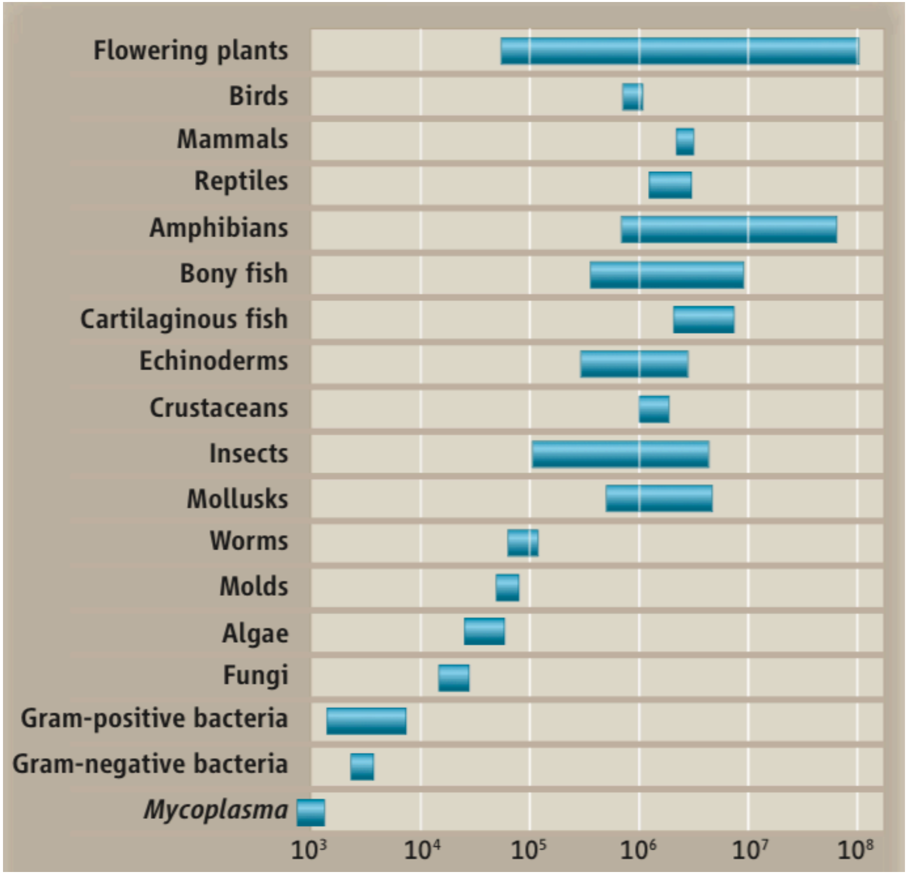
Genome Evolution

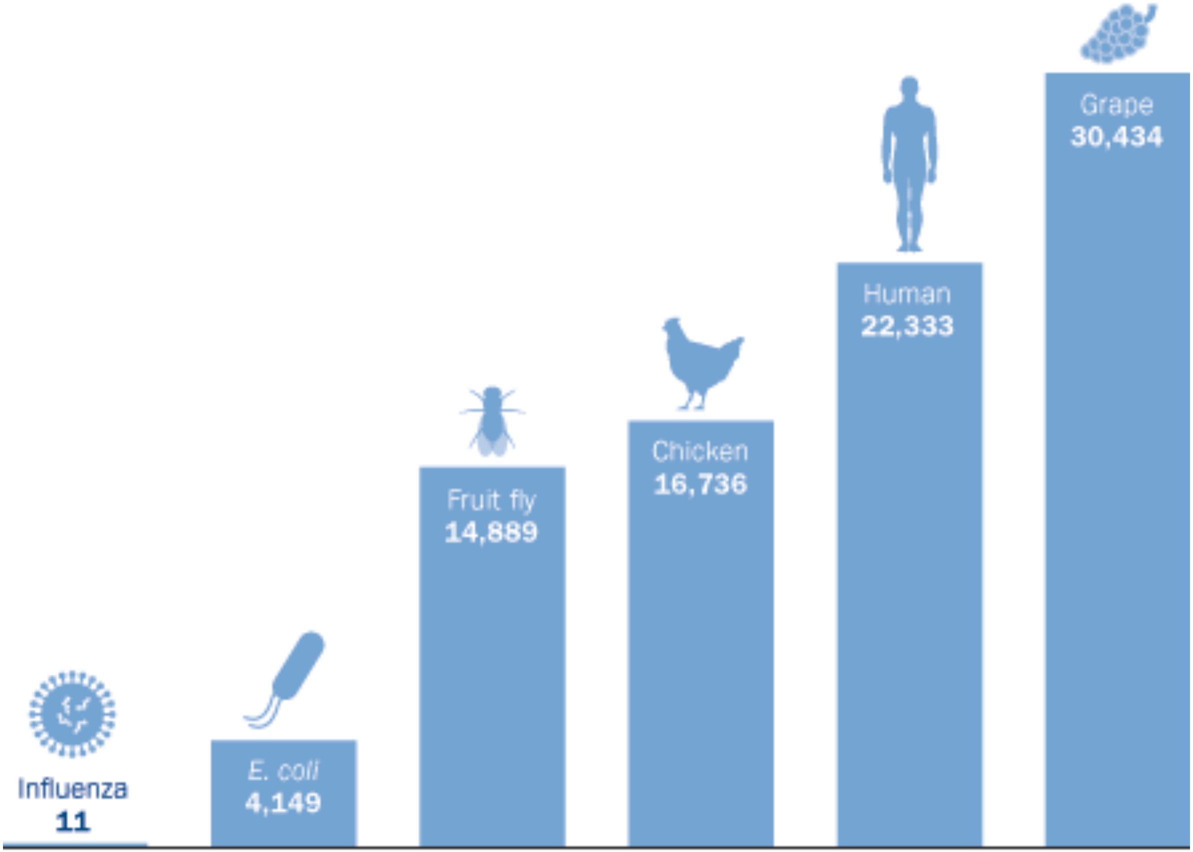
16Nov15

Announcements

Genome Diversity: size

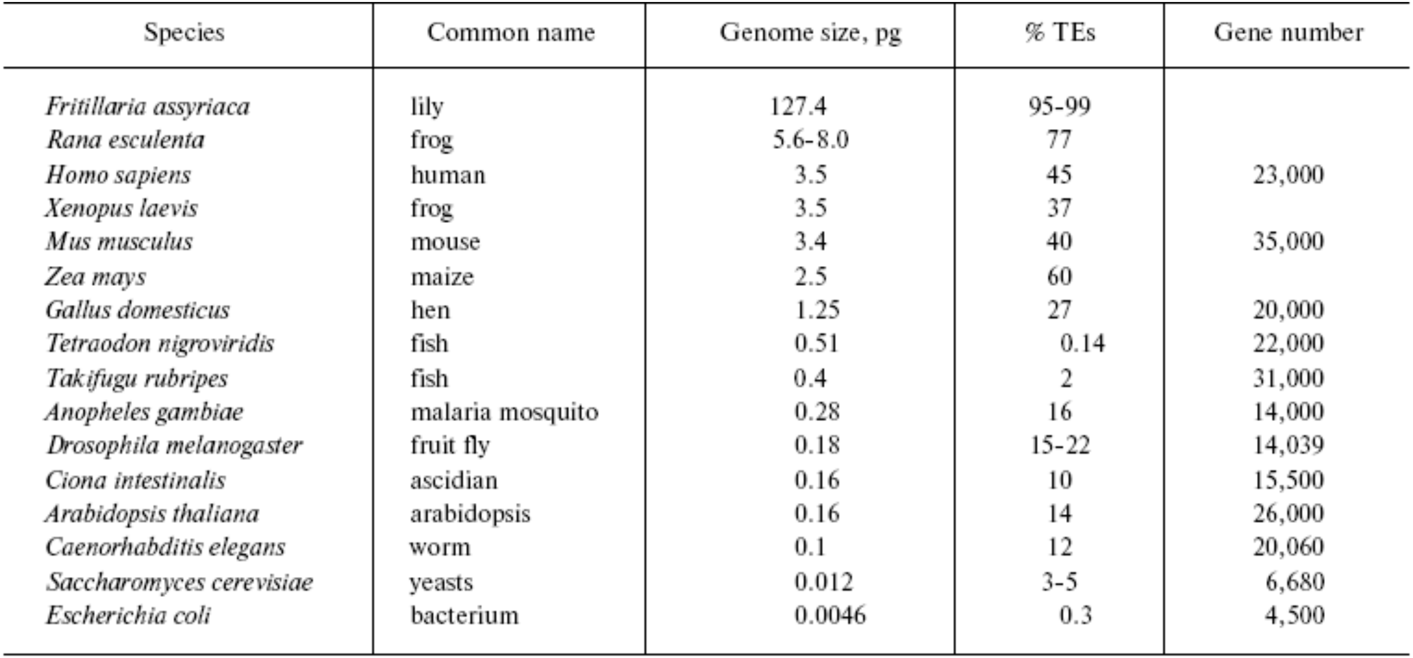


Genome Diversity: # genes



GENOME DIVERSITY:

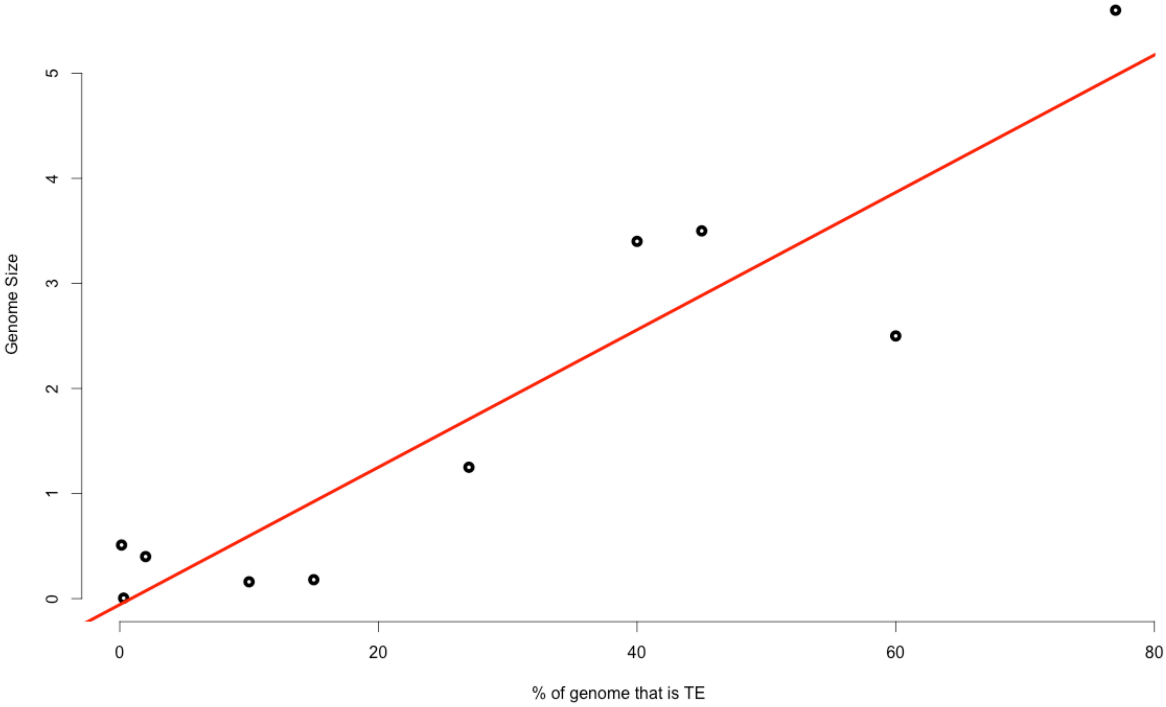
Genome size unrelated to gene number



Species	Common name	Genome size, pg	% TEs	Gene number
Fritillaria assyriaca	lily	127.4	95-99	
Rana esculenta	frog	5.6-8.0	77	
Homo sapiens	human	3.5	45	23,000
Xenopus laevis	frog	3.5	37	
Mus musculus	mouse	3.4	40	35,000
Zea mays	maize	2.5	60	
Gallus domesticus	hen	1.25	27	20,000
Tetraodon nigroviridis	fish	0.51	0.14	22,000
Takifugu rubripes	fish	0.4	2	31,000
Anopheles gambiae	malaria mosquito	0.28	16	14,000
Drosophila melanogaster	fruit fly	0.18	15-22	14,039
Ciona intestinalis	ascidian	0.16	10	15,500
Arabidopsis thaliana	arabidopsis	0.16	14	26,000
Caenorhabditis elegans	worm	0.1	12	20,060
Saccharomyces cerevisiae	yeasts	0.012	3-5	6,680
Escherichia coli	bacterium	0.0046	0.3	4,500

Genome Diversity:

Genome size unrelated to gene number



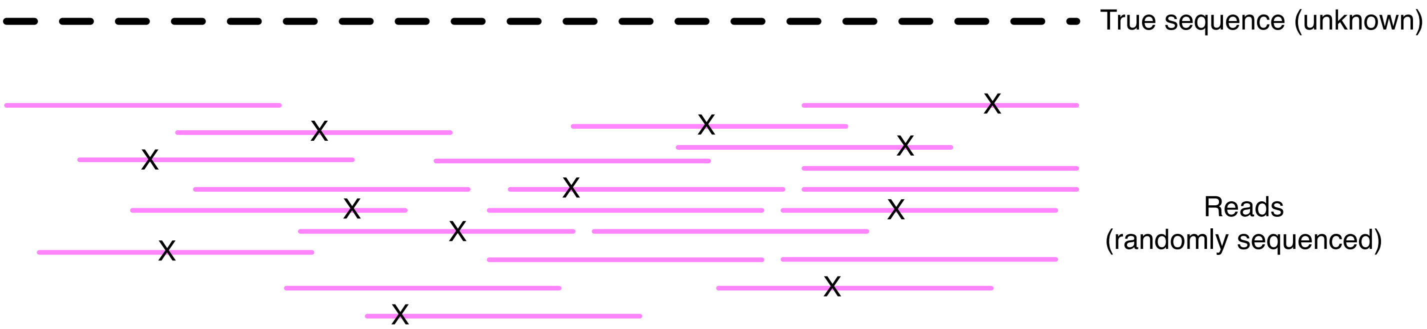
Genome Diversity:

Genome size unrelated to gene number

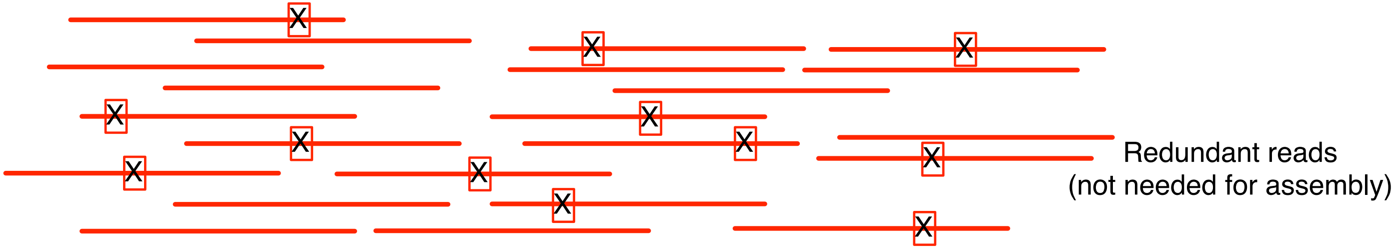
REVIEW:

Diginorm

DIGITAL NORMALIZATION

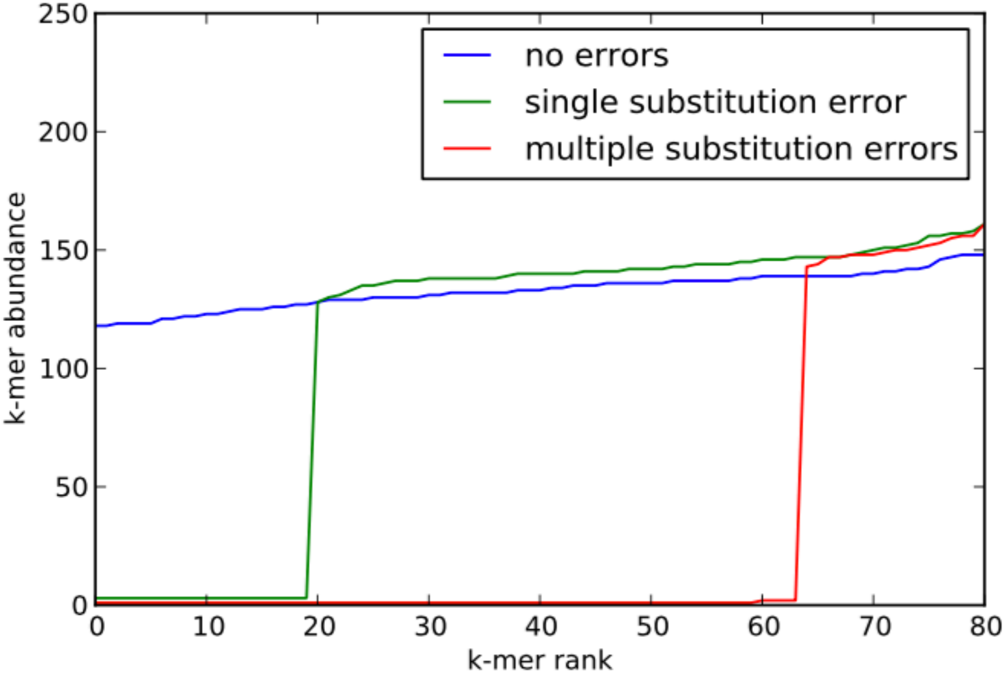


```
for read in dataset:
    if estimated_coverage(read) < C:
        accept(read)
    else:
        discard(read)
```



Digital Normalization

```
for read in dataset:
   if estimated_coverage(read) < C:
      accept(read)
   else:
      discard(read)
```



Review:

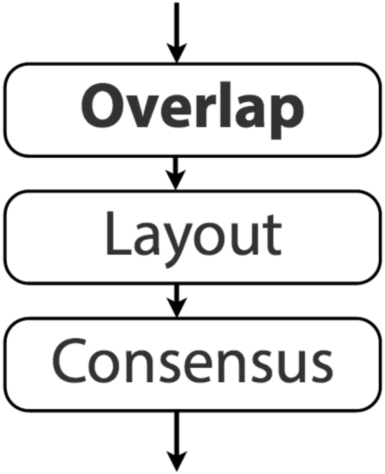
Genome and Transcriptome Assembly

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

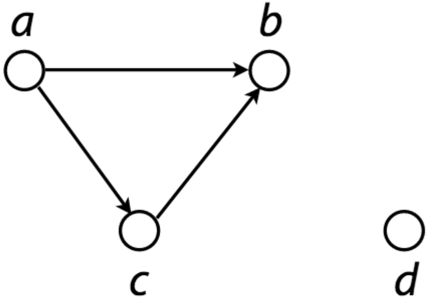
Assembly

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*

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

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

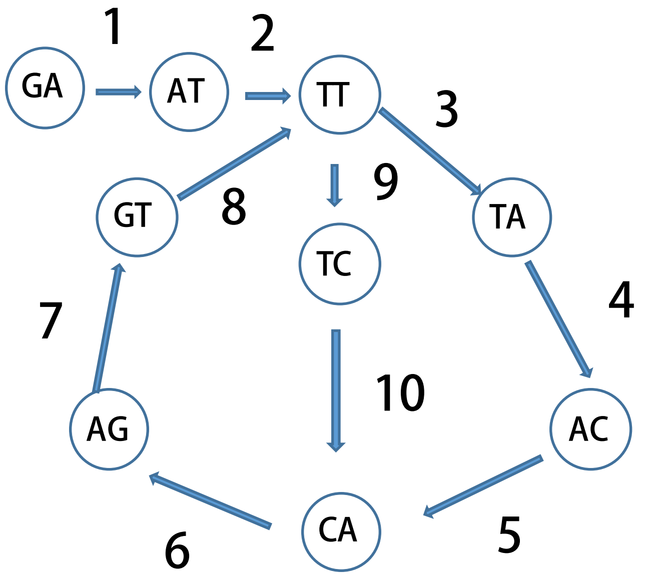
Assembly – De Bruijn

Hamiltonian Path Problem

Eulerian Path Problem

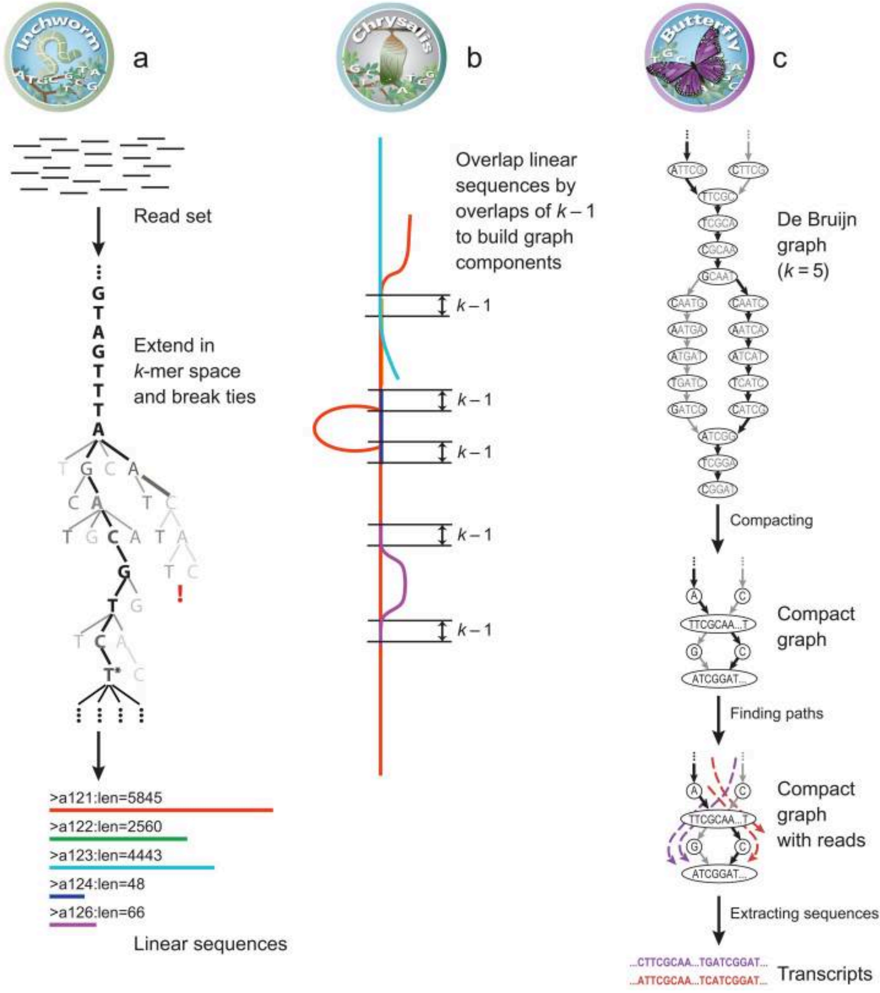
Assembly – De Bruijn

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