

# **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

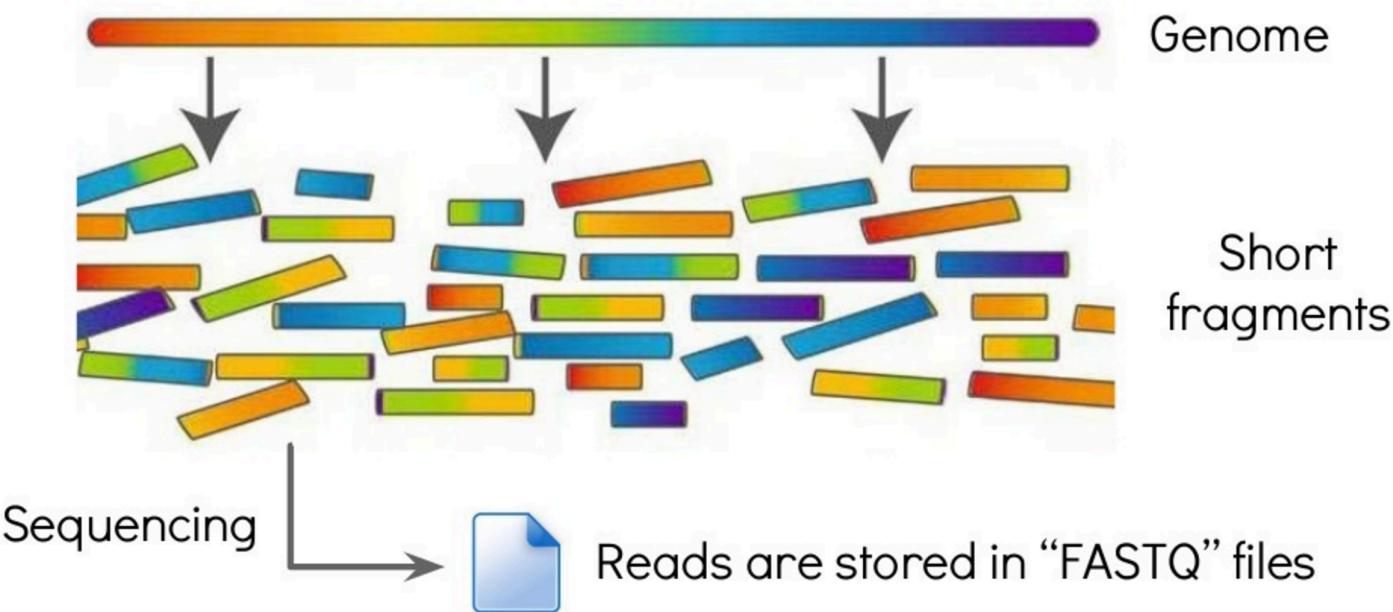
iSequencer™

AGTCTAGGATTCGCTATAG  
ATTCAGGCTCTGATATATT  
TCGCAGCATTAGCTAGAGA  
TCTCGAGATTGTCCCAGT  
CTAGGATTGCTAT  
AAGTCTAAGATTCAAG...  
...

46 chromosomal  
and 1 mitochondrial  
sequences

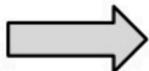
# ASSEMBLE A GENOME?

The real world (for now)



# ASSEMBLE A GENOME?

## *De novo* genome assembly



“From scratch”

# ASSEMBLE A GENOME?

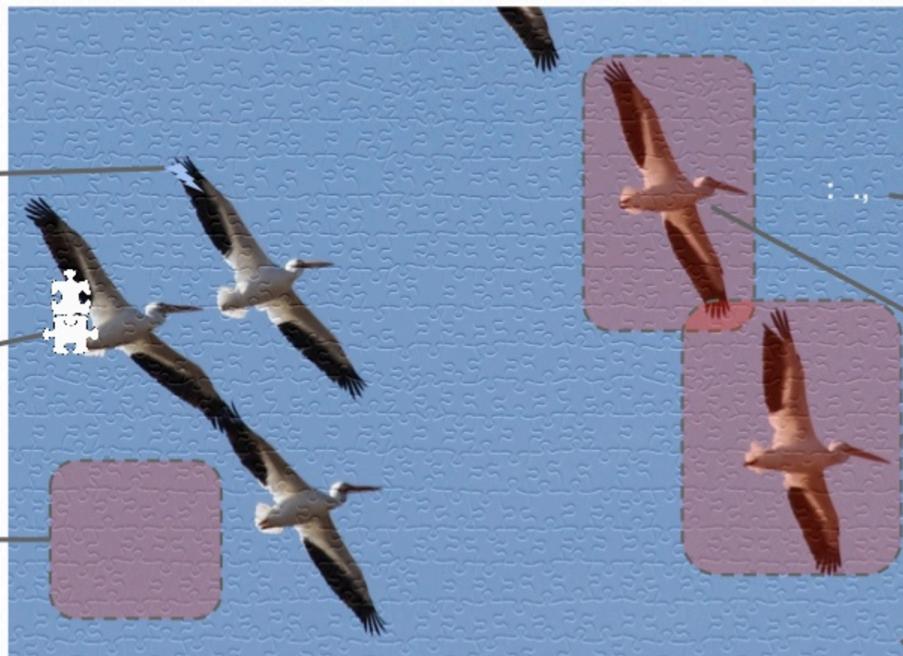
## What makes a jigsaw puzzle hard?

No box

Frayed pieces

Missing pieces

Repetitive regions



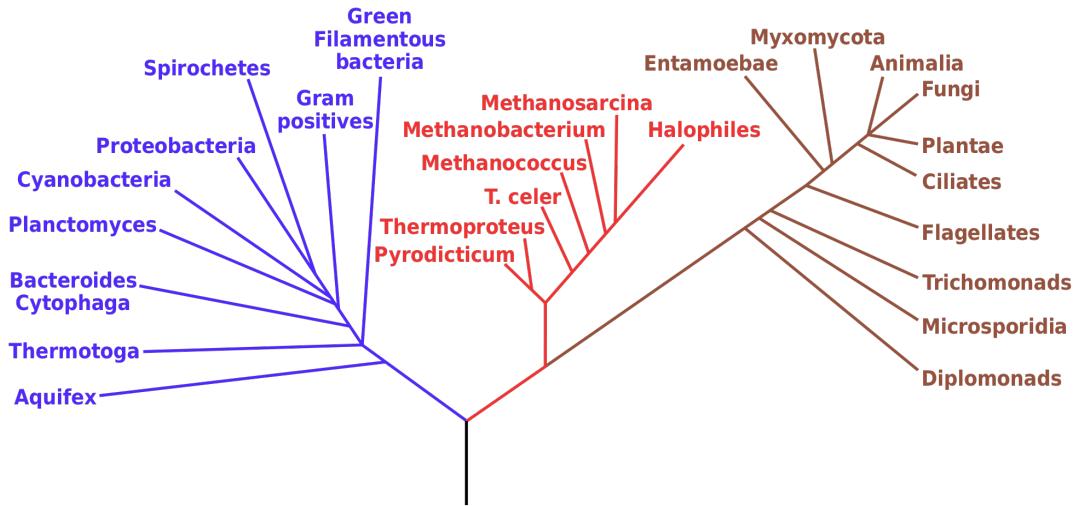
Lots of pieces

Dirty pieces

Multiple copies

No corners (circular genomes)

# 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...

Reconstruct  
this

CTAGGCCCTCAATTTT  
CTCTAGGCCCTCAATTTT  
GGCTCTAGGCCCTCATT  
CTCGGCTCTAGCCCCTCATT  
TATCTCGACTCTAGGCCCTCA  
TATCTCGACTCTAGGCC  
TCTATATCTCGGCTCTAGG  
GGCGTCTATATCTCG  
GGCGTCGATATCT  
GGCGTCTATATCT  
→ GGCGTCTATATCTCGGCTCTAGGCCCTCATT

From these

# ASSEMBLY

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

Reconstruct  
this

CTAGGCCCTCAATTTT  
GGCGTCTATATCT  
CTCTAGGCCCTCAATTTT  
TCTATATCTGGCTCTAGG  
GGCTCTAGGCCCTCATTTTT  
CTCGGCTCTAGGCCCTCATTTT  
TATCTCGACTCTAGGCCCTCA  
GGCGTCGATATCT  
TATCTCGACTCTAGGCC  
GGCGTCTATATCTCG

From these

→ GGCGTCTATATCTGGCTCTAGGCCCTCATTTTT

# ASSEMBLY

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

CTAGGCCCTCAATTTT	
CTCTAGGCCCTCAATTTT	
GGCTCTAGGCCCTCATTTTT	
CTCGGCTCTAGCCCCTCATTTT	
TATCTCGACTCTAGGCCCTCA	177 nucleotides
TATCTCGACTCTAGGCC	
TCTATATCTCGGCTCTAGG	
GGCGTCTATATCTCG	
GGCGTCGATATCT	
GGCGTCTATATCT	
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTT	35 nucleotides

$$\text{Average coverage} = 177 / 35 \approx 7x$$

# **OTHER ASSEMBLY TERMS**

Unitig

Contig

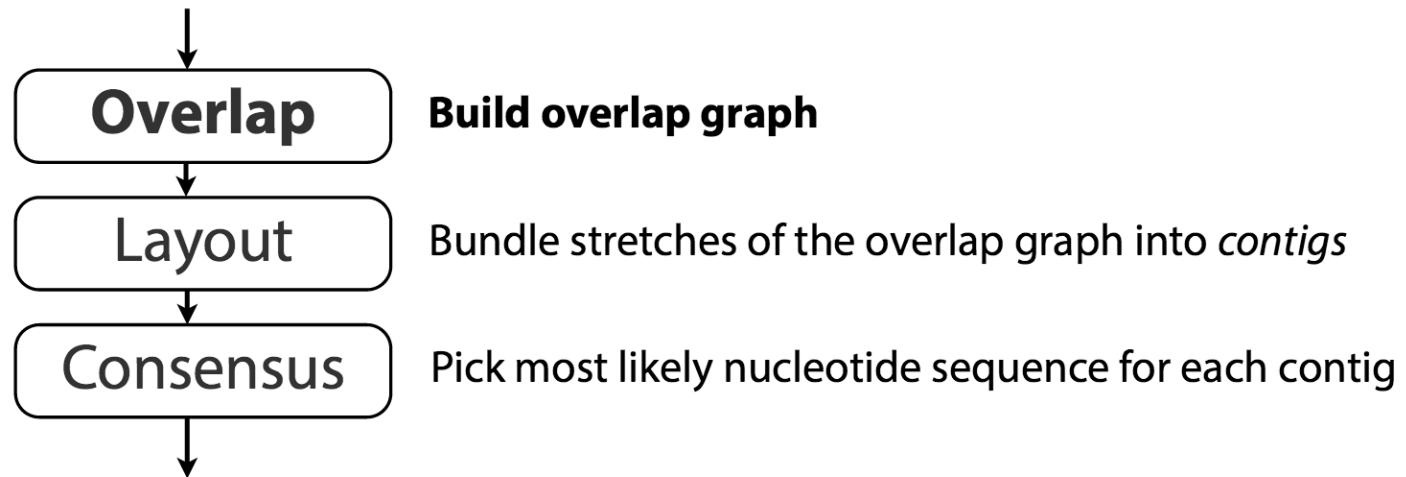
scaffold

# **ASSEMBLY**

- 3 assembly strategies:

# ASSEMBLY

- OLC Assembly



# **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*.

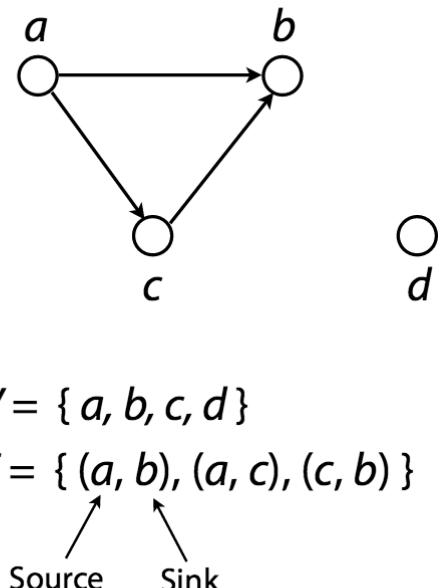
Vertex is drawn as a circle

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*

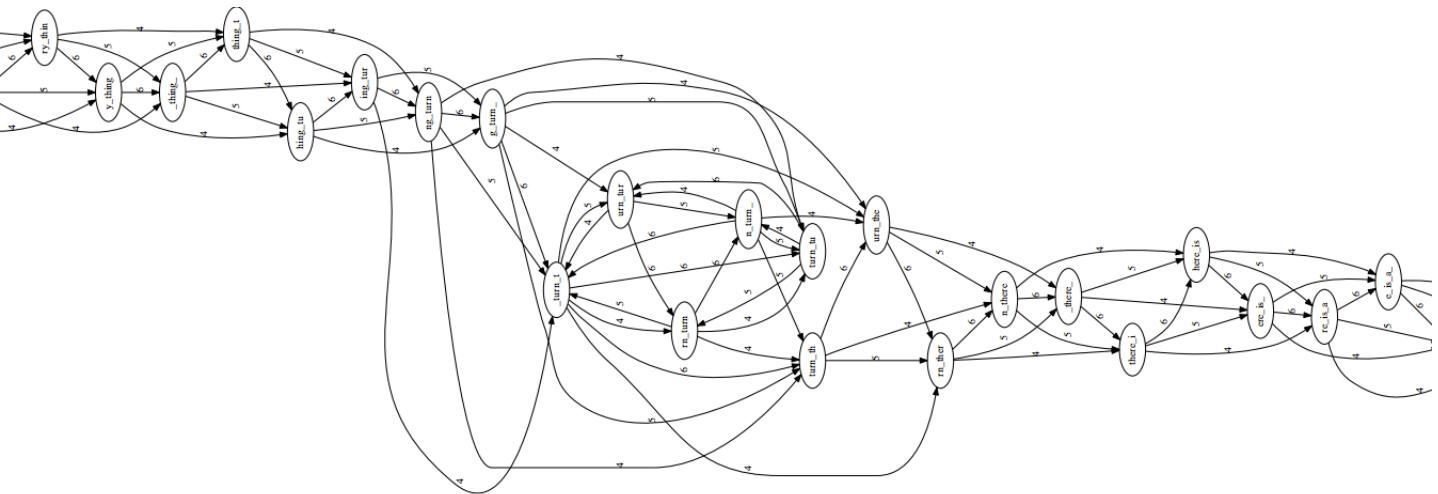


# ASSEMBLY

## Overlap

Build overlap graph

to\_every\_thing\_turn\_turn\_turn\_there\_is\_a\_season  
 $L=4, k=7$



# ASSEMBLY

## Overlap

Build overlap graph



Vertices (reads): {  $a: \text{CTCTAGGCC}$ ,  $b: \text{GCCCTCAAT}$ ,  $c: \text{CAATTTTT}$  }

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

$a: \text{CTCTAGGCC}$

3

$b: \text{GCCCTCAAT}$

4

$c: \text{CAATTTTT}$

CTCTAGGCC

|||

GCCCTCAAT

GCCCTCAAT

||||

CAATTTTT