Metagenome assembly

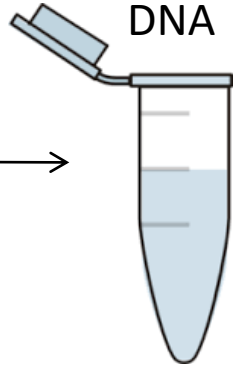
Jenni Hultman

Reconstruct the original genome from the sequence reads

De novo genome assembly

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

bacteria



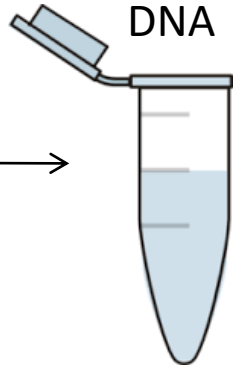
Sequencer

Genome sequence of the chromosome

De novo metagenome assembly

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

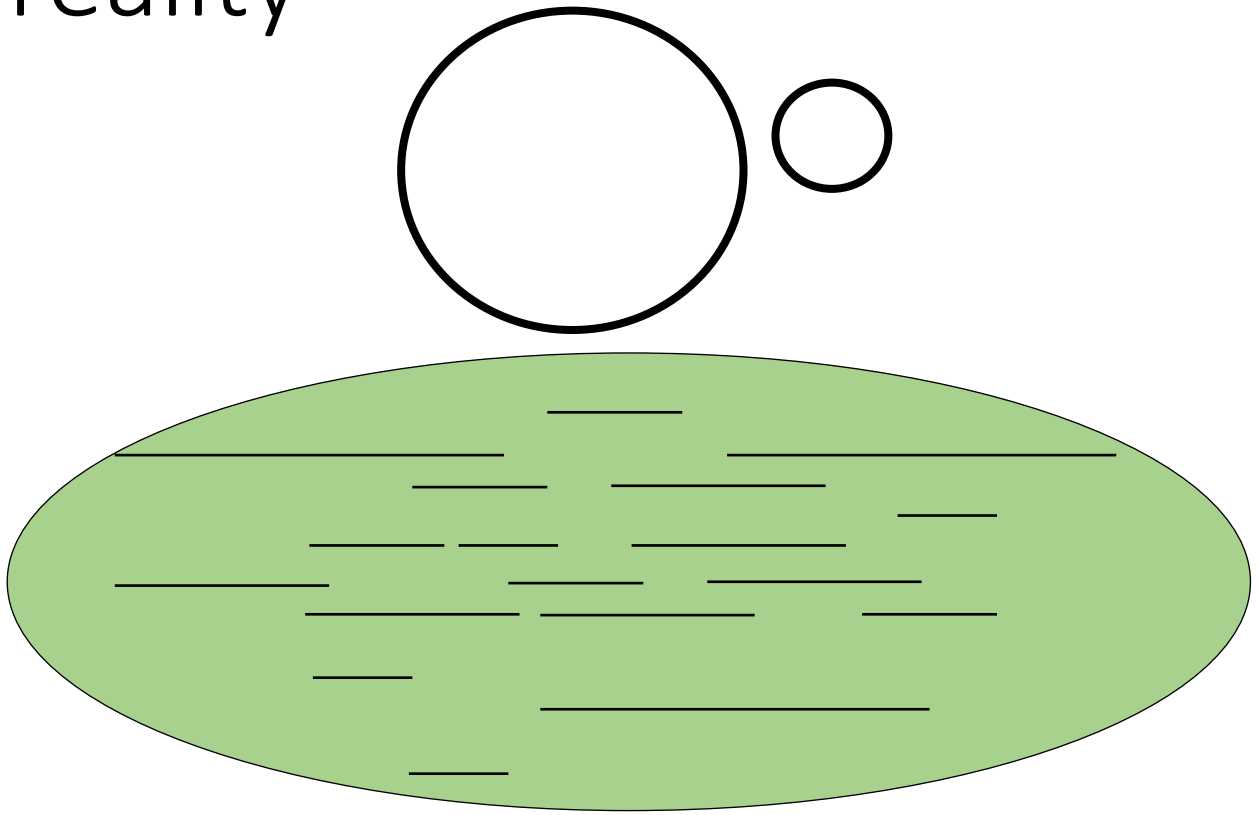
community



Sequencer

Genome sequence of the community

In reality

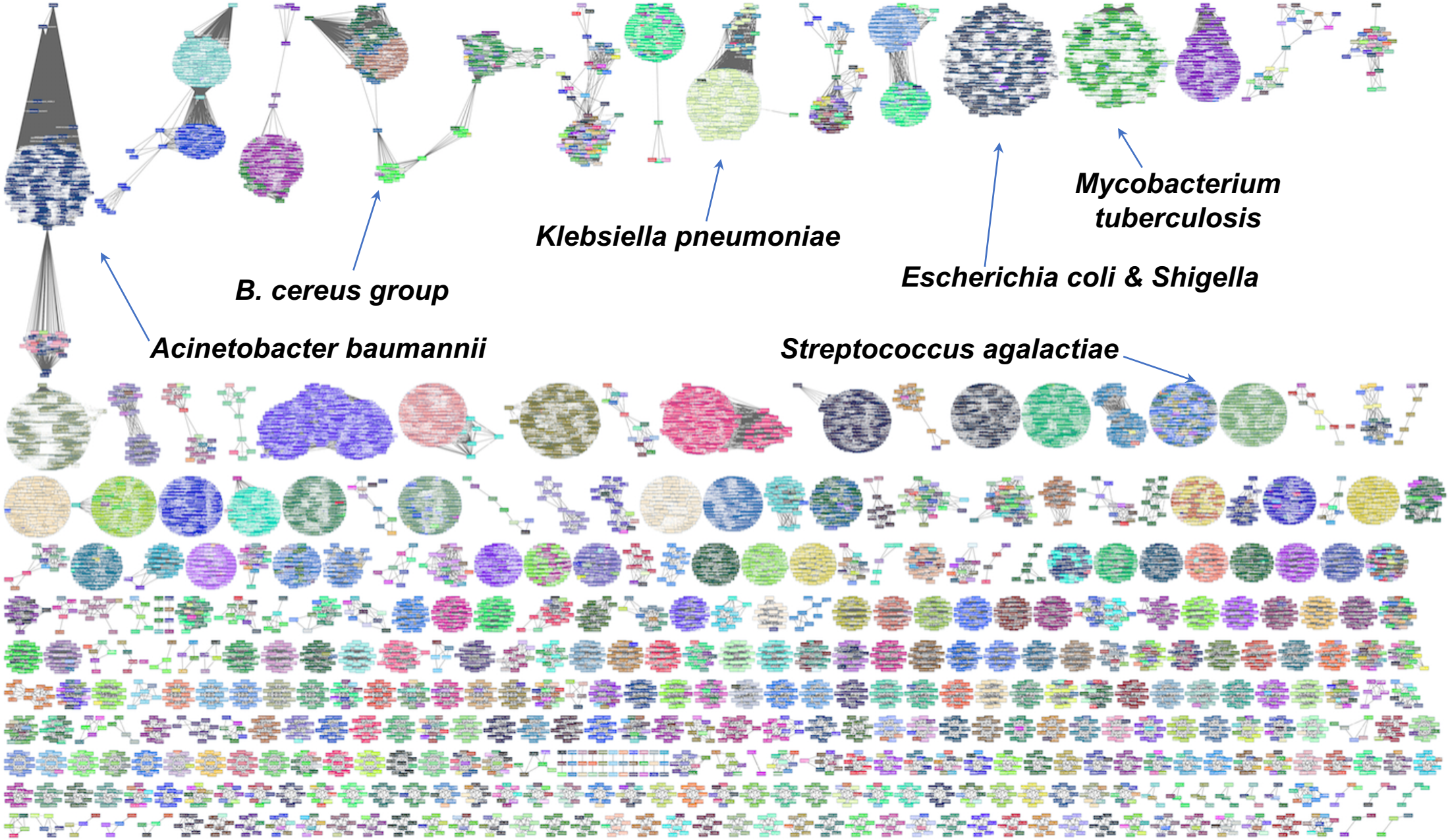


Genome

Fragmentation to short pieces

File

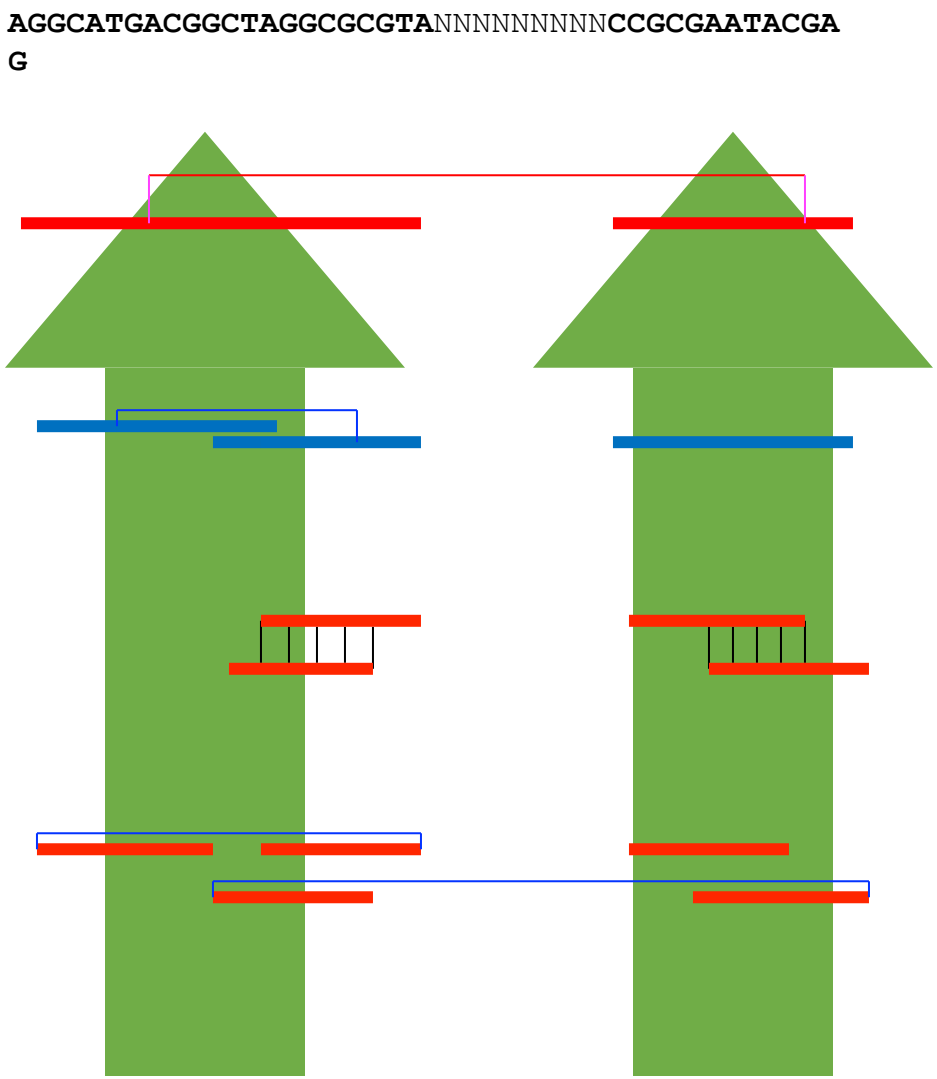
Fastq-file with sequences of the fragments



Slide from Todd Treangen

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”

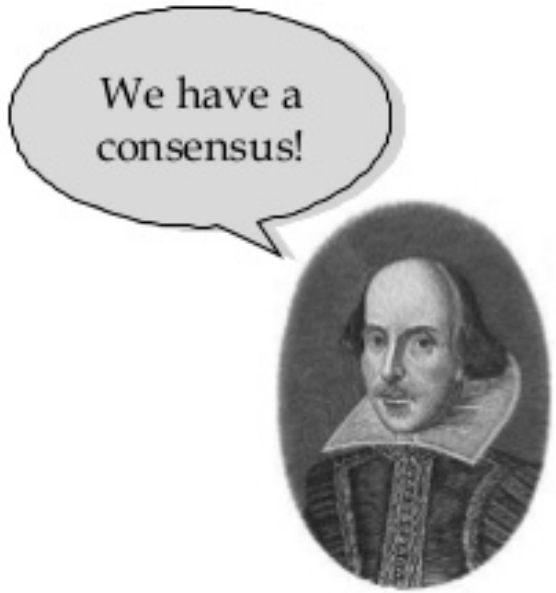

```
Friends,
   Romans,
      countrymen,
         lend me your ears;
```

Slide from Torsten Seemann

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;
```

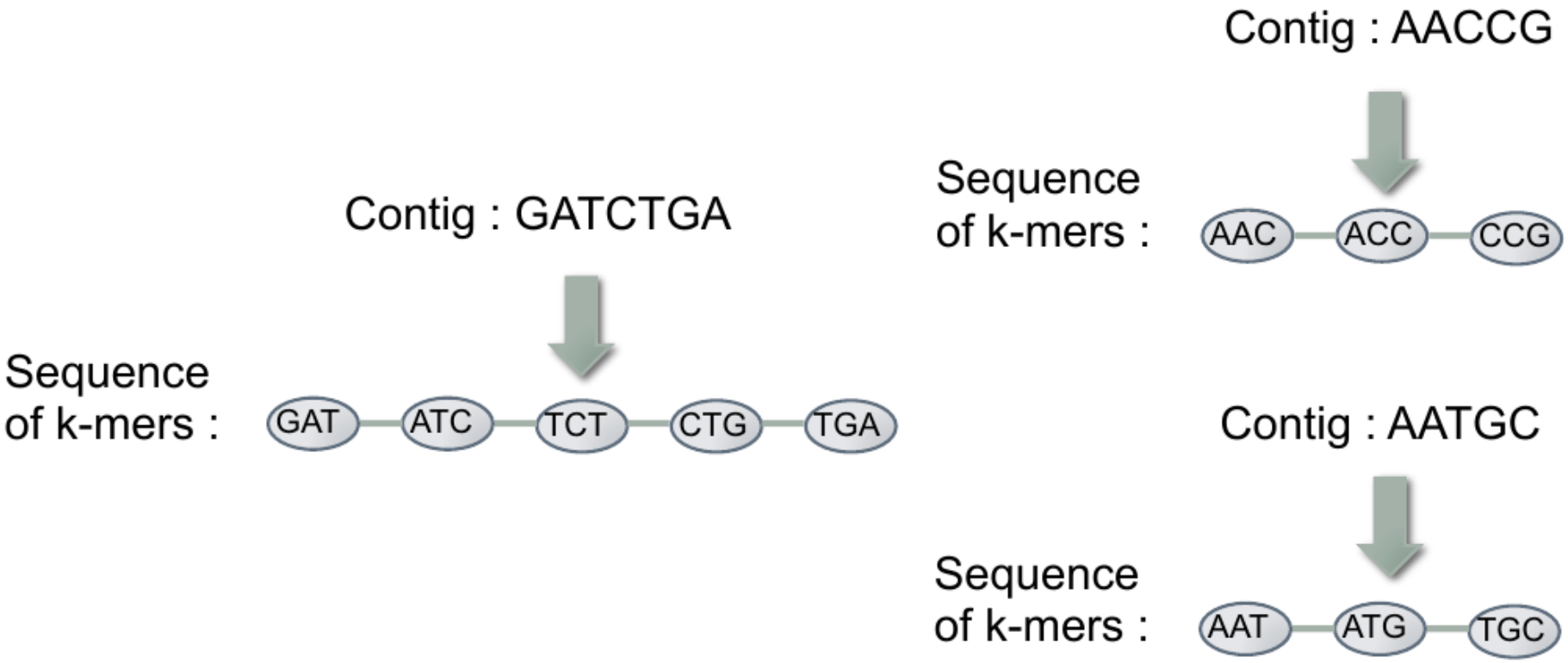
- ***Majority consensus***

```
Friends, Romans, countrymen, lend me your ears;
```

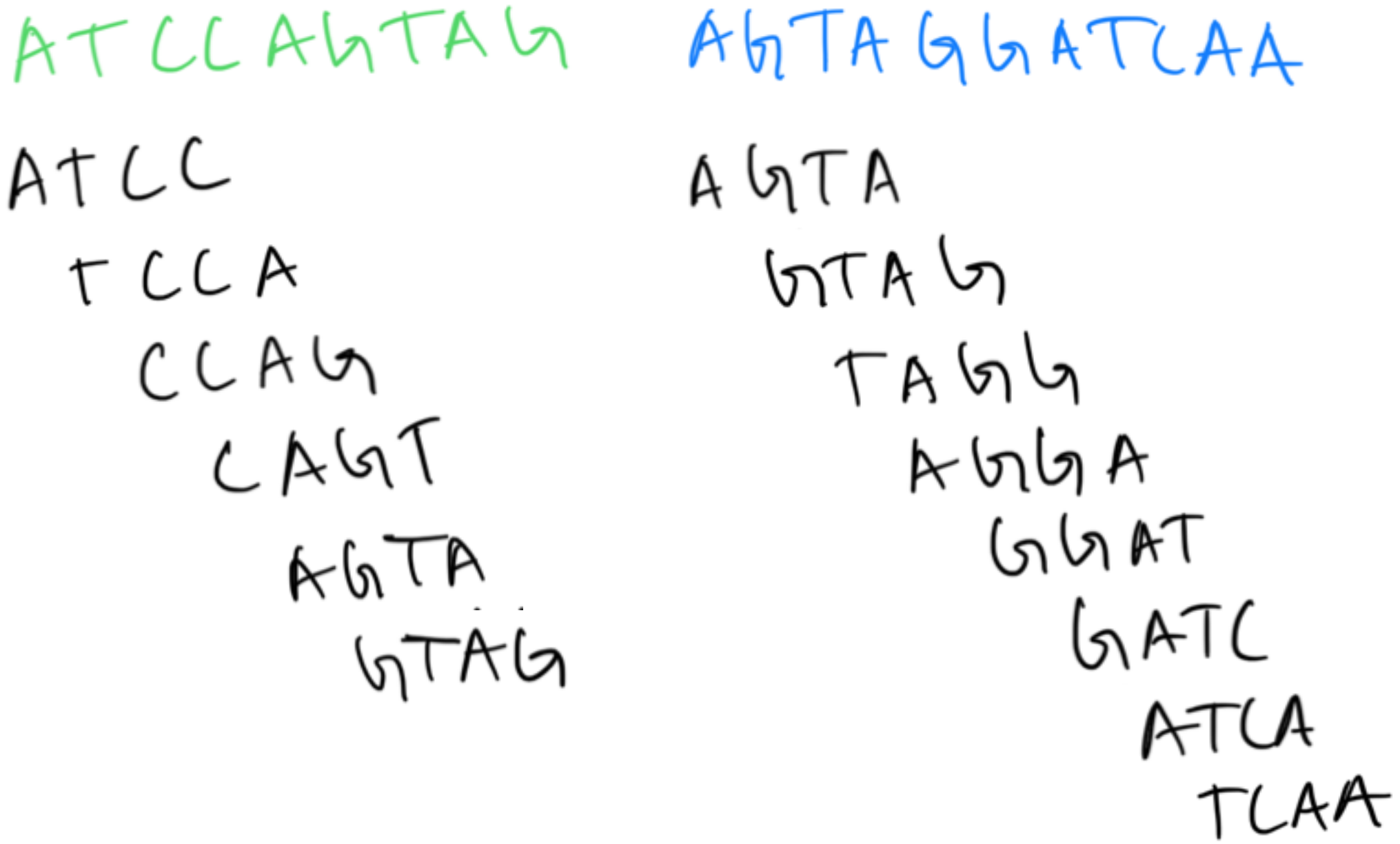
Slide from Torsten Seemann

de Bruijin graph

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



How k-mers work in assembly kmer=4




ATCCAGTAG
AGTAGGATCAA
ATCC
TCCA
CCAG
CAGT
AGTA
GTAG
AGTA
GTAG
TAGG
AGGA
GGAT
GATC
ATCA
TCAA

ATCCAGTAG

ATCC

TCCA

CCAG

CAGT

AGTA

GTAG

AGTAGGATCAA

AGTA

GTAG

TAGG

AGGA

GGAT

GATC

ATCA

TCAA

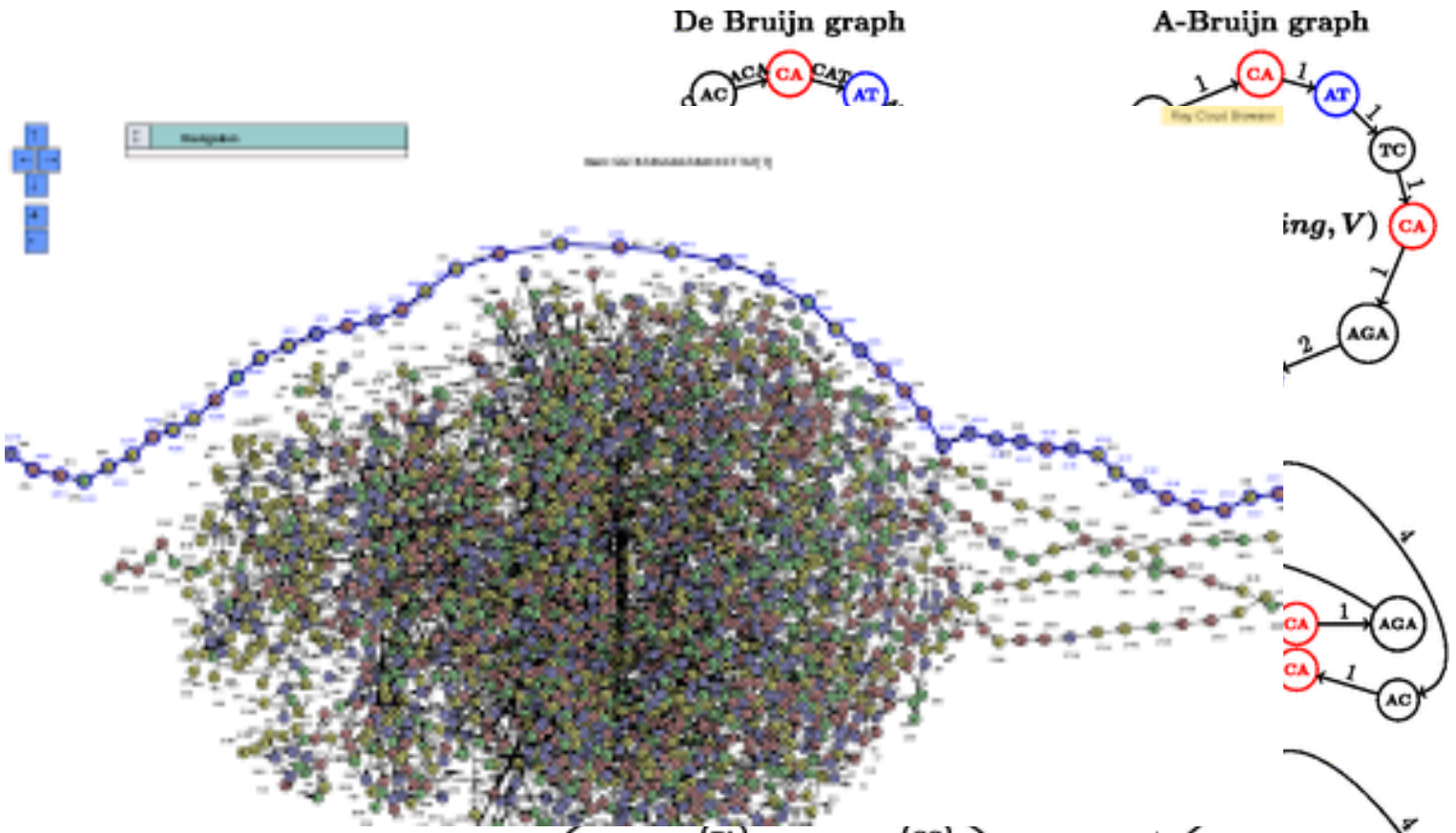
ATCCAGTAGGATCAA

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



Bandage – /media/ryan/Data/Bandage_demo/07_NW1_metagenome/NW1_LastGraph

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

Draw graph

Graph display

Zoom: 2.6%

Uniform colour

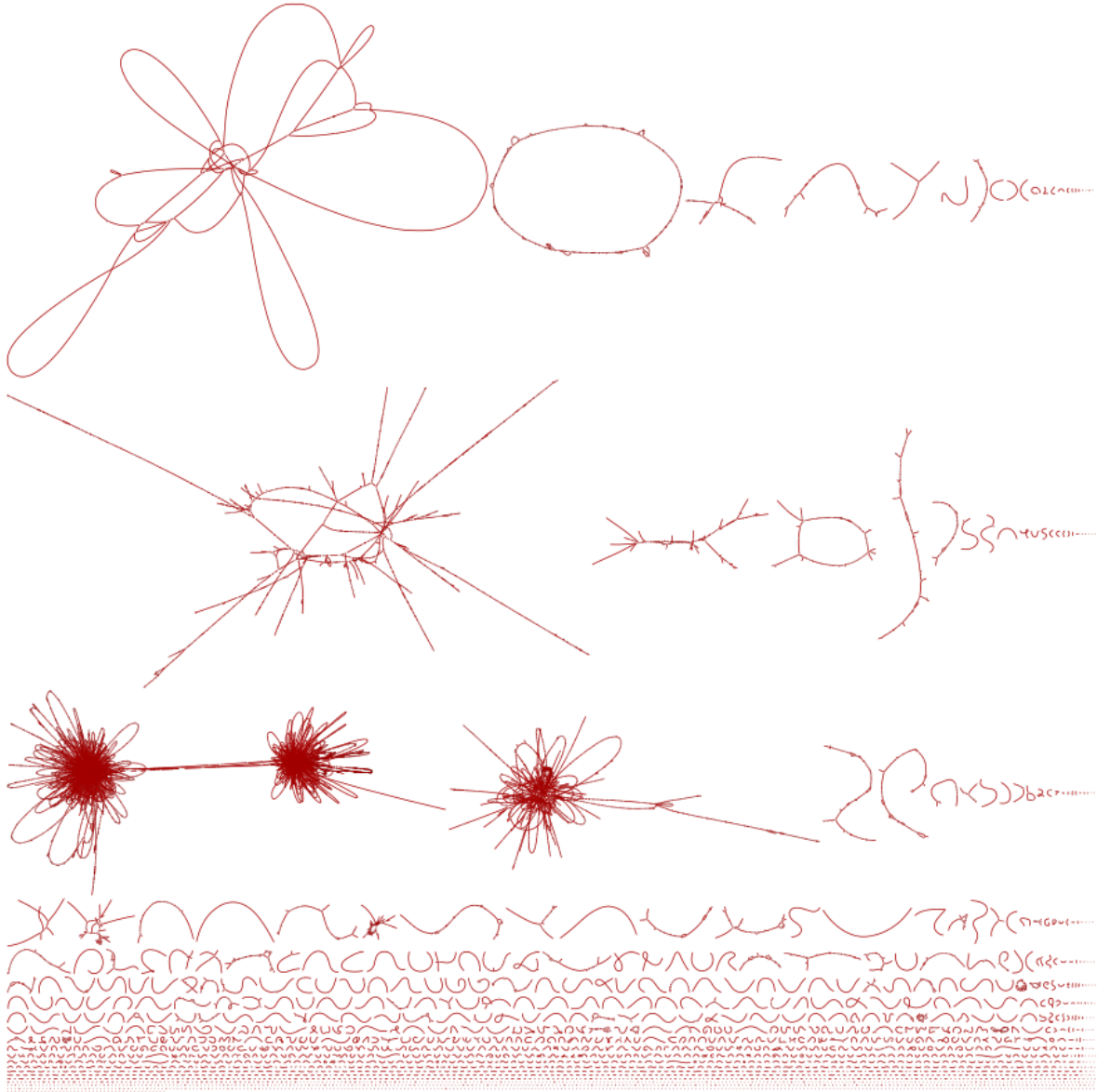
Node labels

Custom Number
Length Coverage
Font Text outline

BLAST

Create/view BLAST search

Query:

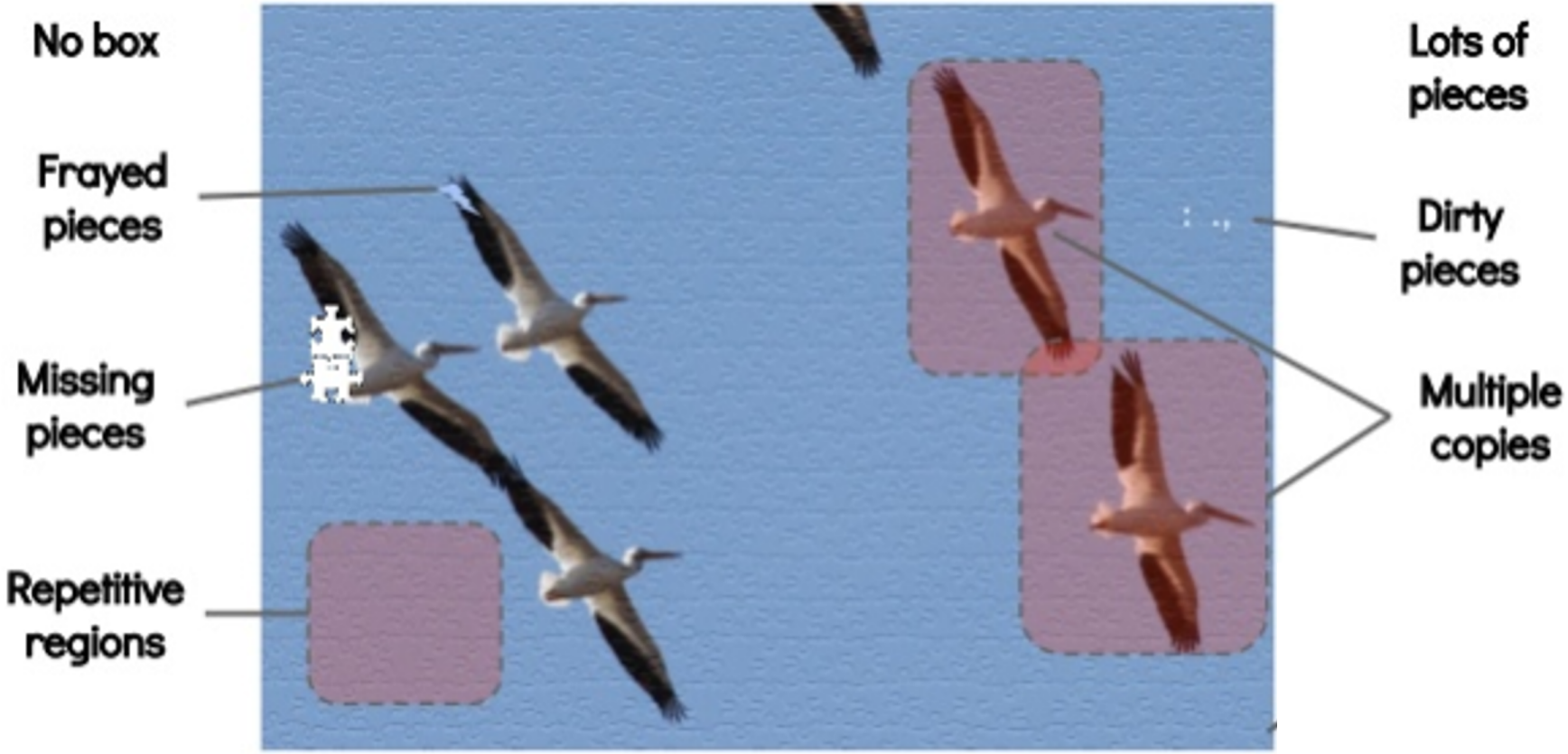


Find nodes

Node(s):

Find node(s)

What makes a puzzle hard?



Slide taken from Torsten Seemann

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

What makes **metagenome** 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
- Conserved genes
- Community composes of strains, species, genera with high similarity
- Diverse environments, how much data is needed?

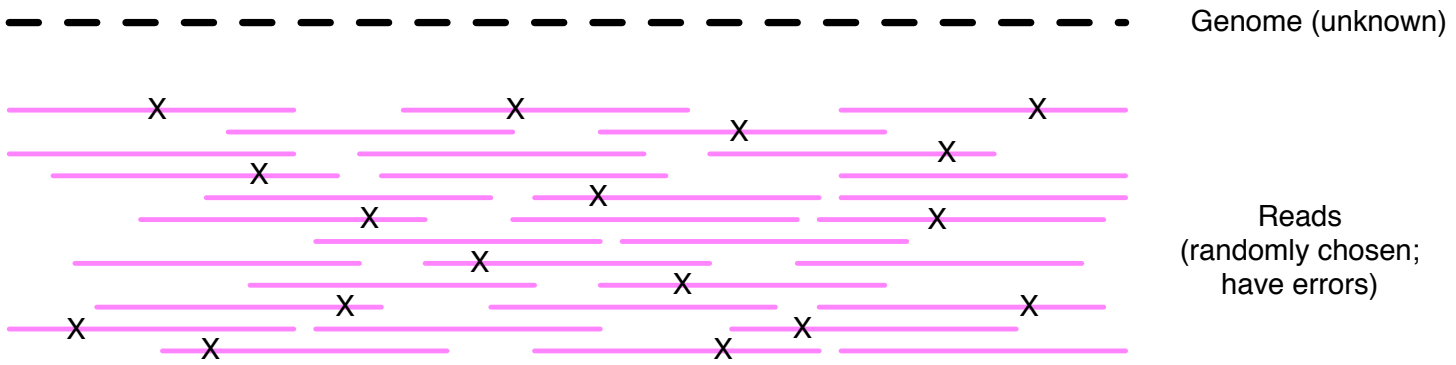
Metagenomic assembly

- Only fraction of reads assembles
  - Does not represent the whole community
- Can and will contain errors
- Testing different assemblers
- More or less sequences
- Co-assembly in some cases

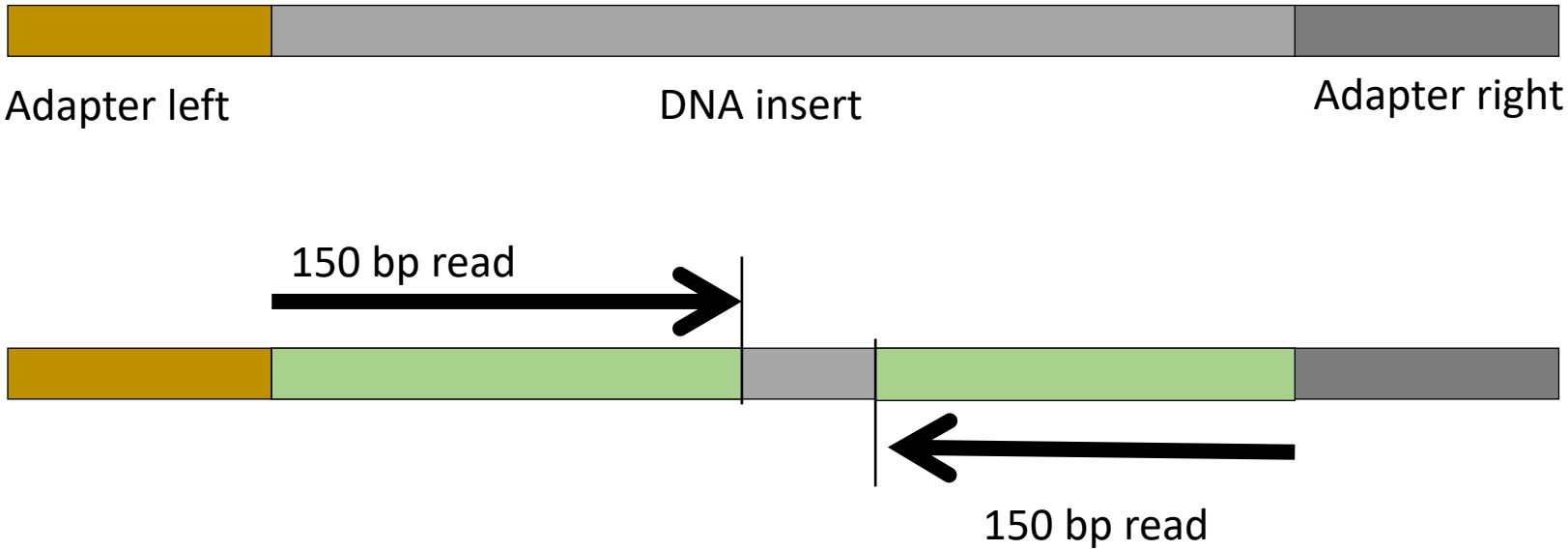
Coverage

- 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