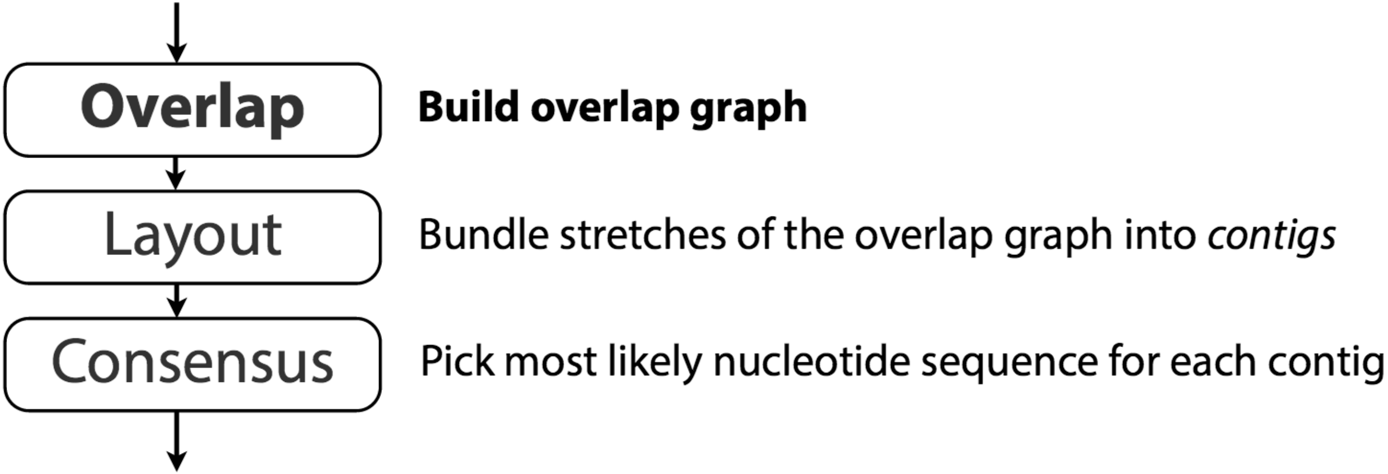
Review:

Assemble a genome? General Strategies

Genome size	Unlimited $$	Typical
>10Mb		
10Mb - 100Mb		
> 100 Mb		

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

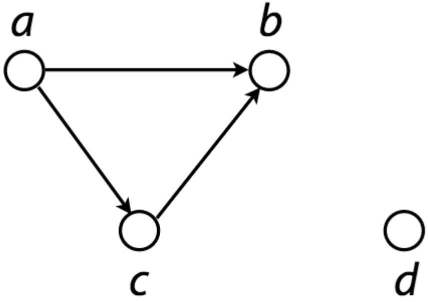
ASSEMBLY

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*

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

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

Assembly – De Bruijn

Hamiltonian Path Problem

Eulerian Path Problem

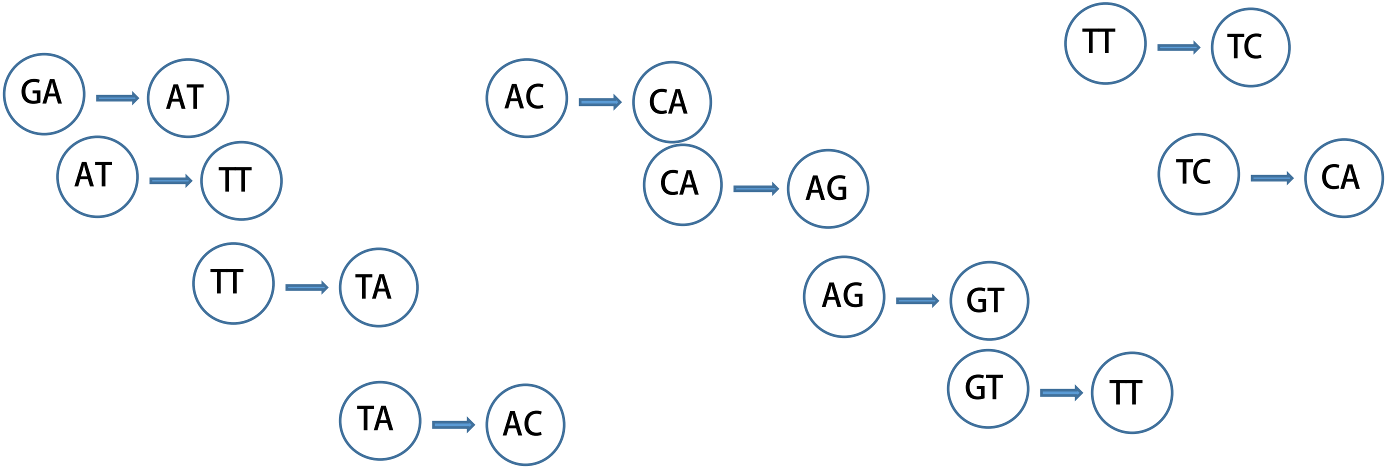
Assembly – De Bruijn

```
GATTAC
GAT
 ATT
  TTA
   TAC
```

```
ACAGTTCA
ACA
 CAG
  AGT
   GTT
    TTC
     TCA
```

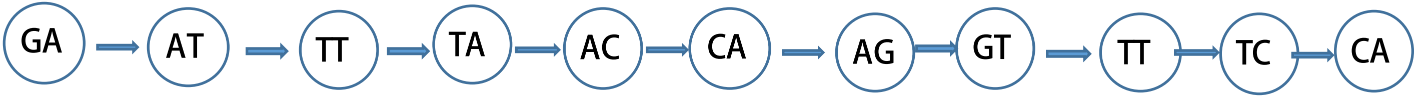
Assembly – De Bruijn

GAT ATT TTA TAC ACA CAG AGT GTT TTC TCA



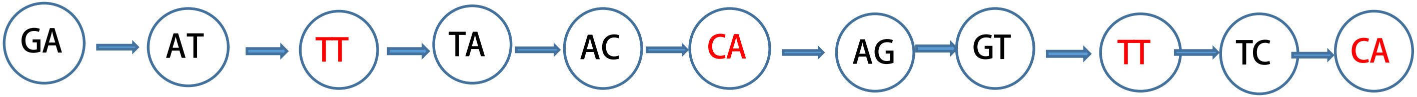
Assembly – De Bruijn

GAT ATT TTA TAC ACA CAG AGT GTT TTC TCA



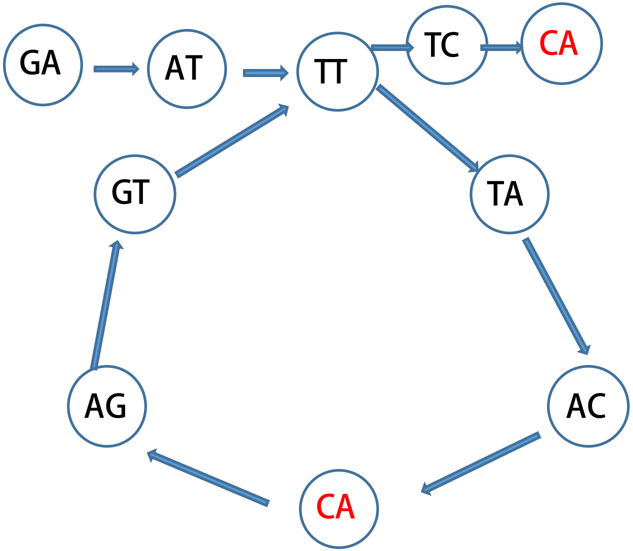
Assembly – De Bruijn

GAT ATT TTA TAC ACA CAG AGT GTT TTC TCA



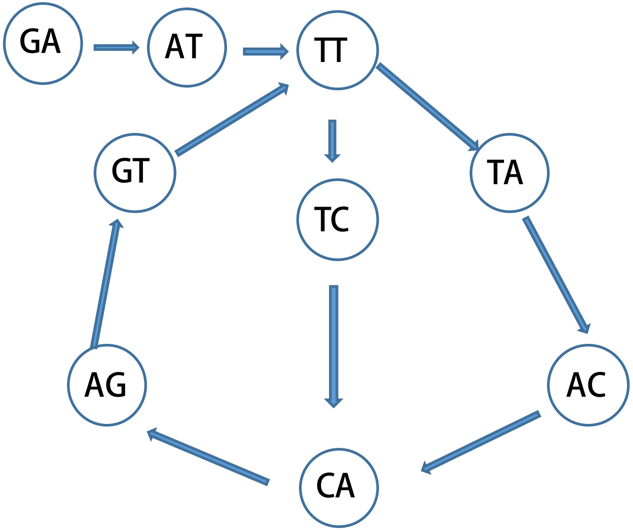
Assembly – De Bruijn

GAT ATT TTA TAC ACA CAG AGT GTT TTC TCA



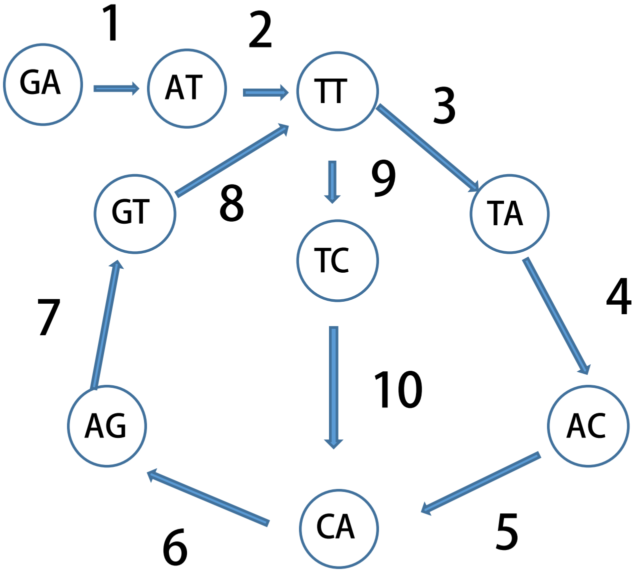
Assembly – De Bruijn

GAT ATT TTA TAC ACA CAG AGT GTT TTC TCA



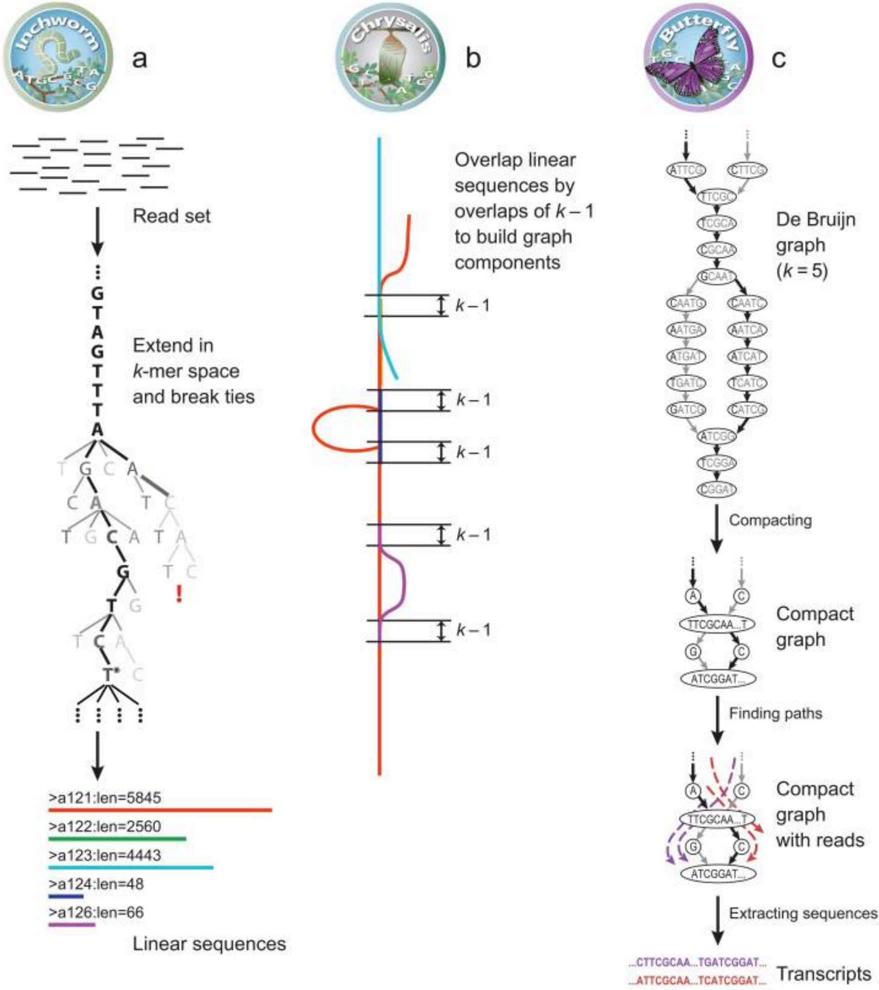
Assembly – De Bruijn

GAT ATT TTA TAC ACA CAG AGT GTT TTC TCA



TRANSCRIPTOME ASSEMBLY

Trinity



Review:

Mapping

Mapping - BWT

A B A A B A

$	*a*	*b*	*a*	*a*	*b*	*a*
a	$	*a*	*b*	*a*	*a*	*b*
a	*a*	*b*	*a*	$	*a*	*b*
a	*b*	*a*	$	*a*	*b*	*a*
a	*b*	*a*	*a*	*b*	*a*	$
b	*a*	$	*a*	*b*	*a*	*a*
b	*a*	*a*	*b*	*a*	$	*a*

Mapping – SAM/BAM

Gene Expression

Inter- versus intra-sample comparison

Gene Expression

$$TPM_i = \frac{X_i}{\tilde{l}_i} * \left(\frac{1}{\sum_j \frac{X_j}{\tilde{l}_j}} \right) * 10^6$$