

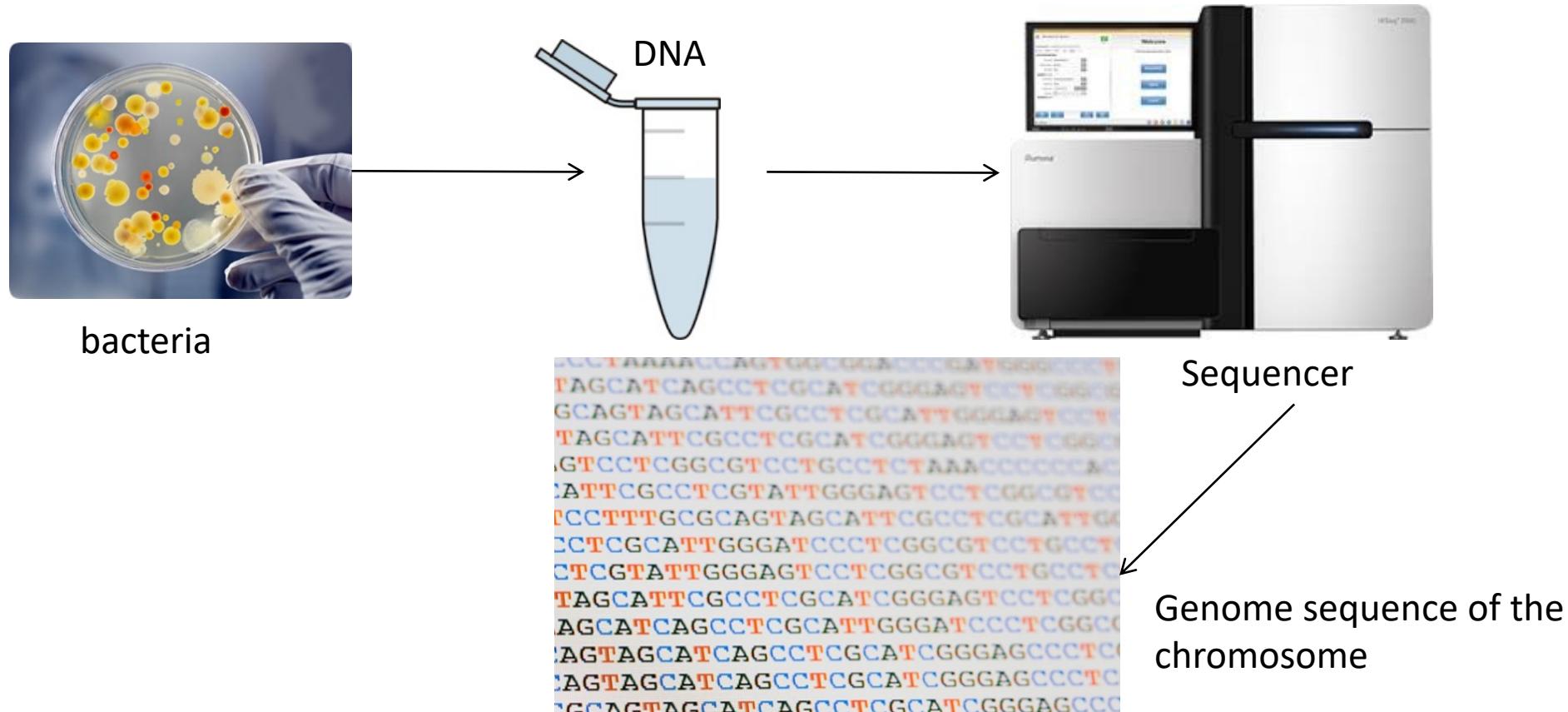
Genome assembly

Jenni Hultman

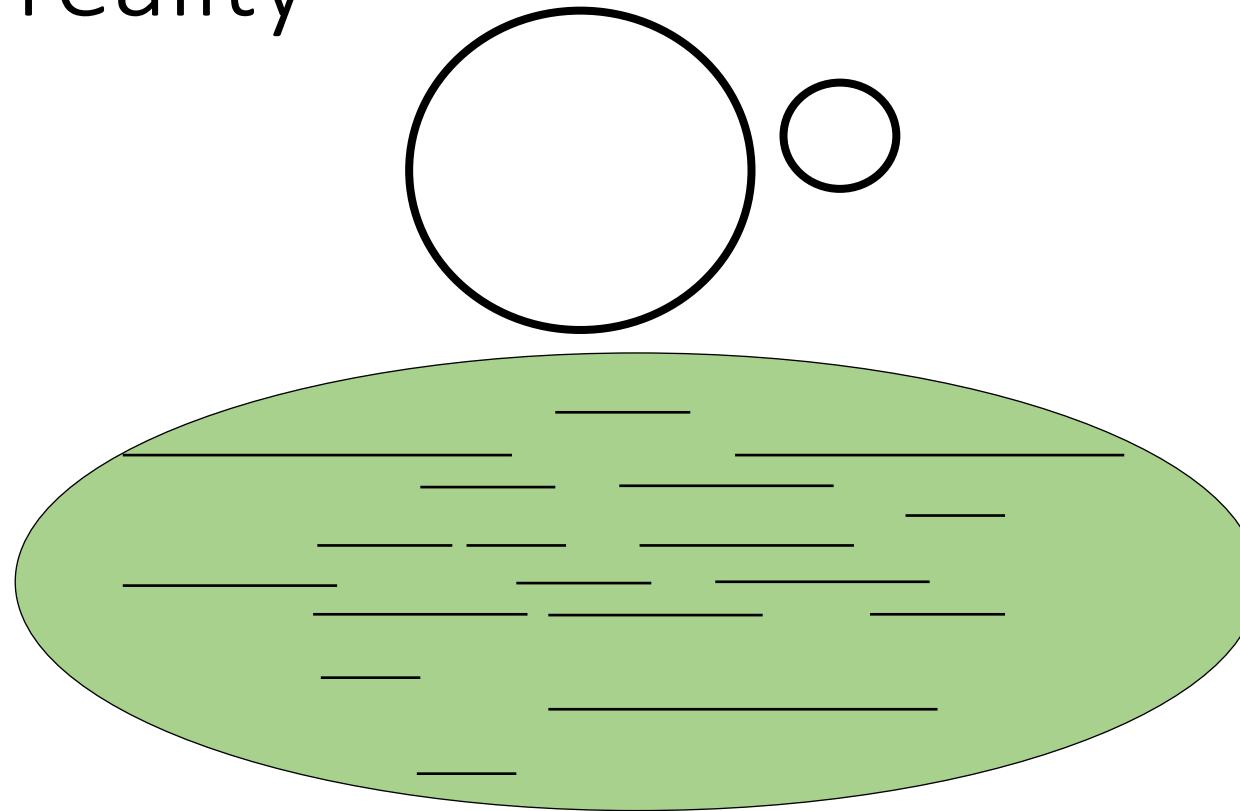
Reconstruct the original genome from
the sequence reads

De novo genome assembly

- Putting the sequence reads together
- In an ideal world:

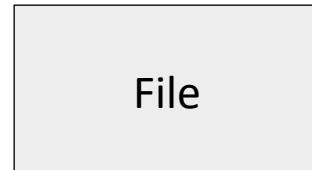


In reality



Genome

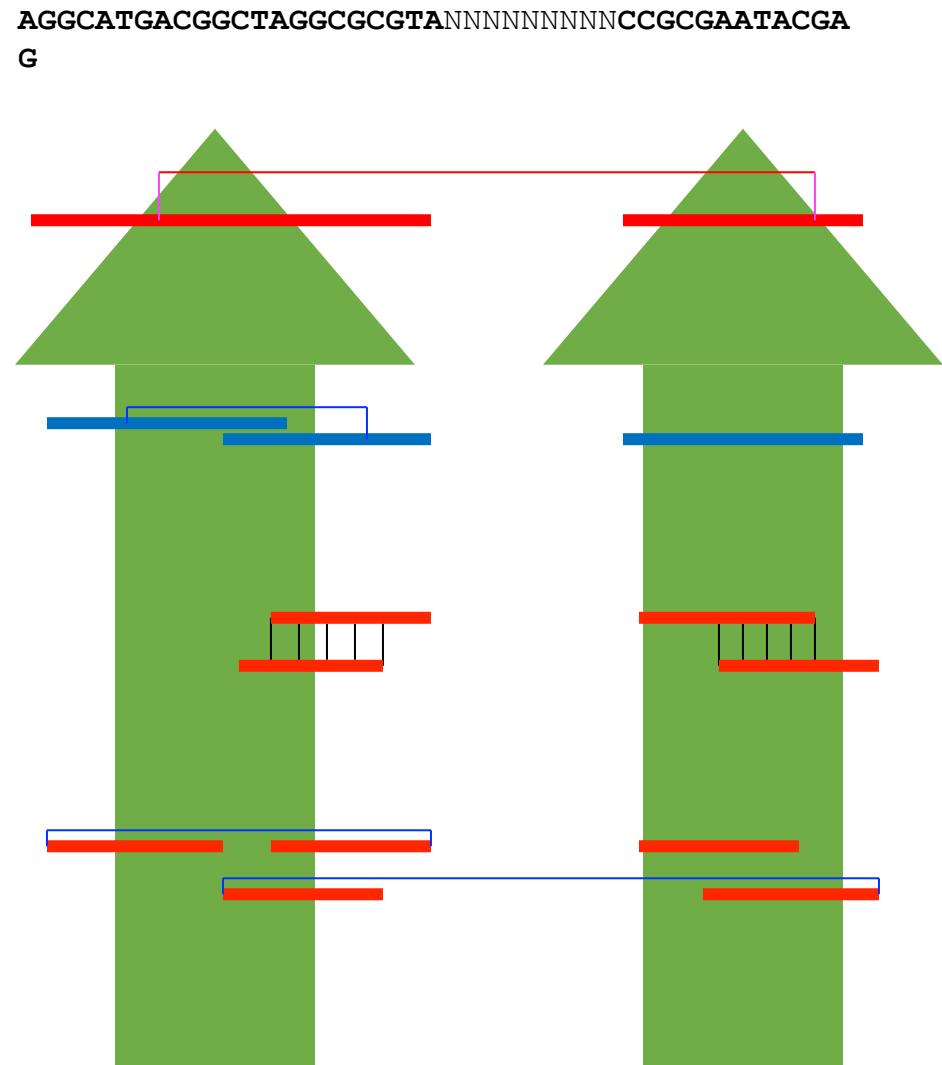
Fragmentation to
short pieces



Fastq-file with sequences
of the fragments

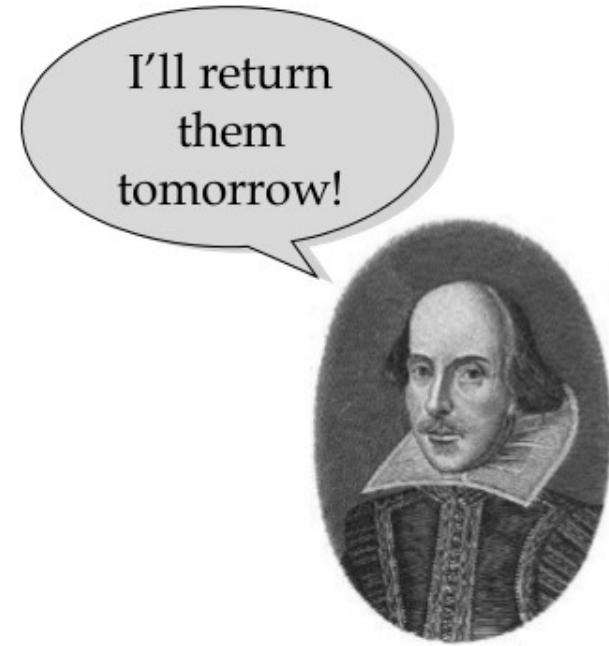
Glossary

- **Consensus**
 - Multiple alignment of sequences
- **Scaffold=contigs + gaps**
 - group of contigs that can be ordered and oriented with respect to each other (usually with the help of mate-pair data)
- **Contigs**
 - contiguous segment of DNA reconstructed (unambiguously) from a set of reads
- **Overlaps**
 - Shared sequences between the suffix of one read and the prefix of another
- **Reads**
 - segment of DNA "read" by a sequencing instrument
- **Mate-pairs, paired ends**
 - pair of reads whose distance from each other within the genome is approximately known



A small “genome”

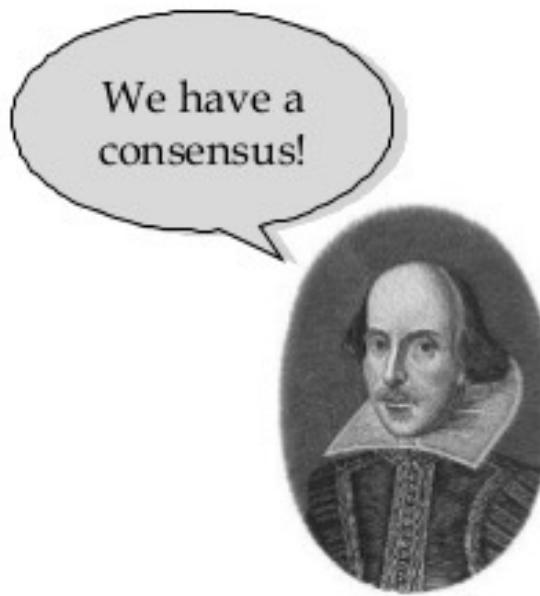
Friends,
Romans,
countrymen,
lend me your ears;



Shakespearomics

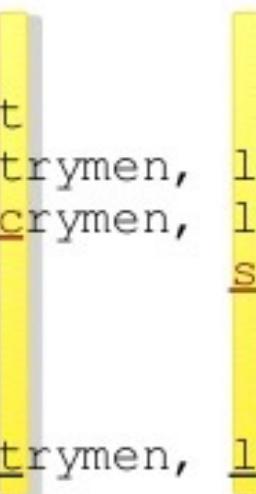
- **Reads**

ds, Romans, count
ns, countrymen, le
Friends, Rom
send me your ears;
crymen, lend me



- **Overlaps**

Friends, Rom
ds, Romans, count
ns, countrymen, le
crymen, lend me
send me your ears;

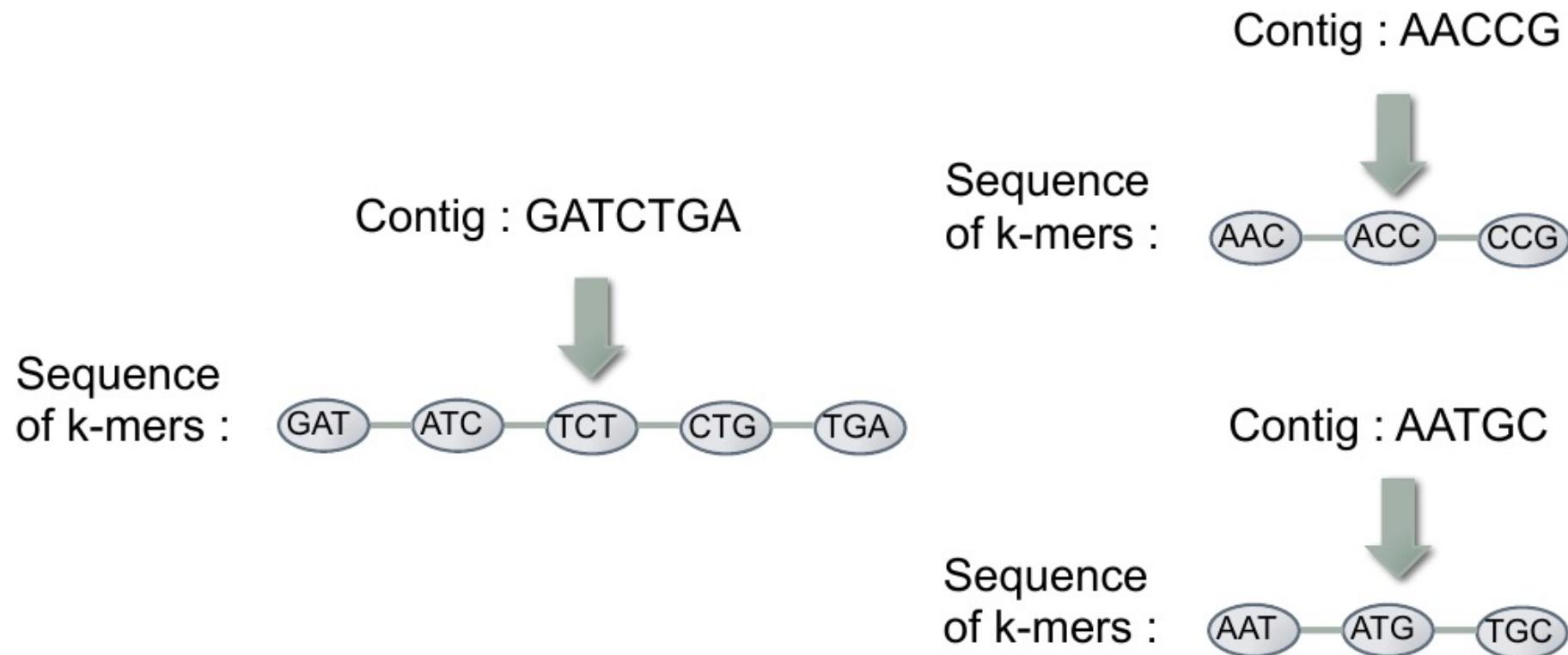


- **Majority consensus**

Friends, Romans, countrymen, lend me your ears;

de Bruijn graph

- In real life 1,000,000 sequences cannot be compared with each read ($10^6 * 10^6$ comparisons)
- Use of k-mers
 - Fragments of k-length



How k-mers work in assembly kmer=4

ATCC A GTA G

A GT A G G A T C A A

AT CC

A G T A

T C C A

G T A G

C C A G

T A G G

C A G T

A G G A

A G T A

G G A T

G T A G

G A T C

A T C A

T C A A

ATCCCAAGTA

A_nTAG_nATCAA

ATCC

TCCA

CCAG

CAGT

A_nTA

G_nTAG

A_nTA

G_nTA

TAG

A_nGA

G_nAT

GATC

ATCA

TCAA

ATCCAGTA

A_GTAGGATCAA

ATCC

TCCA

CCAG

CAGT

AGTA

GTA

A_GTA

GTA

TAG

A_GGA

G_GAT

GATC

ATCA

TCAA

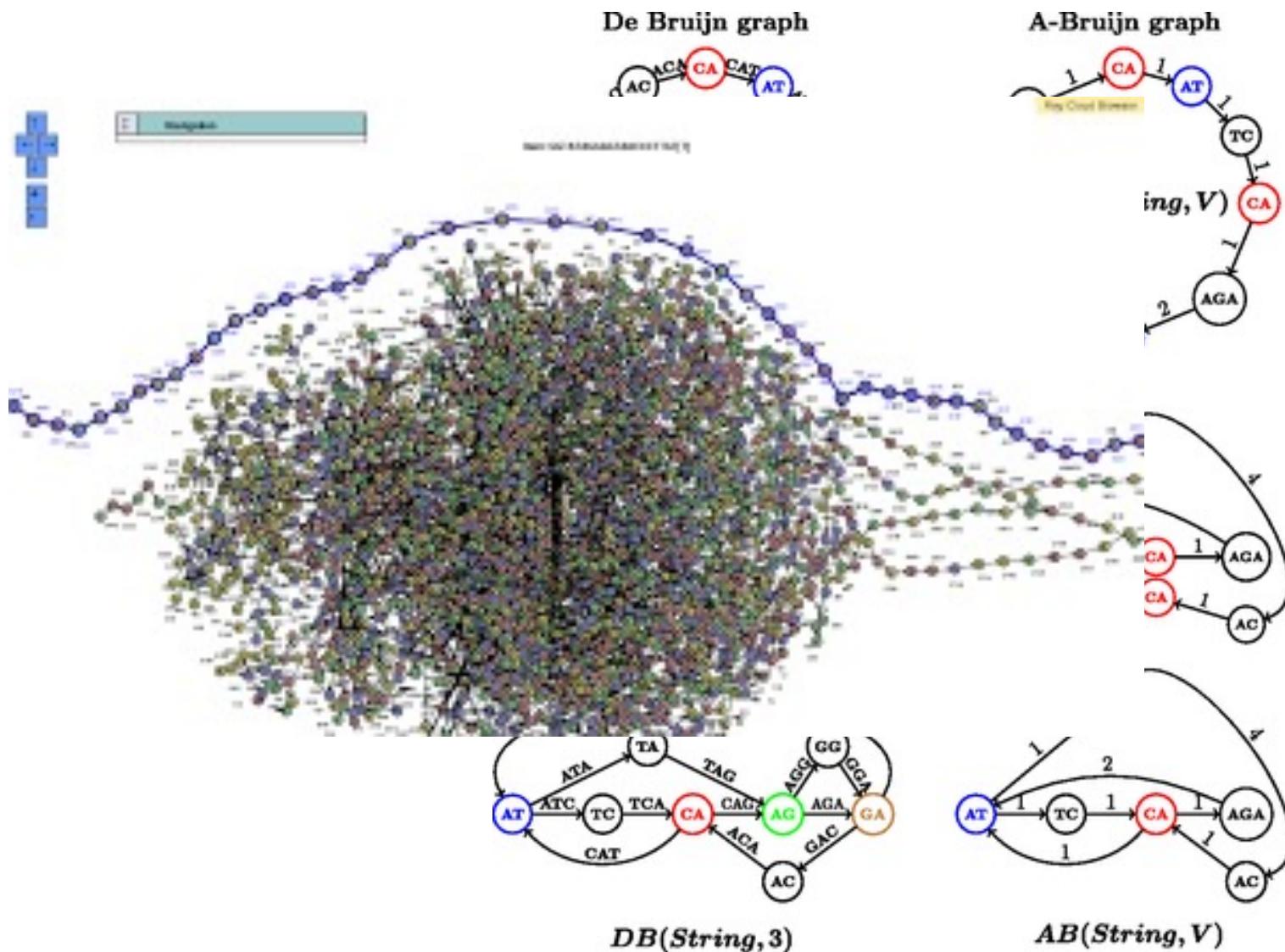
ATCCAGTAGGATCAA

Size of k-
mer has
huge
effect!

Too small -> misassembly
(anything can assemble)

Too long -> no
assembly/misassembly

In real life with 10^6 sequences



File Tools View Help

De Bruijn graph information

Nodes: 51,639
Edges: 65,832
Total length: 18,712,634

Graph drawing

Scope: Entire graph
Style: Single Double

Graph display

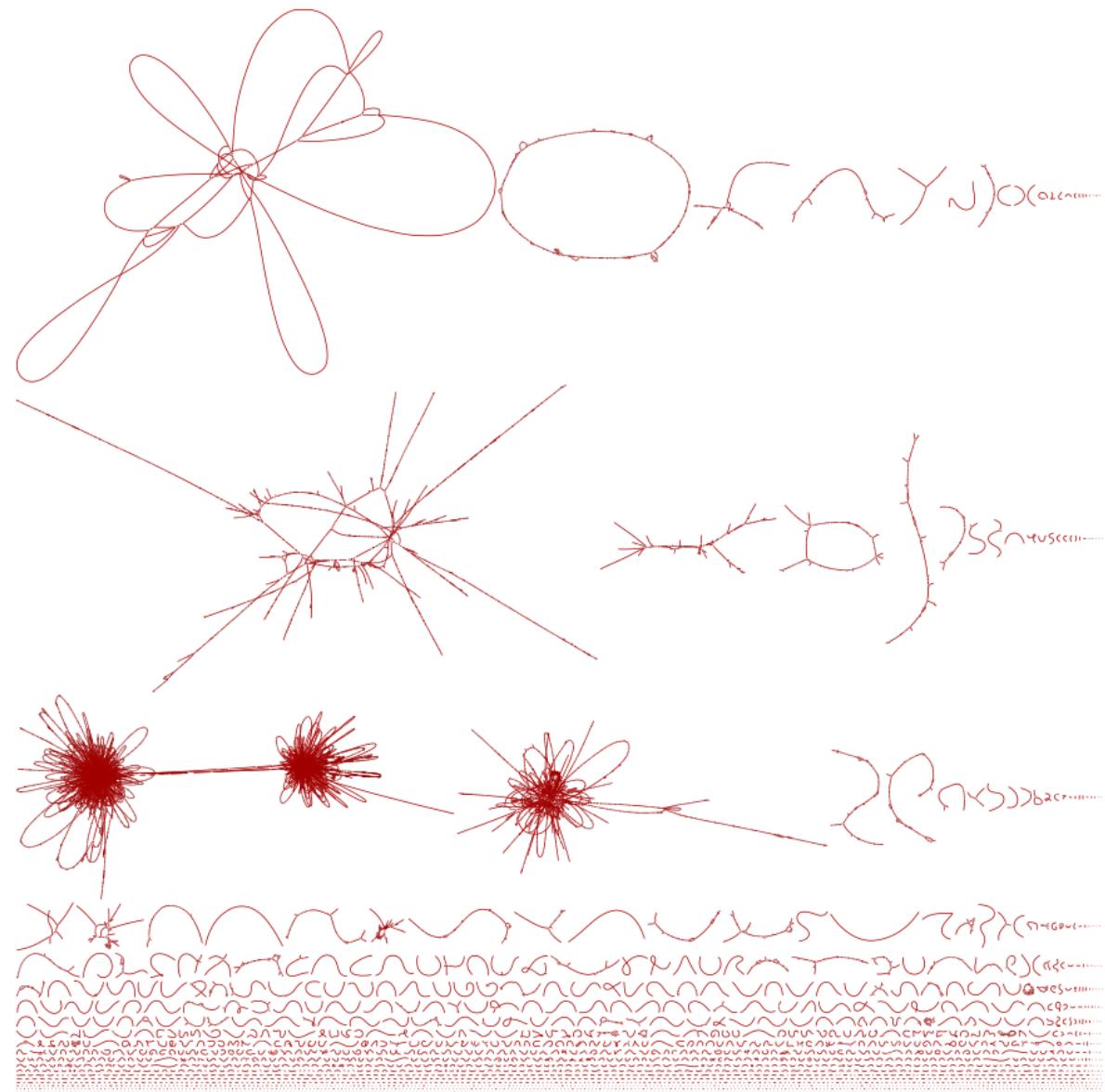
Zoom: 2.6%
Uniform colour

Node labels

Custom Number
 Length Coverage
 Text outline

BLAST

Query:

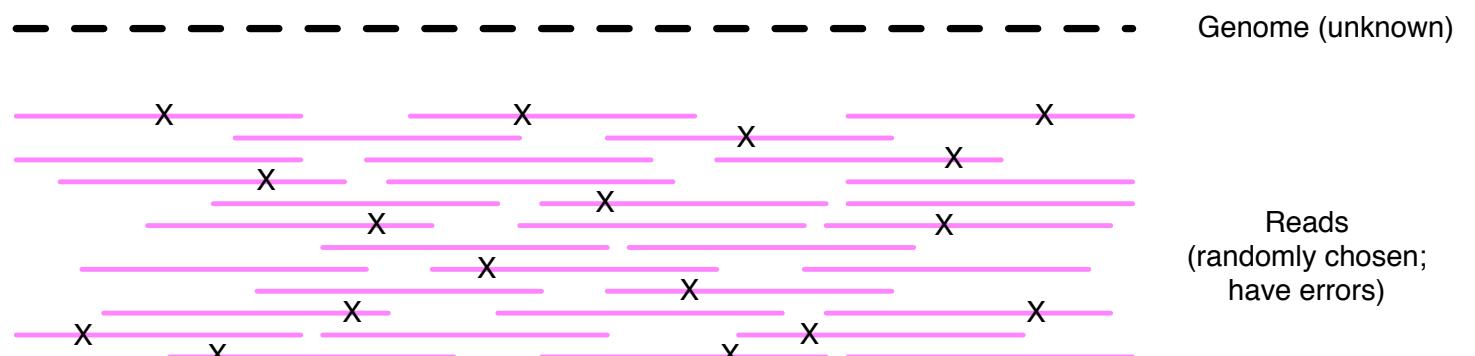


Find nodes

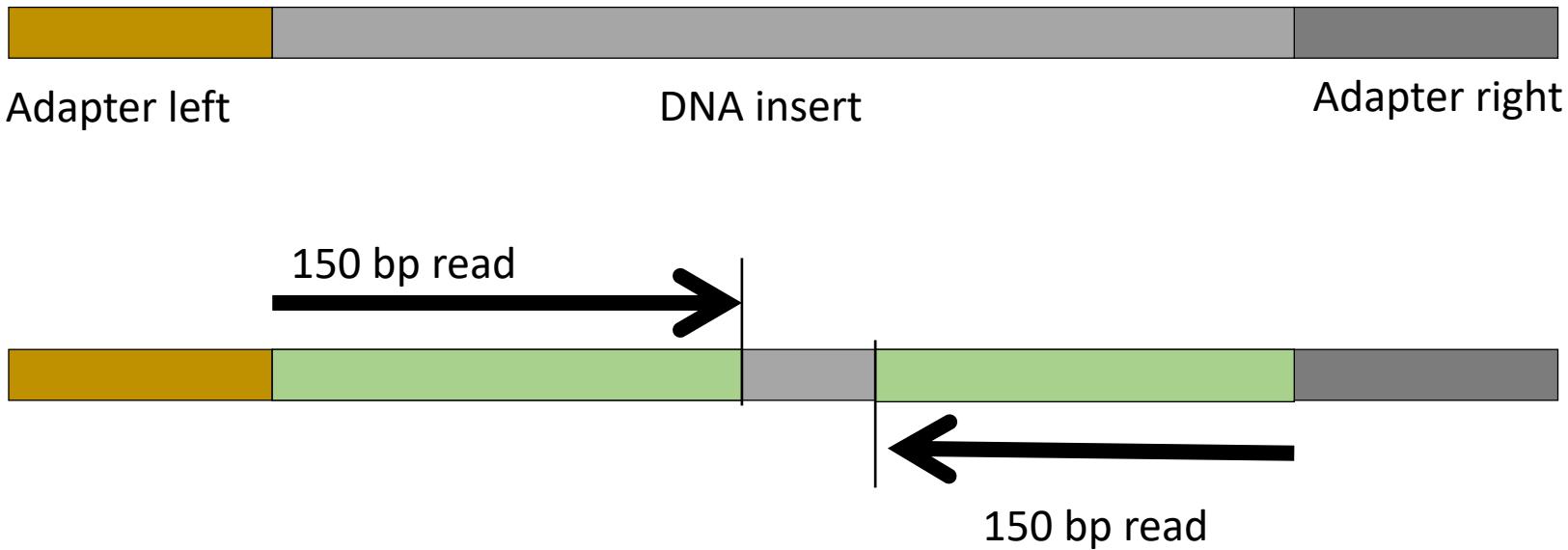
Node(s):

Coverage

- Coverage describes the average number of reads that align to, or "cover," known reference base
- Average coverage

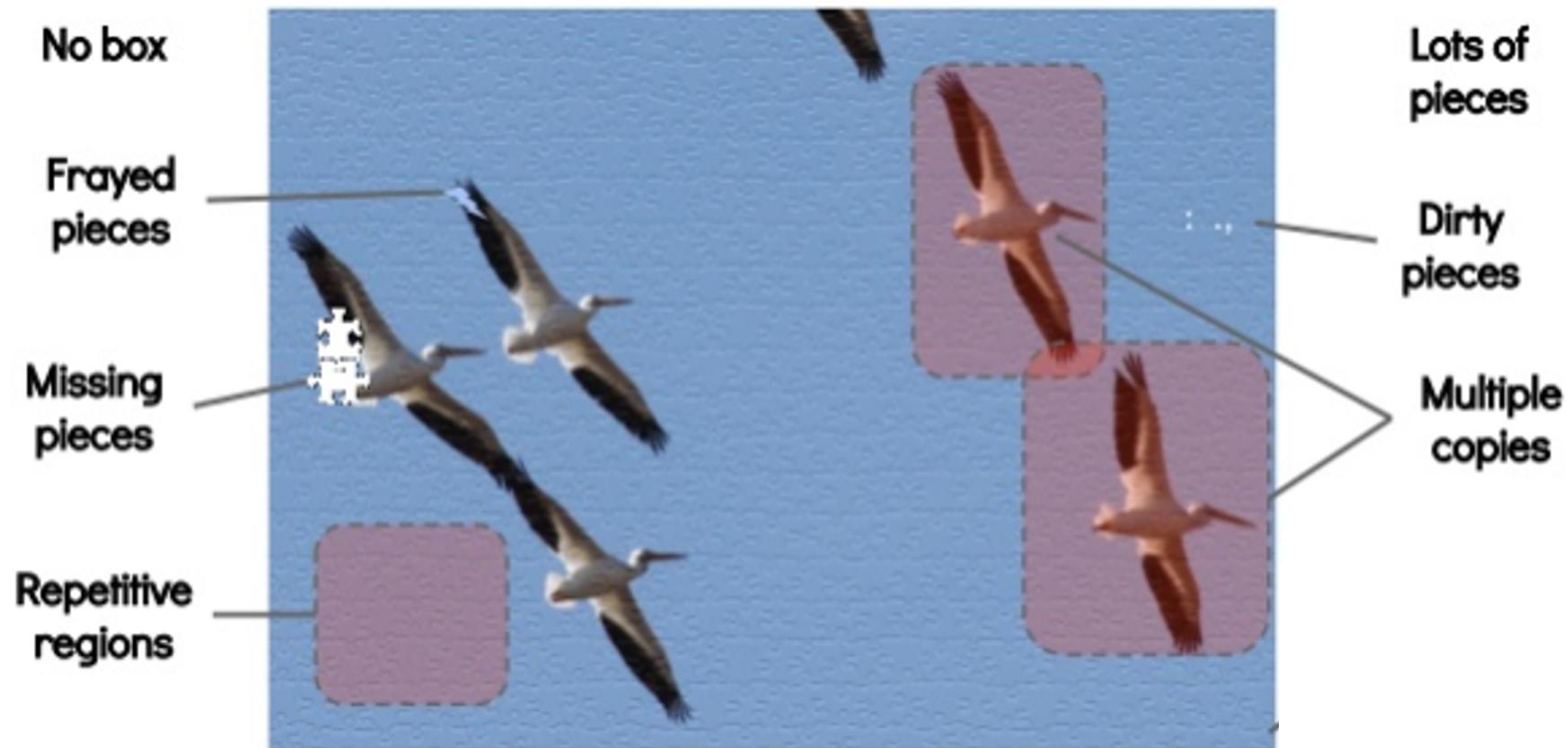


Paired-end sequence



- Helps in assembly, tell to assembler to aid correct assembly
- Also insert size needed

What makes a puzzle hard?



What makes genome assembly tricky?

- Many pieces (computational)
- Errors in sequence (which is correct?)
- Missing fragments
- Repetitive fragments (tandem, interspersed)
- Multiple copies (rRNA gene as an example)
- Circular genome: no starting point
- Non-chromosomal DNA (plamids)

Help from long read sequencing?

- Long read alone?
- Hybrid assembly?

Nanopore data (28x depth) allowed complete and accurate assembly of the *K. pneumoniae* isolate INF177.

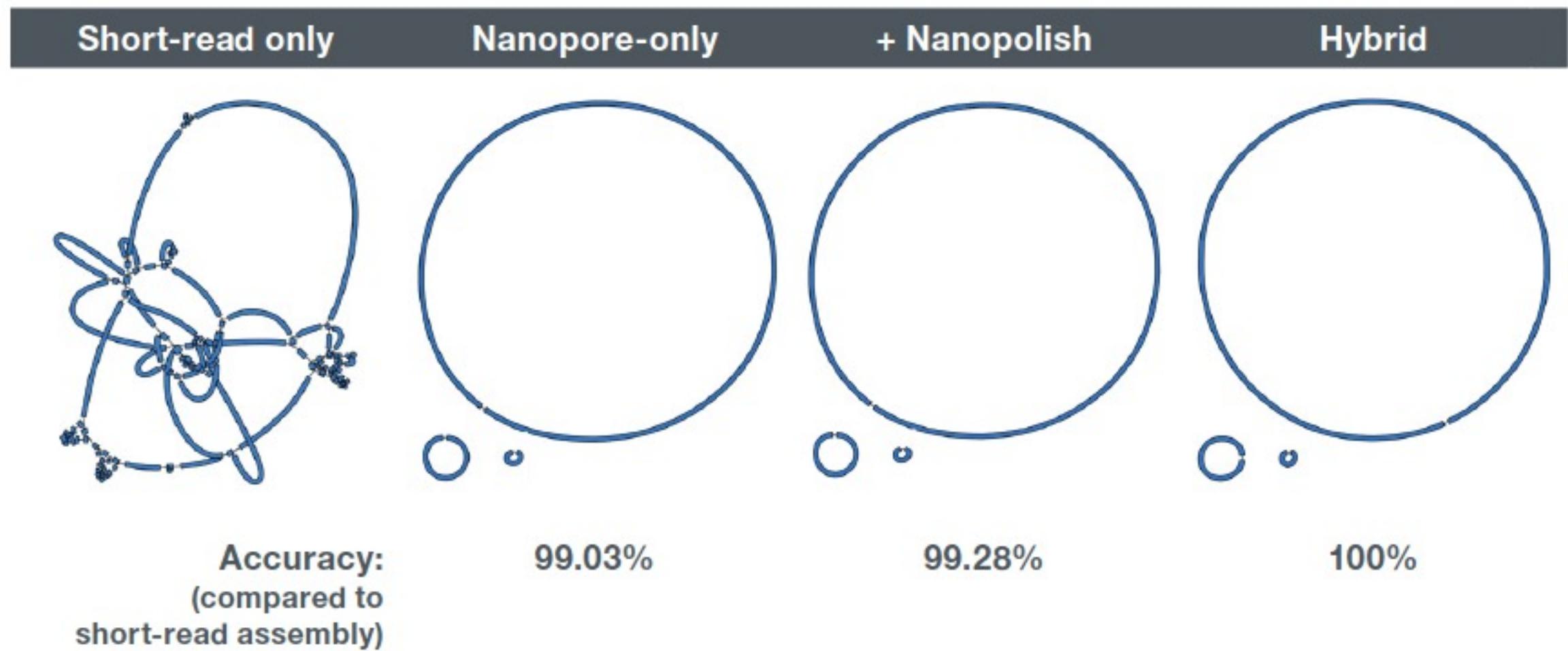
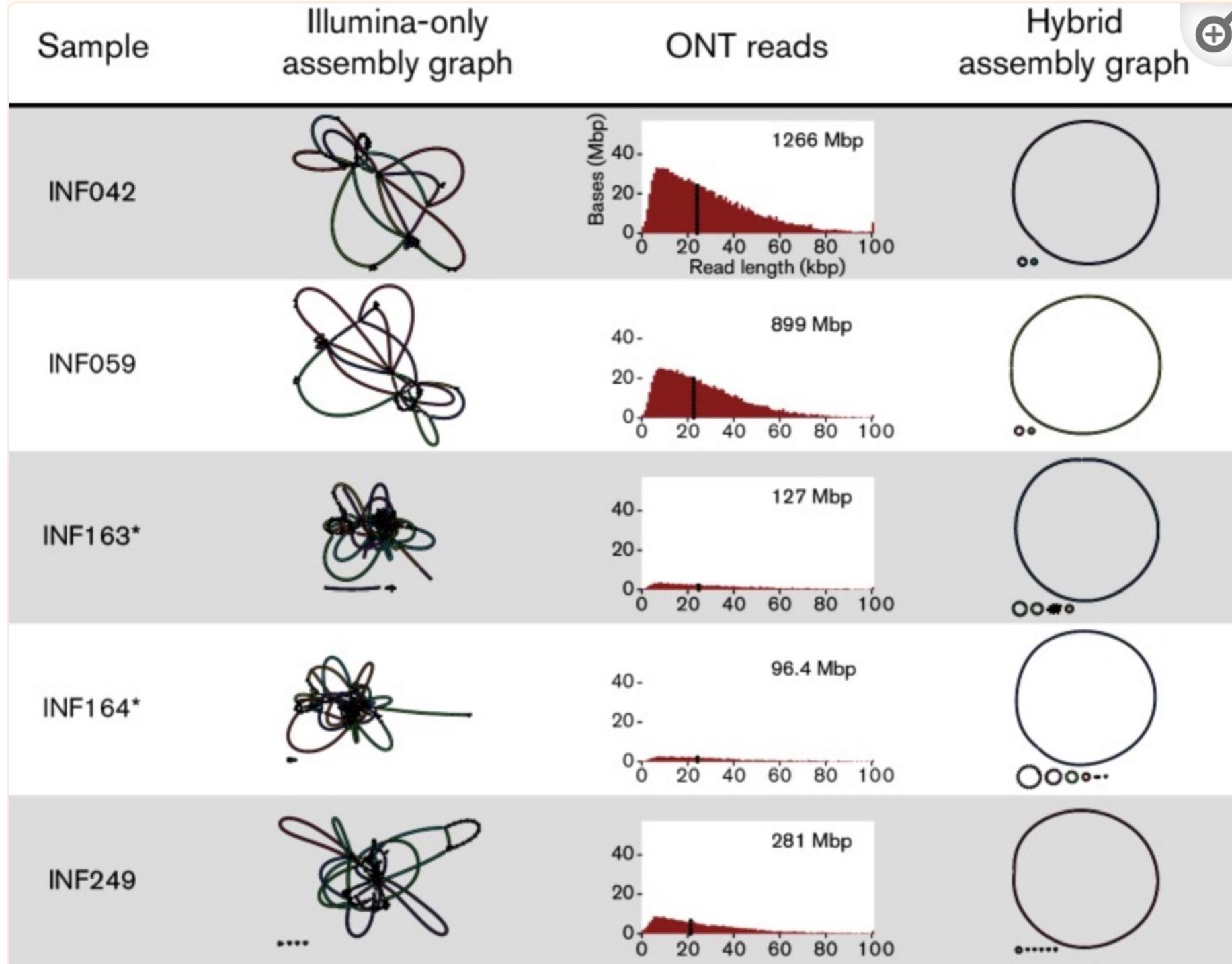


Figure courtesy of Ryan Wick, University of Melbourne, Australia.



Help from long read sequencing?



- Long read alone?
- Hybrid assembly?
- Spades
 - Uses multiple k-mers
 - Hybrid assembly possible, nanopore and Illumina reads assembled together
- If you have time try
 - Nanopore alone
 - Illumina aloneHow did the results differ?