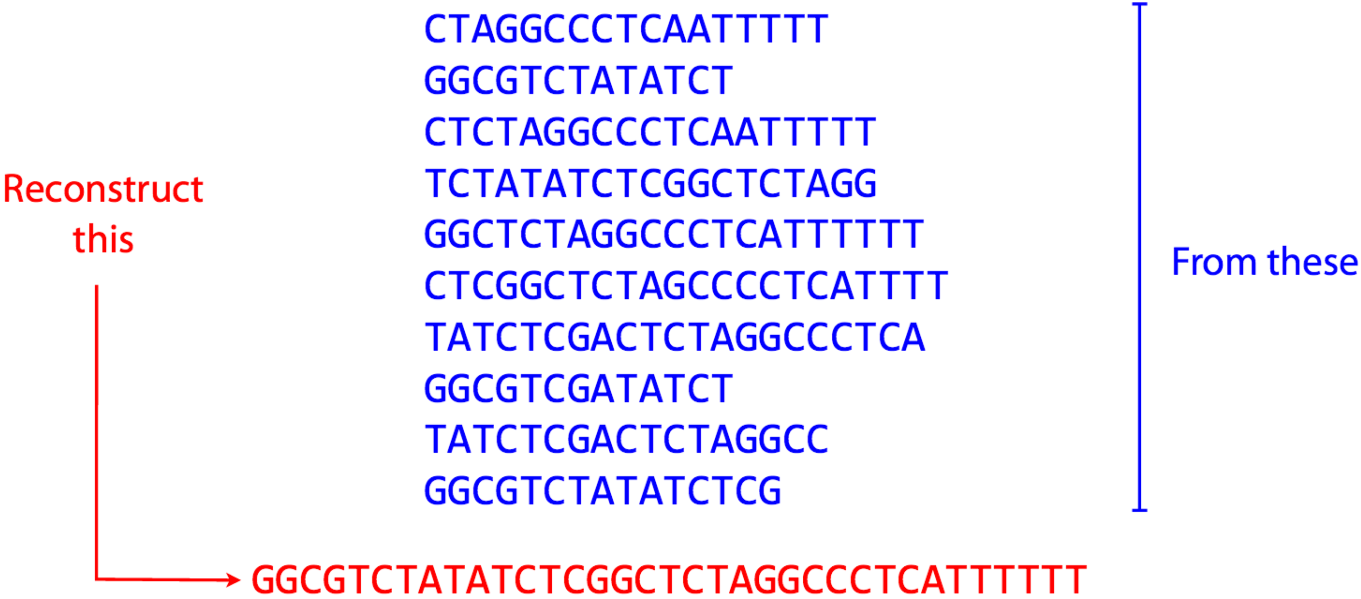
Genome Assembly

Lecture 19
Oct 17, 2016

Announcements

Assembly

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



CTAGGCCCTCAATTTTT
GGCGTCTATATCT
CTCTAGGCCCTCAATTTTT
TCTATATCTCGGCTCTAGG
GGCTCTAGGCCCTCATTTTTT
CTCGGCTCTAGCCCCTCATTTT
TATCTCGACTCTAGGCCCTCA
GGCGTCGATATCT
TATCTCGACTCTAGGCC
GGCGTCTATATCTCG

From these

Reconstruct this

GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT

First law of assembly

If a suffix of read A is similar to a prefix of read B...

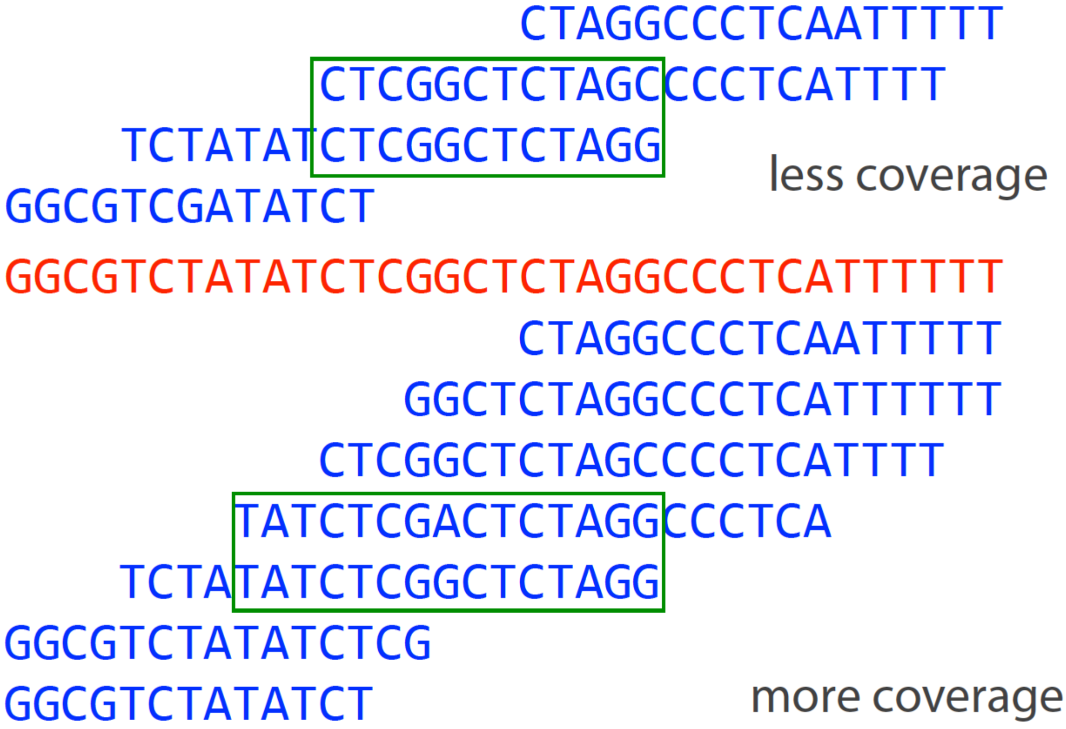
```
TCTATATCTCGGCTCTAGG
    ||||||| |||||||
    TATCTCGACTCTAGGCC
```

...then A and B might *overlap* in the genome

```
    TCTATATCTCGGCTCTAGG
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
        TATCTCGACTCTAGGCC
```

Second law of assembly

More coverage leads to more and longer overlaps



```
                  CTAGGCCCTCAATTTTT
           CTCGGCTCTAGCCCCTCATTTT
    TCTATATCTCGGCTCTAGG
GGCGTCGATATCT
```

less coverage

```
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
```

```
                  CTAGGCCCTCAATTTTT
              GGCTCTAGGCCCTCATTTTTT
           CTCGGCTCTAGCCCCTCATTTT
        TATCTCGACTCTAGGCCCTCA
    TCTATATCTCGGCTCTAGG
GGCGTCTATATCTCG
GGCGTCTATATCT
```

more coverage

Assembly

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

```
                  CTAGGCCCTCAATTTTT
                CTCTAGGCCCTCAATTTTT
              GGCTCTAGGCCCTCATTTTTT
           CTCGGCTCTAGCCCCTCATTTT
        TATCTCGACTCTAGGCCCTCA           177 nucleotides
        TATCTCGACTCTAGGCC
    TCTATATCTCGGCTCTAGG
GGCGTCTATATCTCG
GGCGTCGATATCT
GGCGTCTATATCT
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT     35 nucleotides
```

Average coverage = 177 / 35 ≈ 7x

Other Assembly Terms

Unitig

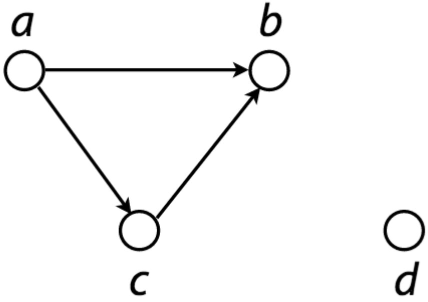
Contig

scaffold

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*

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

$E = \{ (\underset{\text{Source}}{\underset{\nearrow}{a}}, \underset{\text{Sink}}{\underset{\nwarrow}{b}}), (a, c), (c, b) \}$

Edge also called *arc* or *line*

Directed graph also called *digraph*

ASSEMBLY

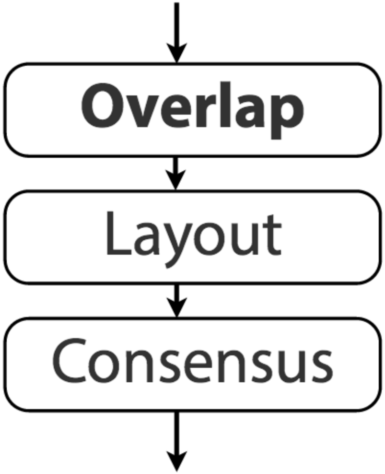
- 2 assembly strategies:

Assembly

- OLC Assembly: Characteristics

ASSEMBLY

- OLC Assembly



Overlap **Build overlap graph**

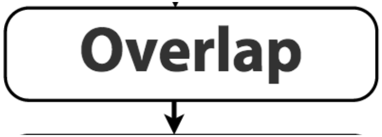
Layout Bundle stretches of the overlap graph into *contigs*

Consensus Pick most likely nucleotide sequence for each contig

Assembly

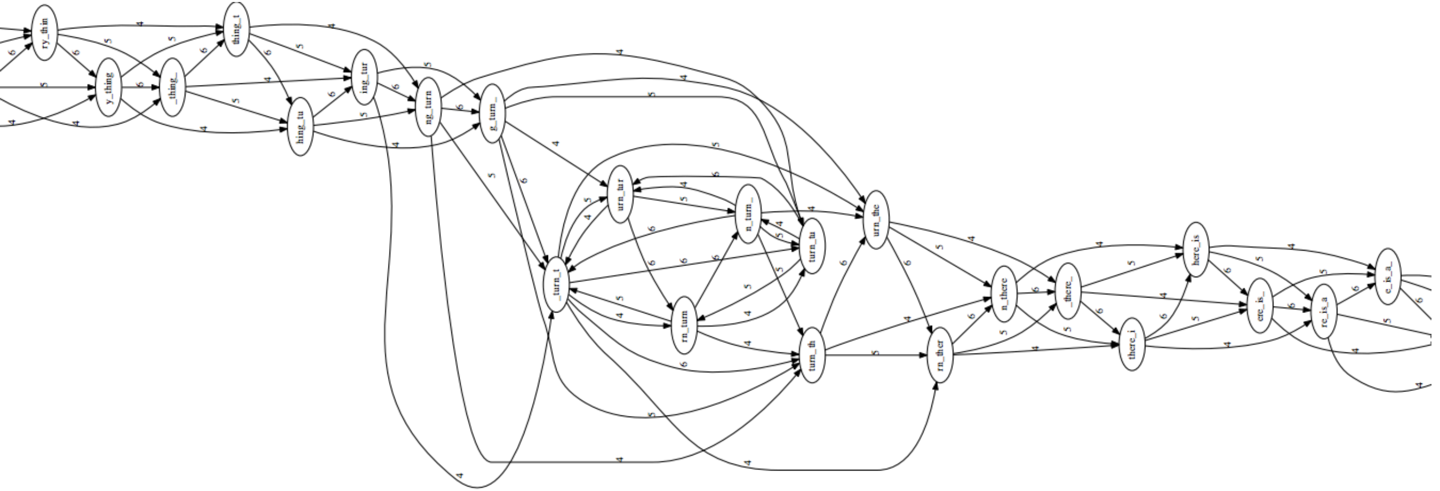
https://youtu.be/yPJ7yHRk2OI

ASSEMBLY

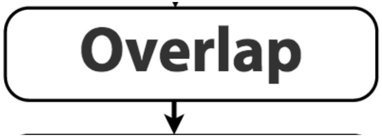


Build overlap graph

to_every_thing_turn_turn_turn_there_is_a_season
L=4, k=7



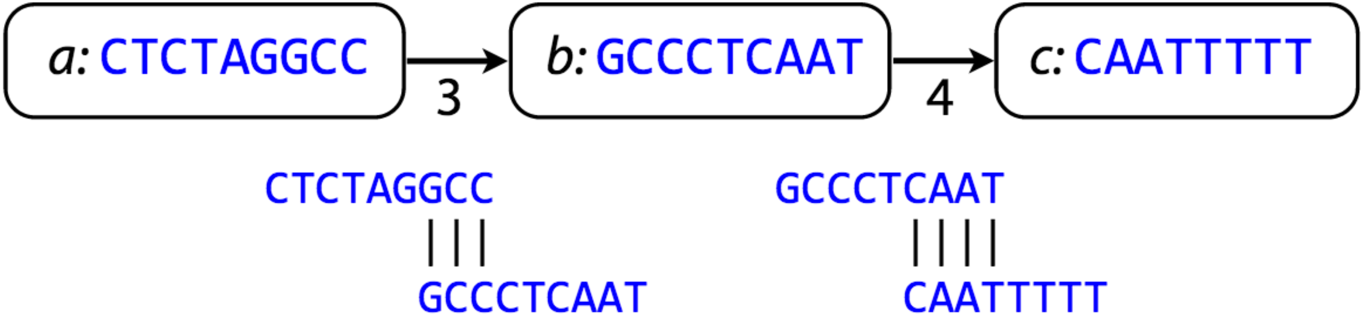
ASSEMBLY



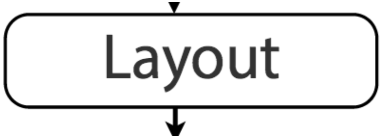
Build overlap graph

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

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



ASSEMBLY - OLC

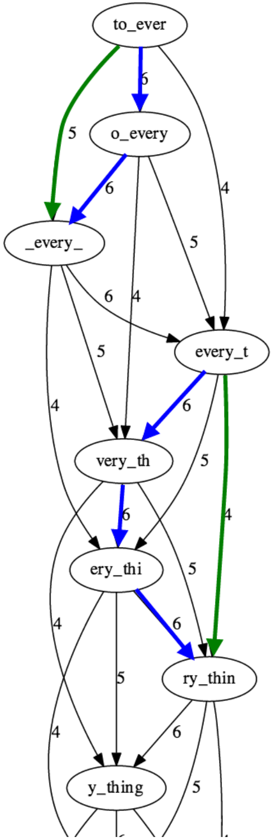


Bundle stretches of the overlap graph into *contigs*

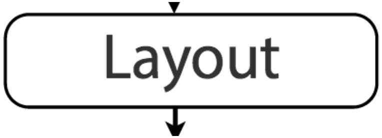
Anything redundant about this part of the overlap graph?

Some edges can be *inferred* (*transitively*) from other edges

E.g. green edge can be inferred from blue



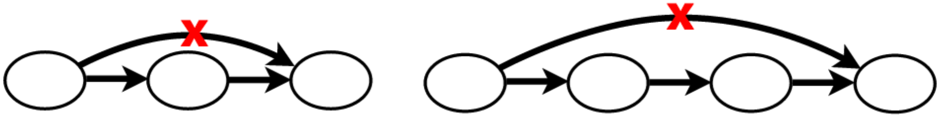
ASSEMBLY - OLC



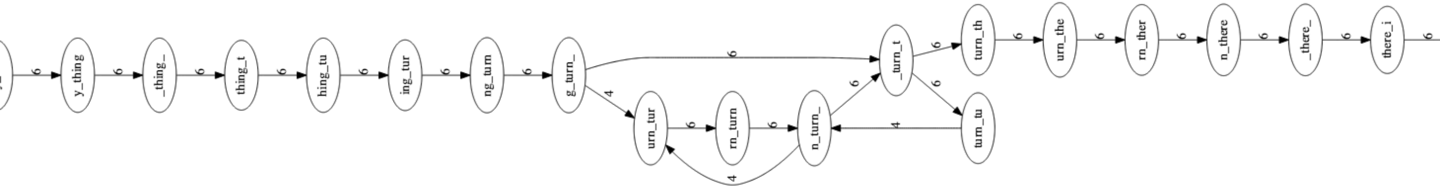
Layout

Bundle stretches of the overlap graph into *contigs*

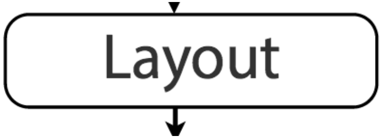
Remove transitively-inferrible edges, starting with edges that skip one *or two* nodes:



After:

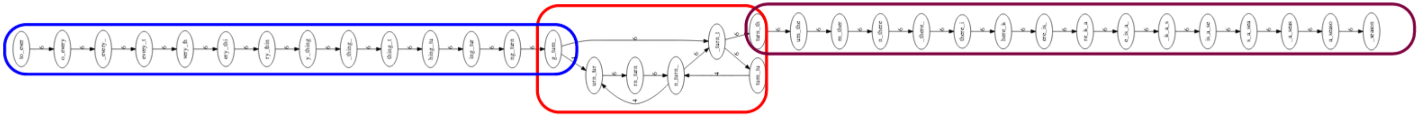


Assembly - OLC



Bundle stretches of the overlap graph into *contigs*

Emit *contigs* corresponding to the non-branching stretches



Contig 1
`to_every_thing_turn_`

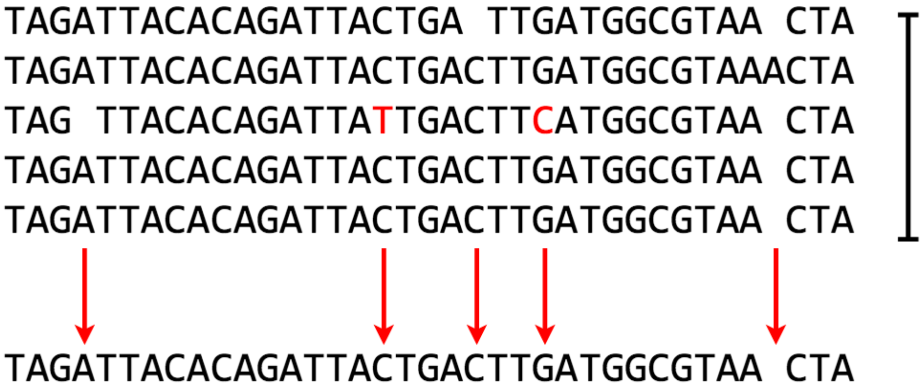
Contig 2
`turn_there_is_a_season`

Unresolvable repeat

Assembly - OLC

Consensus

Pick most likely nucleotide sequence for each contig



```
TAGATTACACAGATTACTGA TTGATGGCGTAA CTA
TAGATTACACAGATTACTGACTTGATGGCGTAAACTA
TAG TTACACAGATTATTGACTTCATGGCGTAA CTA
TAGATTACACAGATTACTGACTTGATGGCGTAA CTA
TAGATTACACAGATTACTGACTTGATGGCGTAA CTA

TAGATTACACAGATTACTGACTTGATGGCGTAA CTA
```

Take reads that make up a contig and line them up

Take *consensus*, i.e. majority vote

At each position, ask: what nucleotide (and/or gap) is here?

Complications: (a) sequencing error, (b) ploidy

Say the true genotype is AG, but we have a high sequencing error rate and only about 6 reads covering the position.