Genome Assembly

Lecture 18
Oct 14, 2016

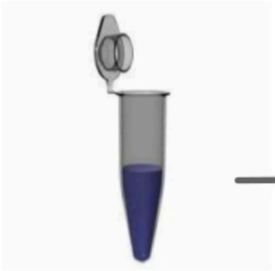
Announcements

Why do you want to Assemble a genome?

What do you need to Assemble a genome?

Assemble a genome?

In an ideal world ...



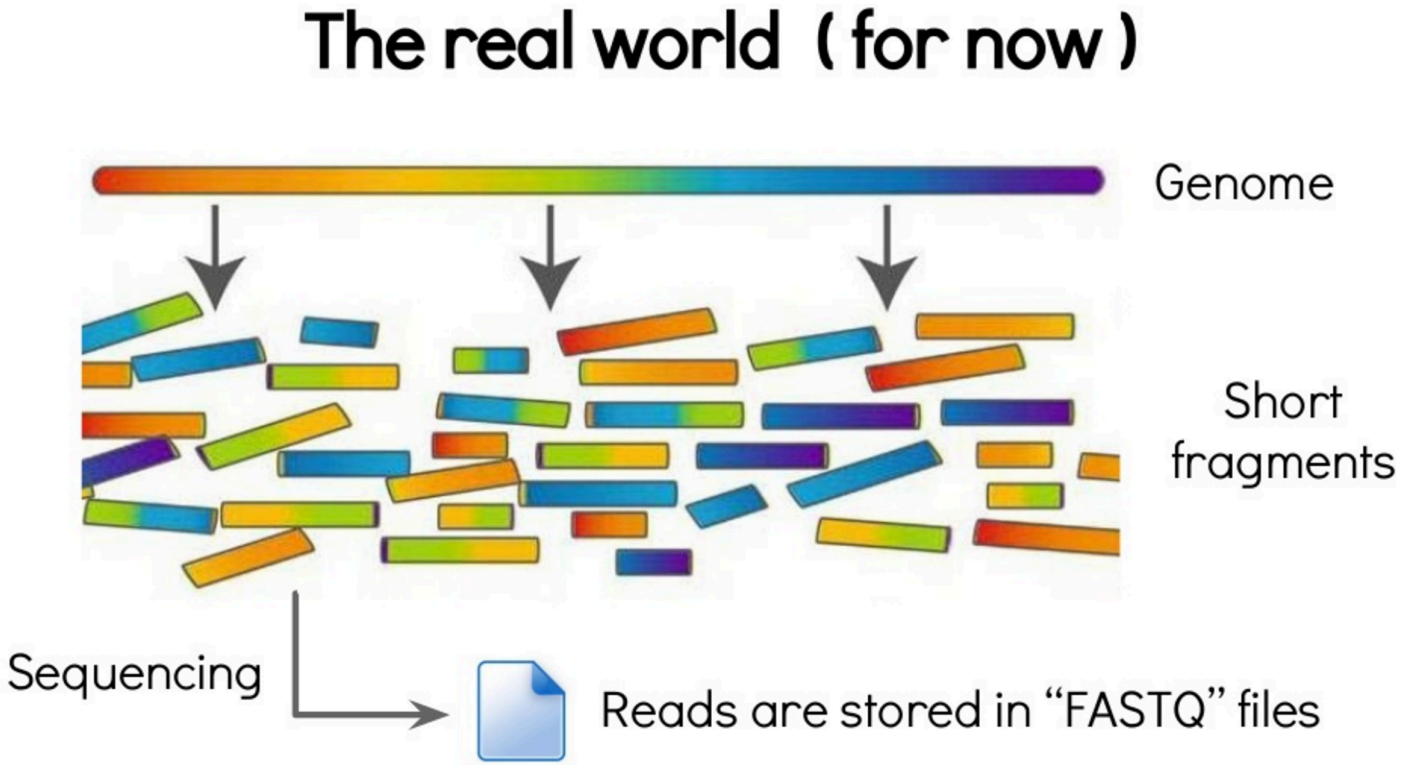
Human DNA

iSequencer ™

```
AGTCTAGGATTCGCTATAG
ATTCAGGCTCTGATATATT
TCGCGGCATTAGCTAGAGA
TCTCGAGATTCGTCCCAGT
CTAGGATTCGCTAT
AAGTCTAAGATTC...
```

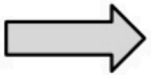
46 chromosomal and1 mitochondrial sequences

Assemble a genome?



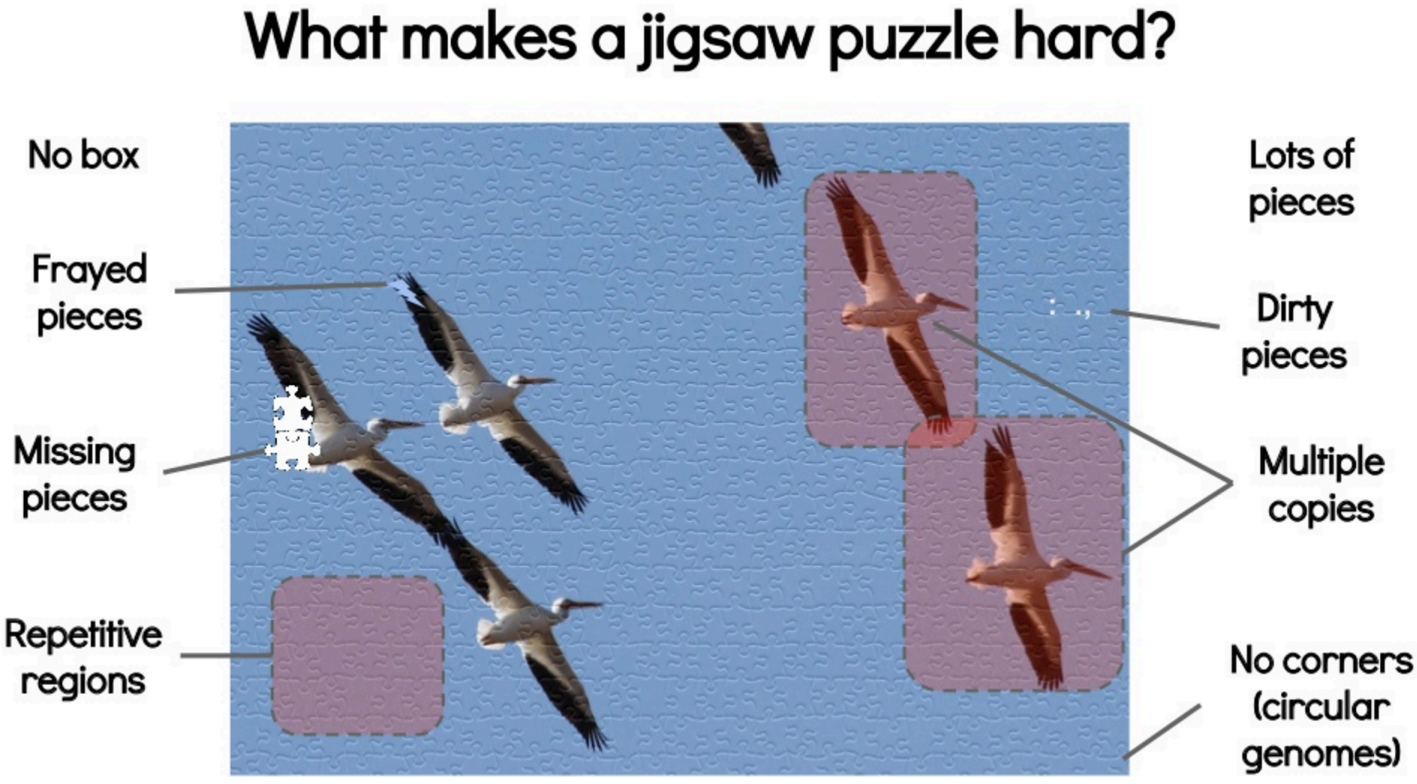
Assemble a genome?

De novo genome assembly

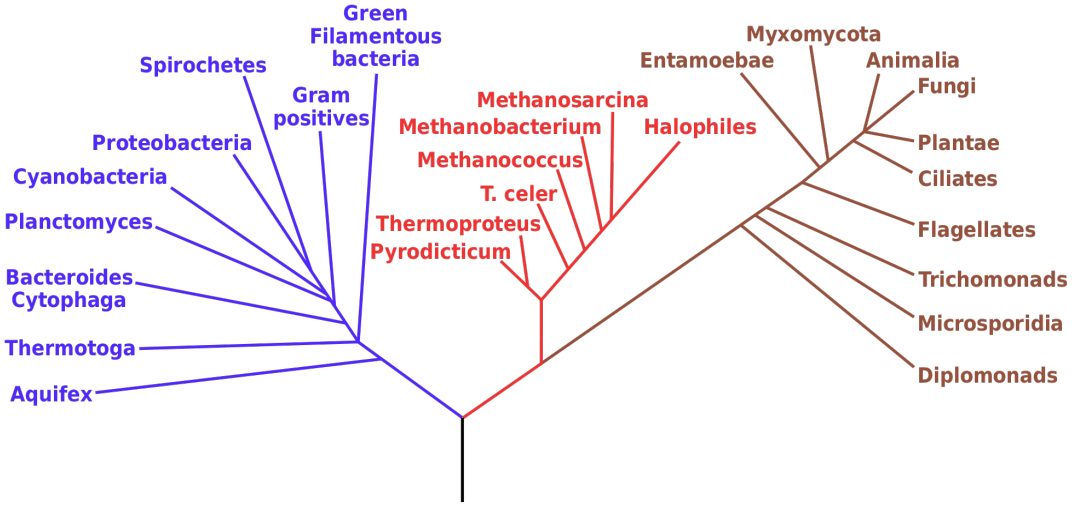


"From scratch"

Assemble a genome?



Genome Sizes



Assemble a genome? General Strategies

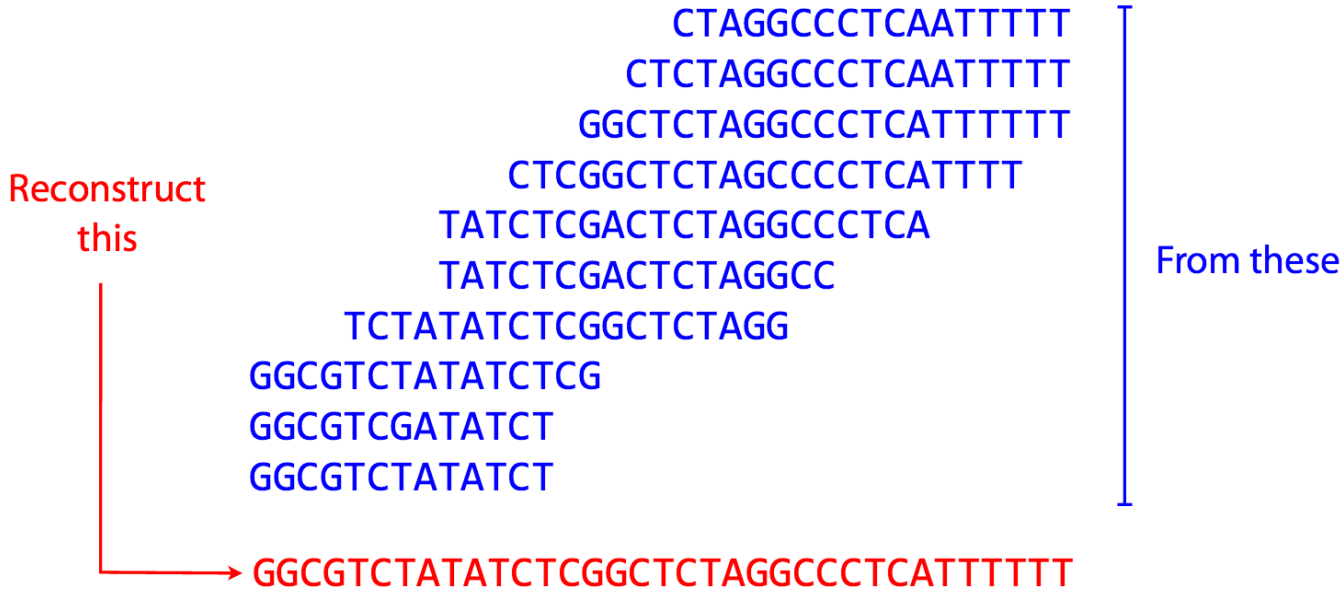
Genome size	Unlimited $$	Typical
>10Mb		
10Mb - 500Mb		
> 500 Mb		

Assembly

- Work flow:

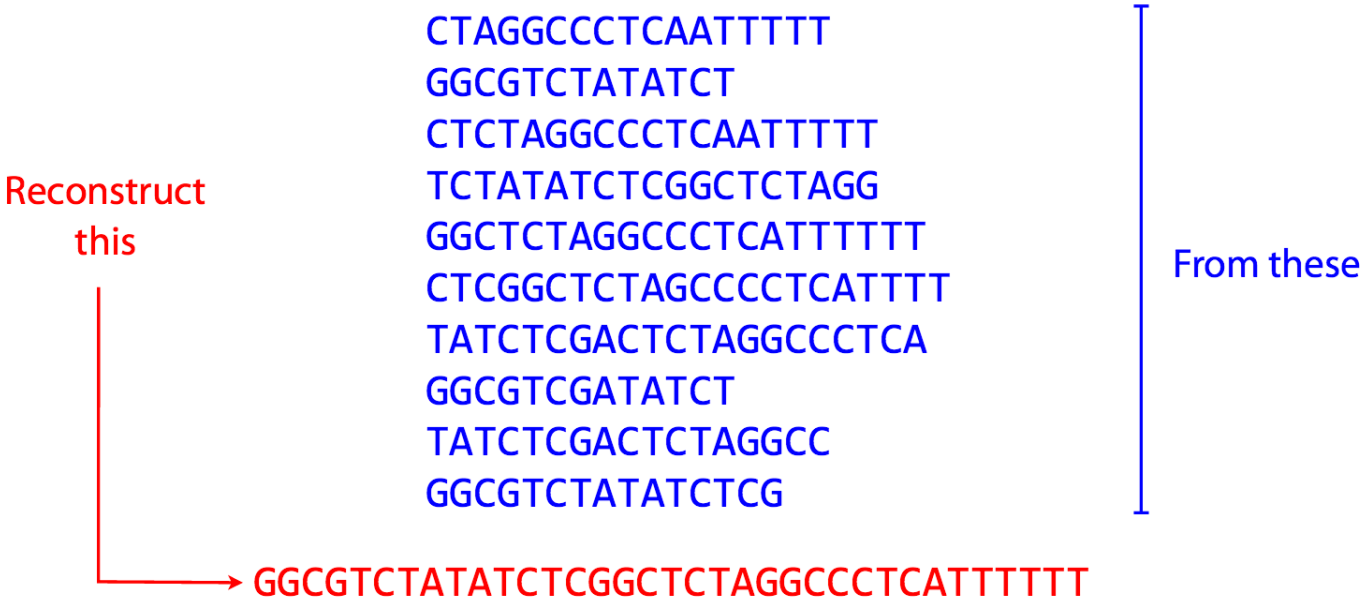
Assembly

Assume sequencing produces such a large # fragments that almost all genome positions are *covered* by many fragments...



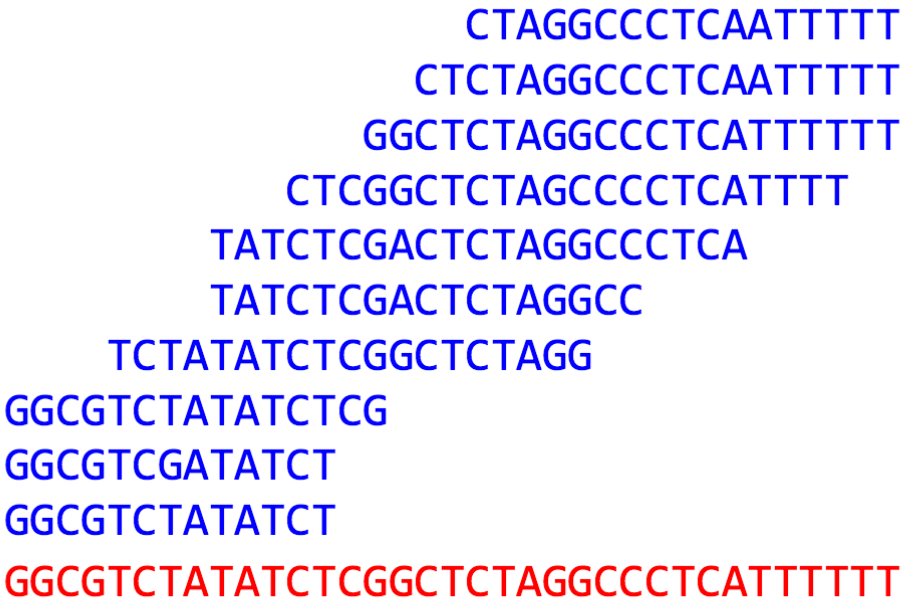
Assembly

...but we don't know what came from where



Assembly

Key term: *coverage*. Usually it's short for *average coverage*: the average number of reads covering a position in the genome.



177 nucleotides

35 nucleotides

Average coverage = 177 / 35 ≈ 7x

Other Assembly Terms

Unitig

Contig

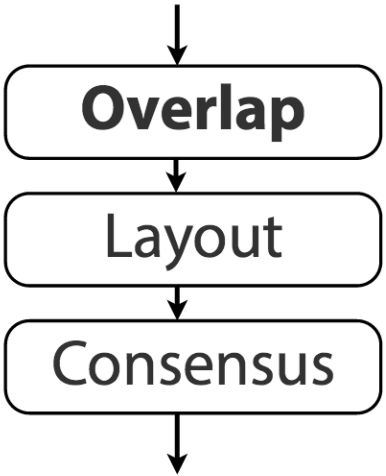
scaffold

Assembly

- 3 assembly strategies:

Assembly

- OLC Assembly



Build overlap graph

Bundle stretches of the overlap graph into *contigs*

Pick most likely nucleotide sequence for each contig

Assembly

- OLC Assembly: Characteristics

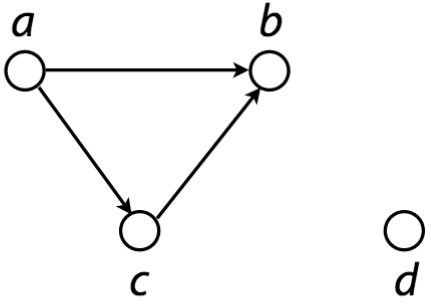
Assembly

Directed graph $G(V, E)$ consists of set of *vertices, V* and set of *directed edges, E*

Directed edge is an *ordered pair* of vertices.
First is the *source*, second is the *sink*.

Vertex is drawn as a circle

Edge is drawn as a line with an arrow connecting two circles



Vertex also called *node* or *point*

Edge also called *arc* or *line*

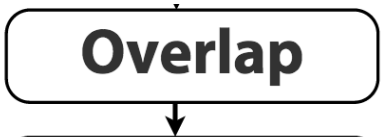
Directed graph also called *digraph*

$V = \{ a, b, c, d \}$

$E = \{ (a, b), (a, c), (c, b) \}$

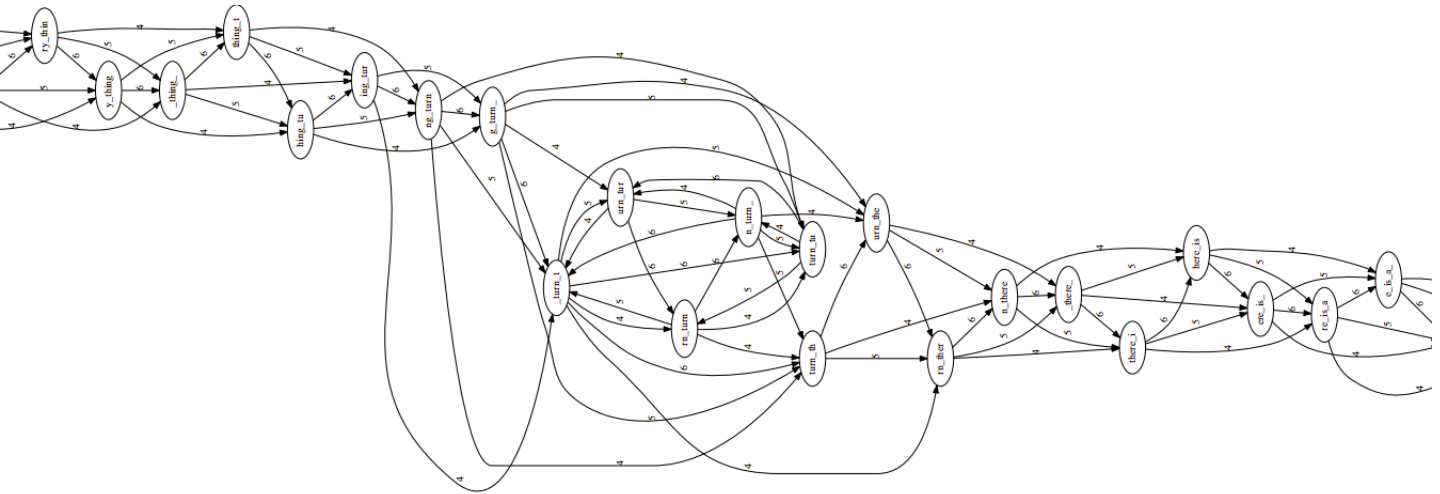
Source Sink

Assembly

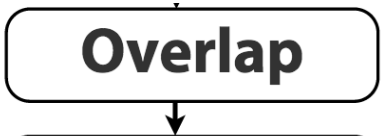


Build overlap graph

to_every_thing_turn_turn_turn_there_is_a_season
L=4, k=7



Assembly



Build overlap graph

Vertices (reads): { *a:* CTCTAGGCC, *b:* GCCCTCAAT, *c:* CAATTTTT }

Edges (overlaps): { (*a*, *b*), (*b*, *c*) }