# CS262: Sequence Assembly

Samantha Zarate

February 16, 2016

**Fun tidbit:** Serafim attended the AGBT conference a few days before this class. Illumina announced a new sequencing machine that does less sequencing for more money and no drop in sequencing cost for this year.

# 1 Sequence Assembly

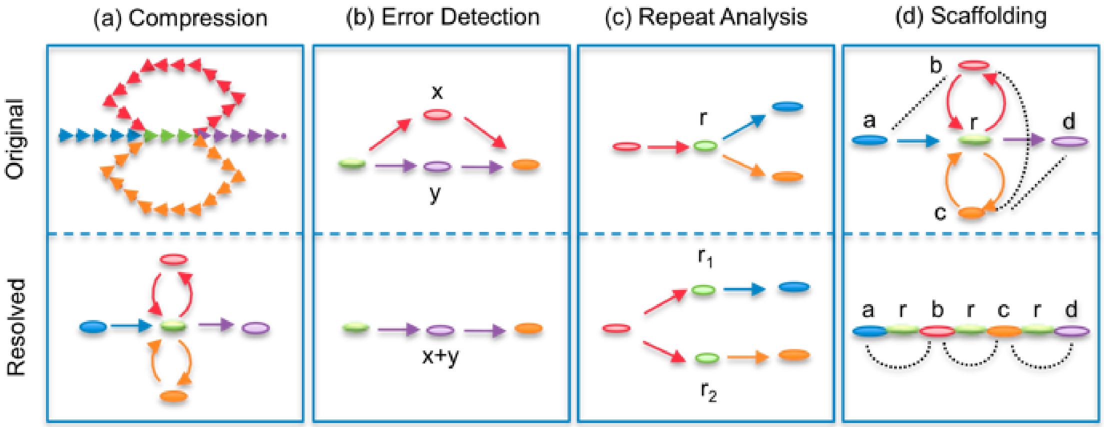
## 1.1 Approaches

The main problem with sequence assembly is that the genome is full of repeats (~ 55% of the genome), which especially causes problems when the repeat is highly similar and longer than the read length (~ $6 - 10\%$ of the genome). There are two main approaches:

- **Overlap-layout-consensus**: First, overlap reads. Then, we lay out the reads and combine those which overlap into contigs. Finally, we determine a consensus and therefore the final assembly. However, this can be computationally expensive, especially with large numbers of reads.
- **De Bruijn graphs**: A very simple and elegant alternative. $K$-mers occurring in the genome (nodes) have edges between them if they occur in reads and overlap by $k - 1$. Furthermore, edges are weighted according to the number of times they occur − if an overlap occurs in more than one read, the weight of the edge is incremented accordingly. A path that uses every edge once will also use every read, so if a path is found that does not comply, at least one read has an error of some kind: In other words, the De Bruijn graph contains all of the same information as the reads.
  There are some special things we can do with De Bruijn graphs:
  - **Compression**: If we have an especially long overlap demonstrated by consecutive edges connecting many reads, we can collapse the connected nodes into a single long contig and treat the contig as a single node.
  - **Error detection**: If there are two possible paths to a single node, we can examine the edge weights to determine which path is more likely. The less likely path is considered a "bubble" and can thus be "popped" or discounted entirely when correcting for errors.

- **Repeat analysis**: When there are many repeats (represented as one node) that have multiple options (outgoing edges), we can treat different instances of repeats differently. For example, if there is a `CAG` repeat and 50% of the time it is connected to `AGT` and the other 50% of the time it is connected to `AGA`, we can split the `CAG` node into `CAG1` that connects to `AGT` and `CAG2` that connects to `AGA`.
- **Scaffolding**: When structures are somewhat ambiguous, we can give approximate edges (denoted by dotted lines).

These concepts are demonstrated below:



**N50 contig length**: If we sort contigs from largest to smallest, and start covering the genome in that order, N50 is the length of the contig that just covers the 50th percentile. For a mouse, this is 26 kb.

## 1.2 A Brief History

The BGI (formerly known as Beijing Genome Institute) were essentially given an infinite budget, so they hired a bunch of undergrads to help the first conservation-motivated genome sequencing effort, using SOAP assemblers (short-read). It was also the first genome generated using Illumina (then Solexa) technology, using reads only 25 base pairs long.

An Assemblathon was held a few years ago to compare qualities of different assemblers with a known-to-be correct assembly. Sequence assembly is a very challenging problem; there have been many approaches taken, and quality can vary depending on the organism to which the genome belongs.

### 1.2.1 The Human Genome Project

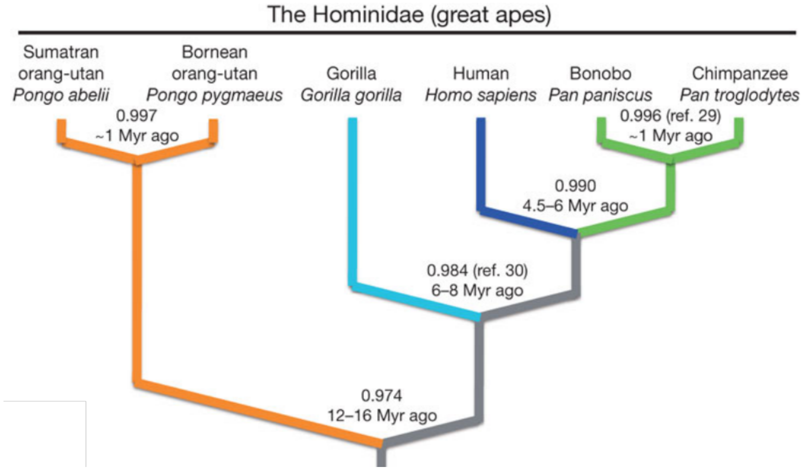
In 1997/8, Gene Myers suggested sequencing the full human genome using the whole-genome shotgun strategy, but Phil Green shot back with a review dismissing the concept as a bad idea (both papers were published in the same edition of Genome Review).

Craig Venter read both papers, decided to believe Gene Myers, and founded Celera to sequence the human genome using the shotgun strategy in competition with the government's pre-existing, slow-going effort. The private competitor placed pressure on the public effort to finish the project faster, and ultimately both the NIH and Celera finished in 2000.

# 2 Human Genome Diversity

## 2.1 Human Evolution

Humans diverged from our closest relatives (chimpanzees and bonobos) approximately 5 million years ago (see below for the ape "family tree"). Chimpanzees and bonobos, though relatively closely related and of similar intelligence, are very different. The former have a patriarchal society and are very aggressive, resolving conflicts through violence, while the latter have a matriarchal society, are nonviolent, and resolve conflicts primarily through sex. Interestingly, humans lie somewhere between those two extremes – what causes this genetically? In this context, the gorillas are the *outgroup*: they are not as closely related to humans, chimpanzees, and bonobos as they all are to each other.

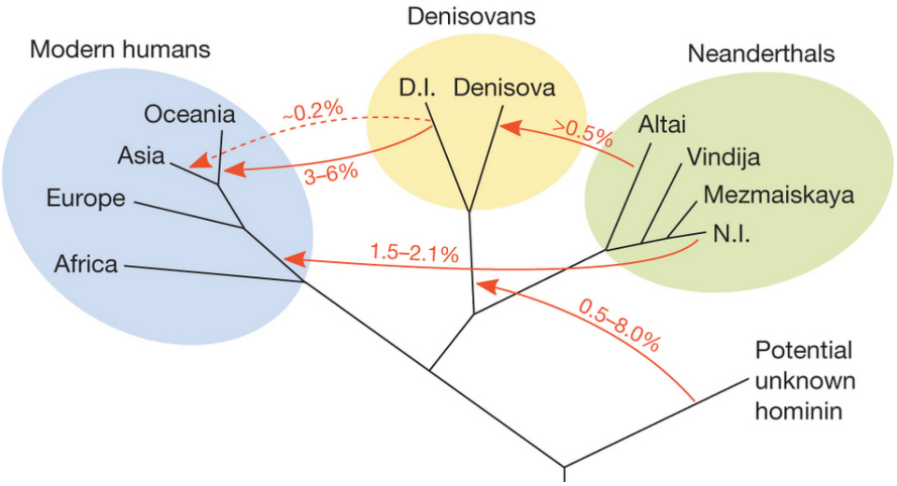

The Hominidae (great apes)
Sumatran orang-utan *Pongo abelii*
Bornean orang-utan *Pongo pygmaeus*
Gorilla *Gorilla gorilla*
Human *Homo sapiens*
Bonobo *Pan paniscus*
Chimpanzee *Pan troglodytes*
0.997 ~1 Myr ago
0.996 (ref. 29) ~1 Myr ago
0.990 4.5–6 Myr ago
0.984 (ref. 30) 6–8 Myr ago
0.974 12–16 Myr ago

### 2.1.1 Competing Theories

- **Out of Africa/Replacement**: There was a single mother of all humans (a so-called Eve) that existed approximately 99,000-150,000 years ago; and there was a single father of all humans (a so-called Adam) that existed approximately 120,000-340,000 years ago. Humans came out of Africa approximately 50,000 years ago and replaced other hominid species, e.g. the Neanderthals. Before coming out of Africa, humans were relatively homogeneous as a result of being a relatively small population confined to one area for so long; it was only after leaving Africa that human diversity blossomed.

- **Multiregional Evolution**: Humans independently evolved in different regions of the world. Though this is largely debunked, ~ 5% of the genomes of Europeans and Asians come from Neanderthals and Denisovans, other hominid species, demonstrating that different hominid species did in fact coexist and likely interbred.

The "family tree" of the Neanderthals, Denisovans, and modern humans is below.



## 2.2 Coalescence

A demonstrating example with men and Y chromosomes: Some men have brothers and some don't; regardless, all males have one father. If one constructs a family tree of all men, multiple leaf nodes (brothers) will *coalesce* into one shared parent node. This continues all the way up the family tree, through the generations, until we reach one man from which all human men today are descended ("Adam"). Though Adam was not the only male alive at the time, and other men certainly passed on different parts of their genetic material, only Adam's Y chromosome survives (albeit altered due to recombination).

This concept applies not only to Y chromosomes but also to every single specific region of the genome (and can definitively be proven with women as well using the same concept applied to mitochondrial DNA). Mutations accumulate over the generations to account for human diversity, and knowing the number of average mutations per generation as well as the two most different genomes, we can calculate how long ago "Adam" or "Eve" lived.

**Fun side note**: The coalescence of the HLA region of the genome (which determines genetic matching/rejection for organ donation dates back to 40 million years ago! Interestingly, this means that we could be closer to a monkey in the HLA region than to another human. Furthermore, males and females are attracted to each other based on the dissimilarity of the HLA regions (expressed through scent) because the offspring from such a match are more likely to have stronger immune systems.

## 2.3 Heterozygosity

Some definitions:

- **Alleles**: Different versions of genes, e.g. some people have T at a certain locus whereas others may have G. The *major* allele is more common than the *minor* allele.
- **Heterozygosity**: The probability that 2 alleles selected, at random, with replacement

are different.

$$H = \frac{4N\mu}{1 + 4N\mu}$$

- $H$ is heterozygosity
- $N$ is the effective population size
- $\mu$ is the per-generation mutation rate

Mutations introduce variation, and larger populations are more likely to allow mutations to persist. Typically, in humans, $4N\mu$ is quite small and so $H \approx 4N\mu$. Furthermore, $\mu \approx 10^{-8}$.

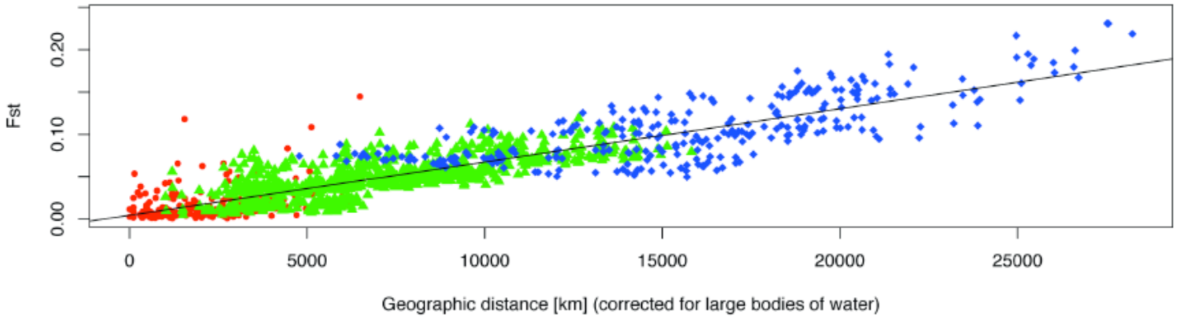
- **Recombination**: Rearrangements in genetic material in chromosomes, typically due to crossing over. Without at least one occurrence per chromosomes, meiosis wouldn't work; however, this is overall a relatively rare occurrence, occurring on ~ 1/100 Mb.
- **Linkage Disequilibrium**: The degree of correlation between two SNP locations. Often, allele association is not random due to large blocks between recombination sites.

### 2.3.1 Fall in Heterozygosity

$$F_{ST} = \frac{H - H_{POP}}{H}$$

- $F_{ST}$ is a measure of similarity within a population
- $H$ is heterozygosity; $H_{POP}$ is heterozygosity within a population

If we plot $F_{ST}$ as a function of geographic distance from sub-Saharan Africa (see below), we see a perfect linear correlation: The farther a population is from sub-Saharan Africa, there is a linear *decrease* in heterozygosity; the more similar they are to each other – they lose variation over time! Not enough mutations have occurred over this time period to compensate.



# 3 References

- Overlap Layout Consensus assembly notes by Ben Langmead, Johns Hopkins.
- Hi-res De Bruijn graph image used in lecture
- GenomeWeb article on the panda genome sequencing effort
- All other information and images is from the lecture audio and notes.