain.
- 4) If we perform sequencing procedure with sufficient copies of the clone, the reaction has a high likelihood to stop at every possible position in the sequence → get products of length 1,2,3,4,5.... n (where n = length of read)
- 5) Separate products using gel electrophoresis (the smaller the fragment, the faster the migration when a current is applied)

likely state (letter) for each position? Today, there are better programs out there than PHRED, but labs usually use PHRED because it's regarded as standard.

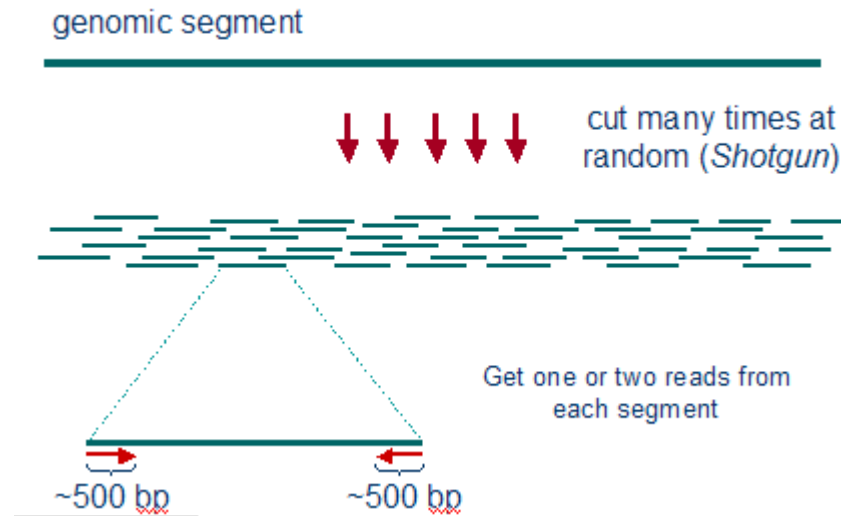
A read is typically 500-1000 nt, and a quality score (lower bounds on the accuracy) is assigned to each read.

Quality Score = $-10 \times \log_{10} \text{Prob}(\text{Error})$

A quality score of 30 translates to a 1/1000 chance of error. A quality score of 40 is the maximum meaningful quality score you can assign to a read.

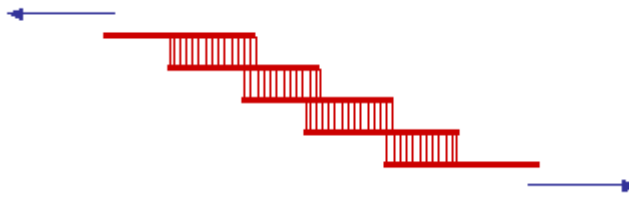
An important method that dramatically changed the way we do sequencing is Double-Barreled Sequencing, which was invented around 1990. This method reads both right and left portions of a fragment, thus getting a paired/two linked reads that came from the same fragment.

3.4 Shotgun Sequencing



Take a DNA segment of genome → excise into fragments at random positions → get 2 reads (500 bp) for each end of each fragment → cover each region with approx. 7 x redundancy → match overlaps of reads (overlap of 50 letters or more is very unlikely to be by chance) → extend into longer and longer sequence





L = length of genomic segment
 N = number of reads
 l = length of each read
 C = coverage
 $C = n l / L$

Reads are distributed randomly, and a good coverage is typically $\geq 7x$. According to the Lander-Waterman Model: if the read distribution is uniform, coverage of $10x$ will give you a gapped region every 1,000,000 nucleotides.

Sequence assembly is difficult due to repeats, which are significant in number for higher organisms. This is because higher organisms have more energy resources to spend in DNA replication without feeling the cost, and therefore incentive to keep their genomes concise is low. 50% of the human DNA is composed of repeats. Repeats may cause two far-away regions to be concatenated together, skipping the region in between during assembly.

There are many different types of repeats -- usually, a repeat is simply DNA that has learned to use the cellular machinery to copy (and transpose) itself.

- **Low-Complexity DNA** (e.g. ATATATATACATA...)
- **Microsatellite repeats** ($a_1 \dots a_k$)^N where $k \sim 3-6$
(e.g. CAGCAGTAGCAGCACCAG)
- **Transposons**
 - **SINE** (Short Interspersed Nuclear Elements)
e.g., ALU: ~300-long, 10^6 copies
 - **LINE** (Long Interspersed Nuclear Elements)
~4000-long, 200,000 copies
 - **LTR retroposons** (Long Terminal Repeats (~700 bp) at each end)
cousins of HIV
- **Gene Families** genes duplicate & then diverge (paralogs)
- **Recent duplications** ~100,000-long, very similar copies

A repeat will be passed onto future generations. Any mutation could either cause it to become more fit or less fit in copying itself → repeat will therefore “learn” to improve the replication of itself via evolution. An example of this is the ALU repeat family (specific to primates). Humans have about 1 million copies of the ALU repeat, which is 300 letters long. Some members of the ALU repeat are alive, while other members are dead. Entire family histories of the ALU repeat can be constructed.

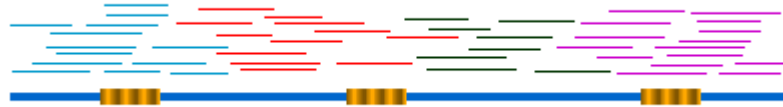
Question: If I gave you a single copy of the ALU repeat, how would you be able to tell whether the ALU repeat is alive or dead?

Answer: If a copy has an exact match somewhere else, the repeat is still alive and can produce "children".

3.5 Solving the Repeat Problem

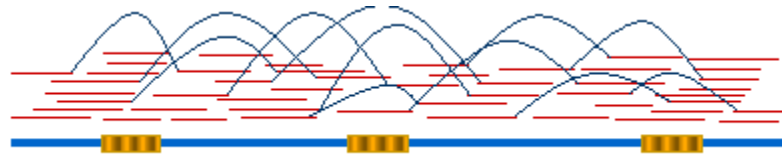
Method 1: Clustering the Reads

Cluster reads within small regions of the DNA that they come from → color them → reads within one cluster should, in theory, be repeat-free since each cluster has only 1 copy of the repeat)



Method 2: Linking the Reads

Every repeat should create a potential point that we can falsely join two or more distant regions of the DNA (every repeat is really a node that connects to many places in the DNA). Linked reads in the two ends of each fragment helps distinguish the order of fragment assembly.



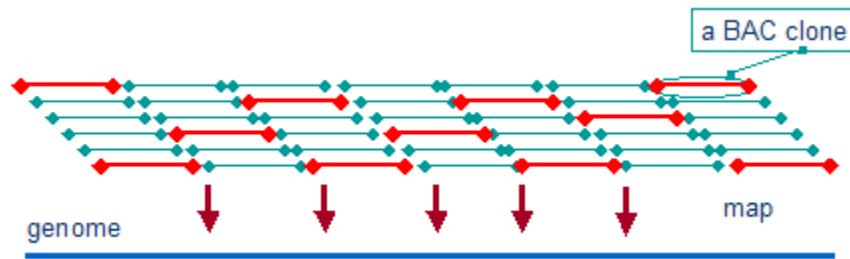
4. Strategies for Whole-Genome Sequencing

4.1 Hierarchical Clone-by-Clone (Clustering approach)

Yeast, human, worm, rat genomes have been sequenced using this approach.

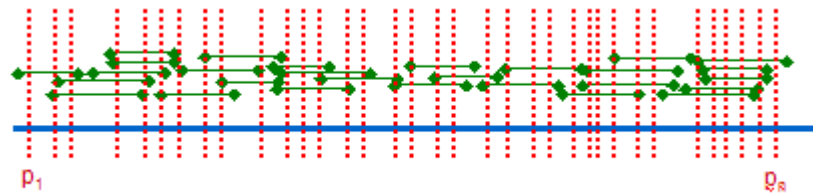
- 1) Break genome into many intermediate length pieces (150 bp or so)
- 2) Take each piece and clone into BAC's and create a library of BAC clones. The BAC clones should cover genome to a redundancy of 20-40x (in other words: the total length of clones is approx 20-40x the length of the original genome).
- 3) Map BACS into the original genome to get a physical map -- so that we know which one overlaps with another. There are two ways to do this: hybridization and digestion. More on these in the section below.
- 4) Select a minimum tiling path based on the physical mapping. This path is a minimum set of clones that cover the genome with smallest overlap between every successive clone. We want the tiling path to be minimum because we have to perform shotgun sequencing for each clone – expensive.

5) Put the shotgun derived sequences together based on the order established from the physical mapping



Hybridization

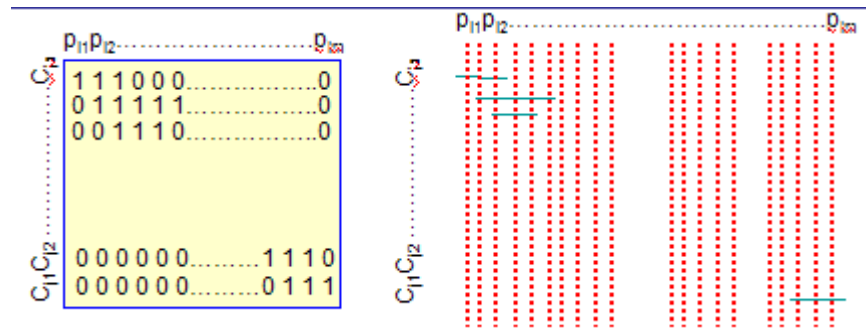
We take single-stranded versions of the clones and we hybridize small probes (8-10 nt) complementary to the clones. A probe that sticks to a particular position in the clone will also stick to all other clones sharing overlap with that position. After blasting many different probes into each of the clones, we can detect and define a set of hybridizing probes for each clone. An intersection of hybridizing probes between two clones represents an overlap between the two clones. The greater the intersection, the greater the overlap.



Computationally, hybridization approach can be resolved using a $m \times n$ probe-clone adjacency matrix, where $m = \# \text{probes}$ and $n = \# \text{clones}$. Matrix describes which probe hybridizes to which clone. Each matrix entry is either a 0 or 1.

$$\begin{aligned} (i, j) &= 1 && \text{if } p_i \text{ hybridizes to } C_j \\ (i, j) &= 0 && \text{otherwise} \end{aligned}$$

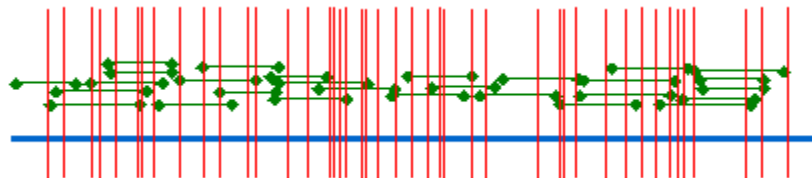
Now, the goal is to reorder the probes into their correct order in the genome. This basically means that for all the rows (clones) in the matrix, all its probes will be consecutive without any interfering zeros. So, we just need to rearrange the columns (probes) in order to get the matrix into consecutive-ones form (every row is of the format 000.....111.....000). The rearranged matrix will provide the correct ordering of the clones and show their overlap. The time complexity for this problem, assuming a consecutive-ones matrix exists, is $O(m^3)$.



A consecutive-ones matrix may not exist if you have problems such as: a probe sticking to > 1 location within the genome. In this case, the problem becomes finding the order of probes resulting in the minimal probe repetition (ie: shortest string of probes such that each clone is a substring). Greedy, probabilistic, as well as human-aided approaches have been applied to solve this more challenging problem. In addition, there may be experimental errors like probes not attaching where they're supposed to attach, probes nonspecifically attaching, etc.

Digestion

Cut the clones with restriction enzymes (sequence-specific scissors). The clones will all be cut in specific places according to their sequence. The product lengths are measured using gel electrophoresis. Two clones that overlap will contain many fragments of the same length. Now, using the overlap information we can infer the original ordering of the clones.



Typically, you cut with 2 restriction enzymes, a and b, as well as a combination of a + b. (Double Digestion)

Disadvantage of hierarchical clone-by-clone: laborious due to the number of biological experiments that have to be done in the physical mapping step. Attempts have been made to generate a more high-throughput method to sequence the genome.

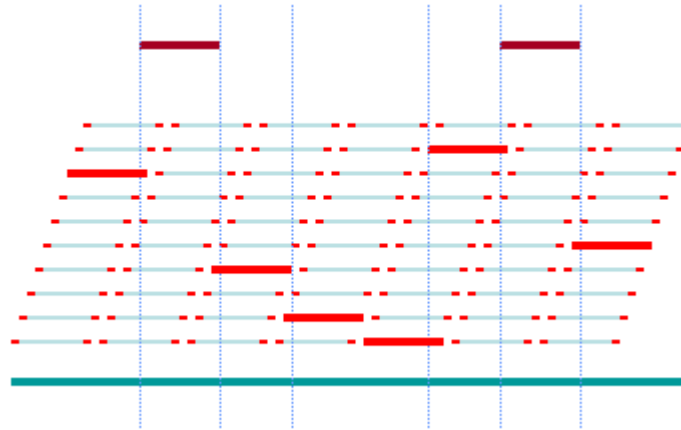
4.2 The Walking Method

The rice genome has been sequenced using this approach.

- 1) Build a highly redundant library of BACS and sequence just the ends of the clones
- 2) Fully sequence (shotgun) a few of the seed clones

3) Walking from the seed:

Using the library of paired ends of BACS, figure out which members of the BAC library will extend (to the left and to the right) the fully sequenced seed clones with minimum overlap. Apply BLAST → look to see which reads fall closest to the endpoints of the seed clone.

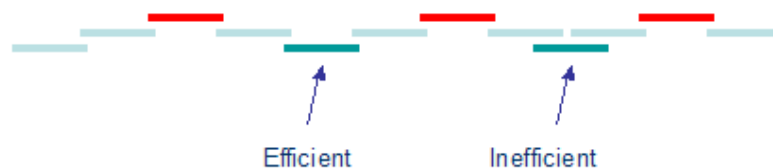


4) Continue to walk until the sequence of the entire genome is complete. If you pick too few original seeds, you will have to do many walking steps. Each walking step cycle spans about 3 months, so selecting the right number of seeds is important.

Walking Step:

1. Grow clone
2. Prepare & Shear DNA
3. Prepare shotgun library & perform shotgun
4. Assemble in a computer
5. Close remaining gaps

Generally speaking, a genome can be sequenced with <20 % redundancy if 5 walking steps are performed.



The walking method is much faster and cheaper than hierarchical approach because physical mapping is not needed. Also, walking from several seeds in parallel ensures that the amount of redundant sequencing is low. However, assembly of BACS and solving repeats are difficult in the walking method.

For whole genome sequencing, the current preferred protocol is Whole Genome Shotgun Sequencing (advocated by Celera), which will be discussed in the next lecture.