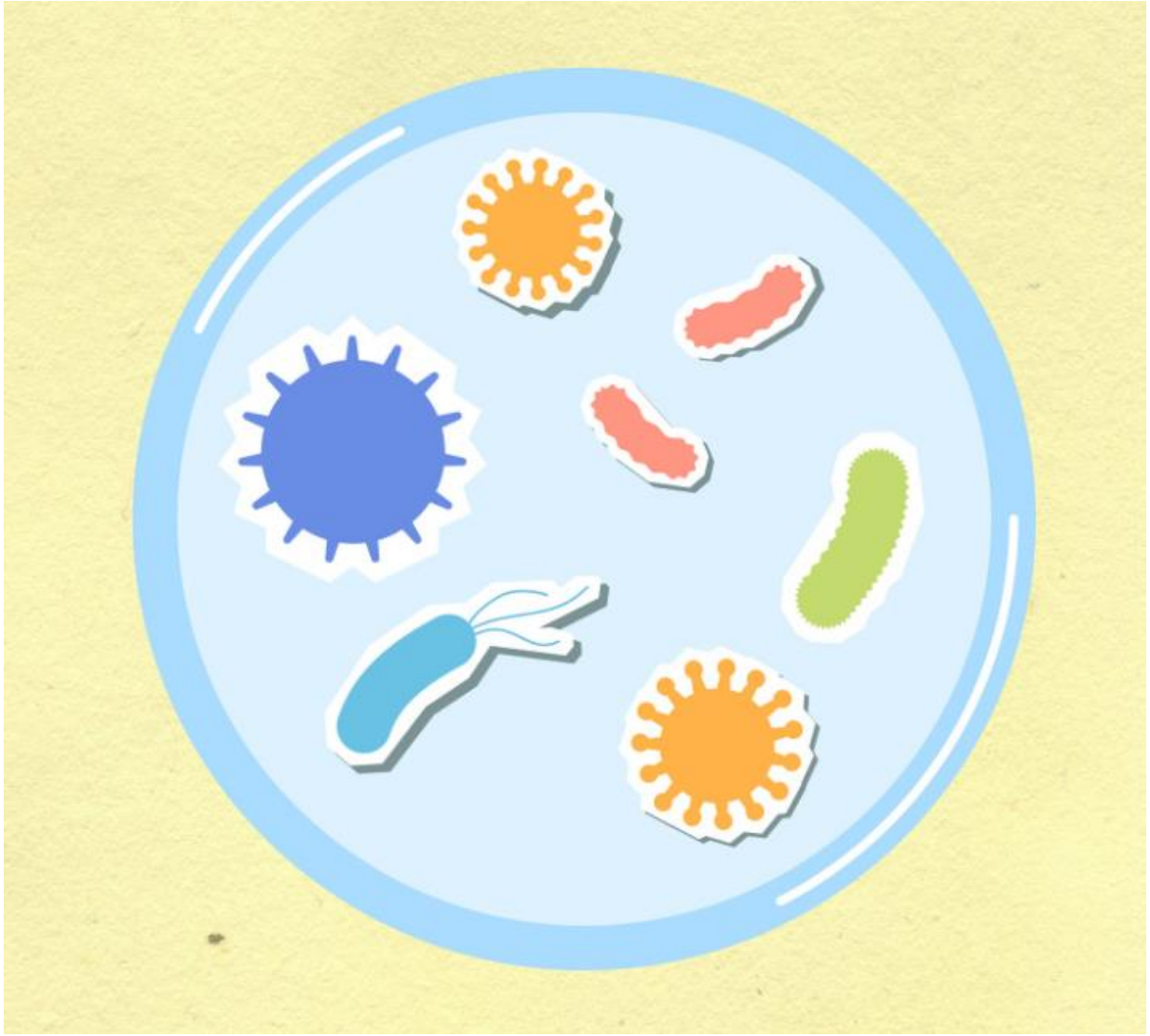

DTU

Inter EURL workshop - 2022

Assembly and assembly statistics

At a glance

- Recap
- De novo assembly vs mapping
- De novo assembly method
  - kmer
  - De bruijn graph
- Assembly statistiscs
  - N50
  - Number of contigs
  - Total base pairs
  - Depth
  - Coverage



Recap

Basecalling

Recap

Basecalling

ADEPTER CCAAGGCCACGTTA GGGGGT

ATGATATTGGCCAA ADEPTER

ADEPTER TTAAGGCCACGTTA ATGGAAA

Trimming

Recap

Basecalling

ADEPTER CCAAGGCCACGTTA GGGGGT

ATGATATTGGCCAA ADEPTER

ADEPTER TTAAGGCCACGTTA ATGGAAA

Trimming

TTAAGGCCACGTTA

Assembly

ATGATATTGGCCAA

CCAAGGCCACGTTA

Recap

Basecalling

ADEPTER CCAAGGCCACGTTA GGGGGT

ATGATATTGGCCAA ADEPTER

ADEPTER TTAAGGCCACGTTA ATGGAAA

Trimming

ATGATATTGGCCAA

CCAAGGCCACGTTA

TTAAGGCCACGTTA

Assembly

TTAAGGCCACGTTA

ATGATATTGGCCAA

CCAAGGCCACGTTA

Recap

Basecalling

ADEPTER CCAAGGCCACGTTA GGGGGT

ATGATATTGGCCAA ADEPTER

ADEPTER TTAAGGCCACGTTA ATGGAAA

Trimming

TTAAGGCCACGTTA

ATGATATTGGCCAA

CCAAGGCCACGTTA

Assembly

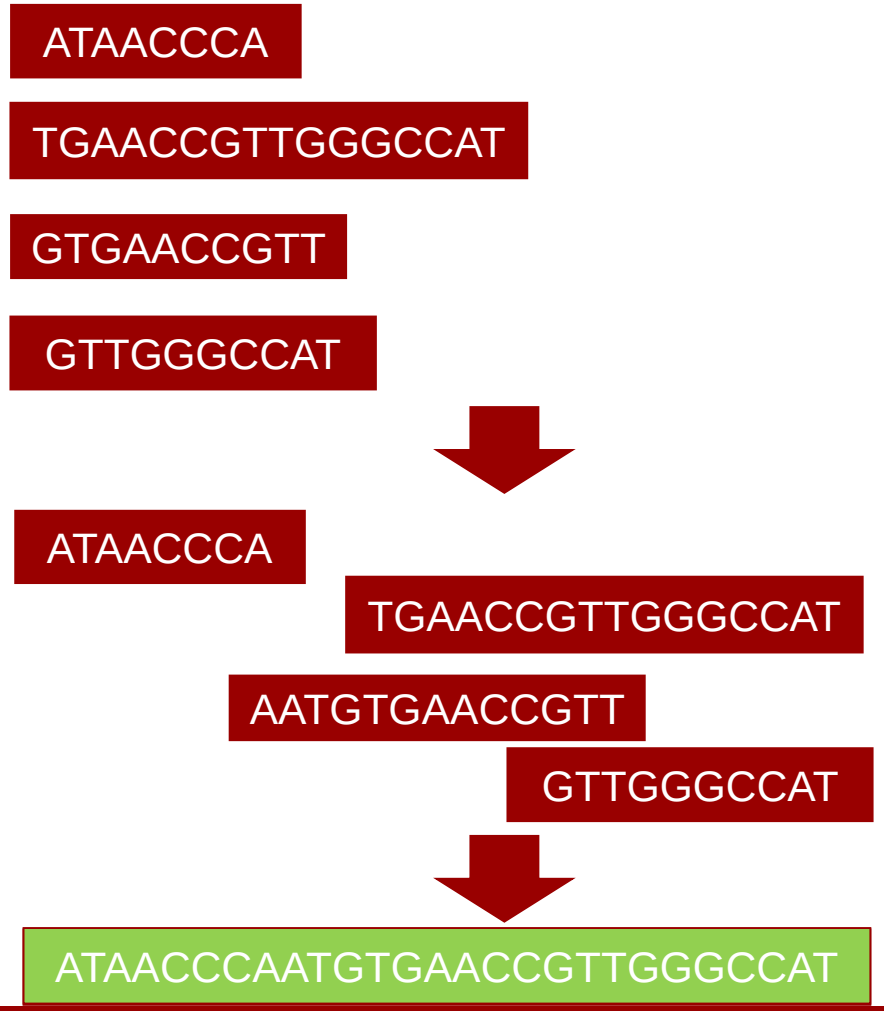
ATGATATTGGCCAA

CCAAGGCCACGTTA

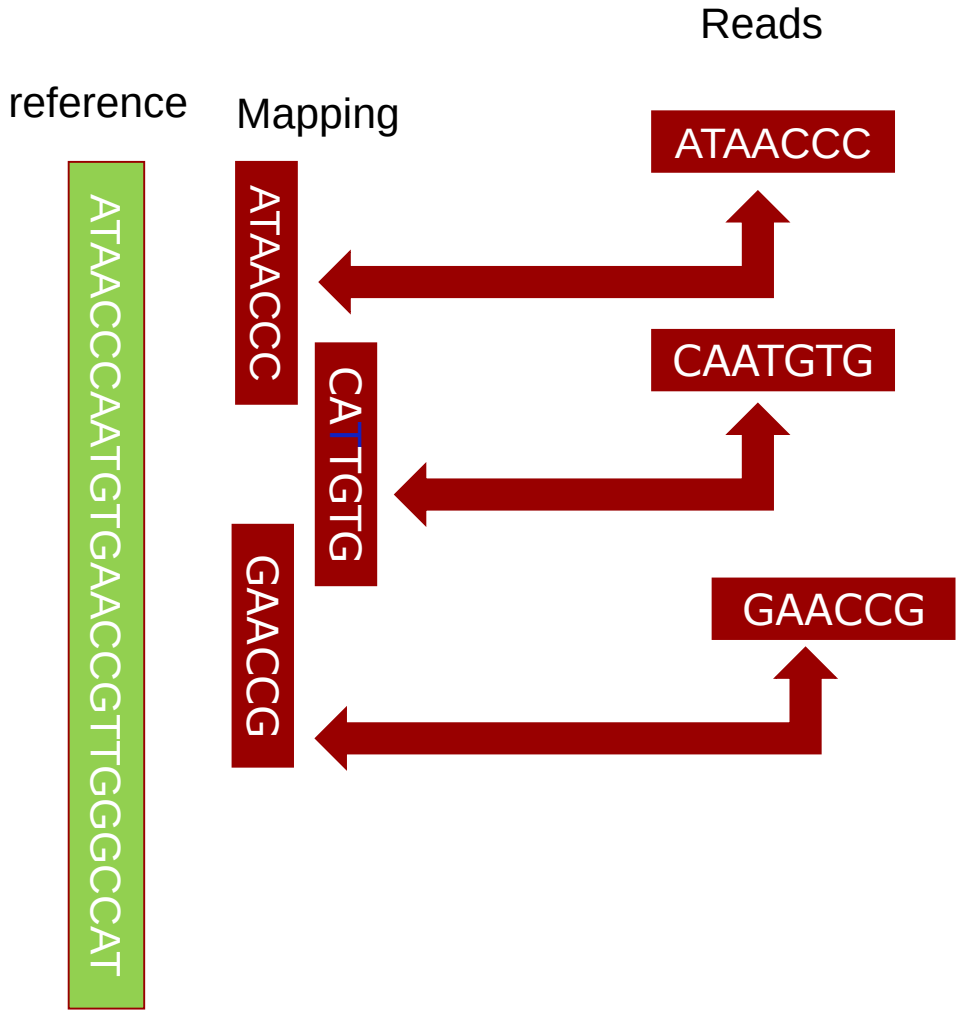
TTAAGGCCACGTTA

ATGATATTGGCCAAGGCCACGTTAAGGCCACGTTA

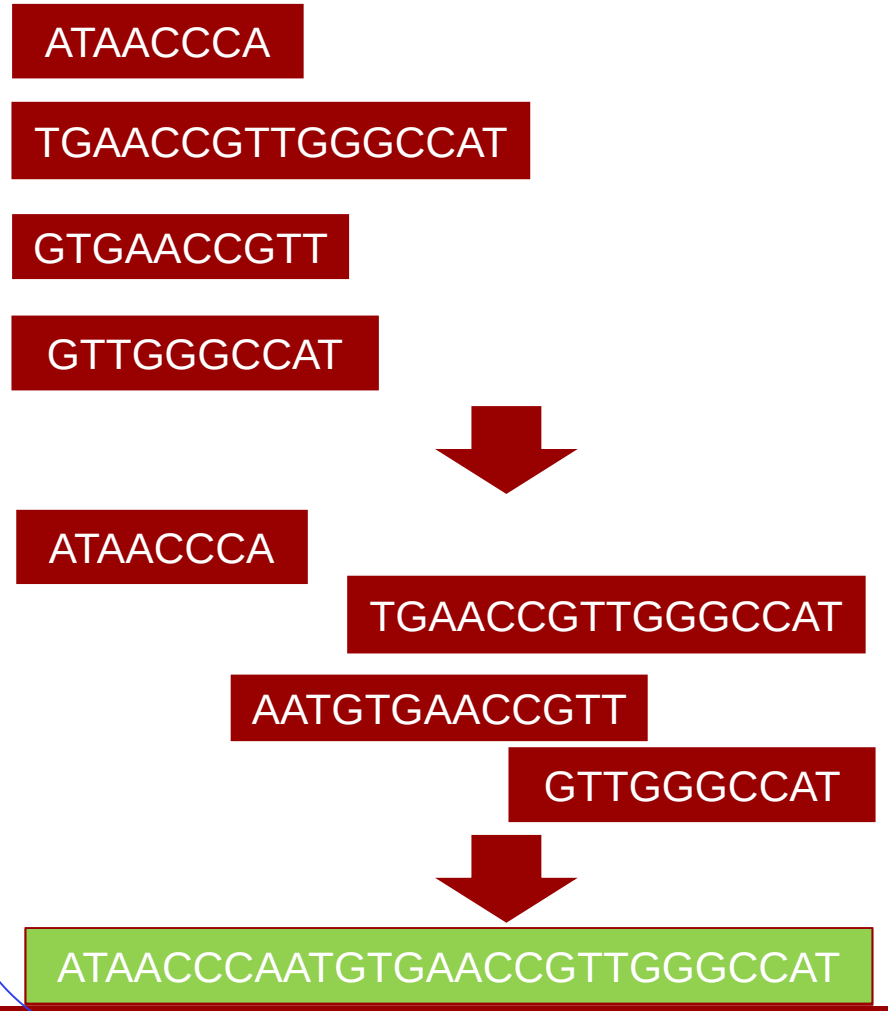
De novo assembly



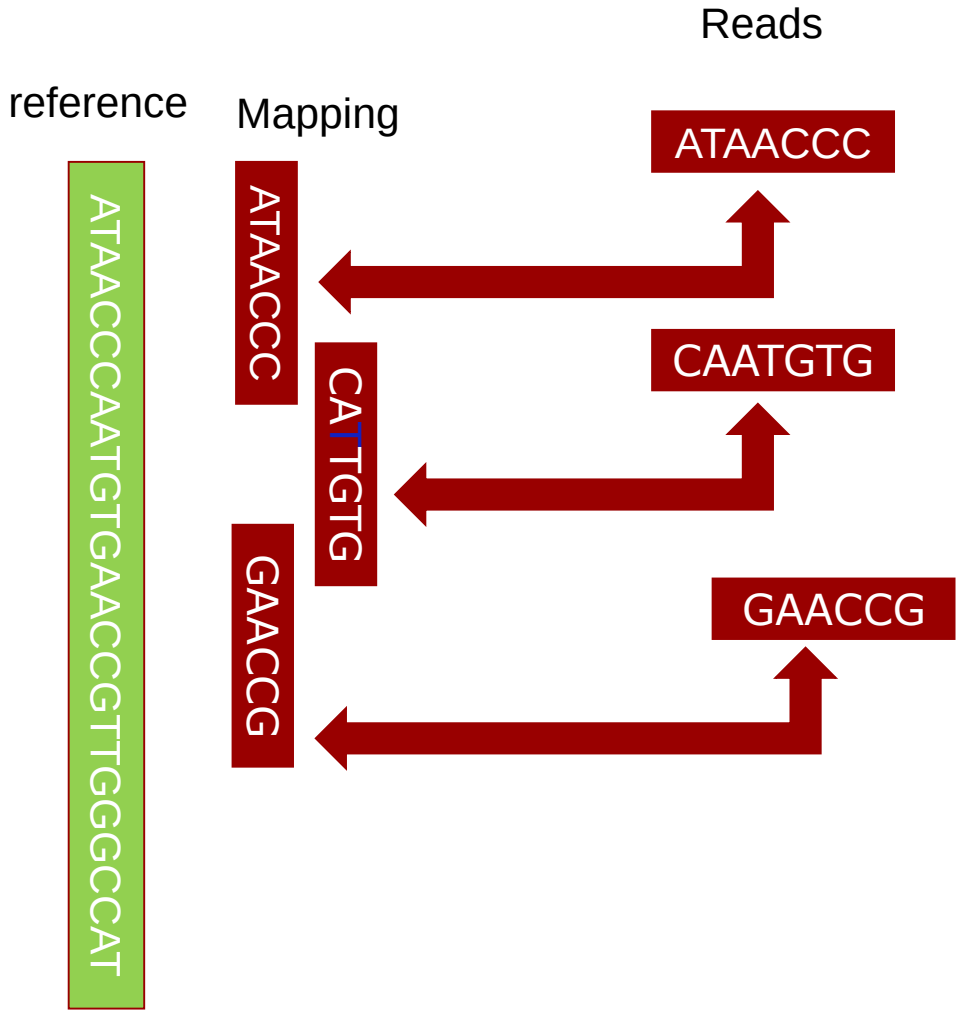
Mapping



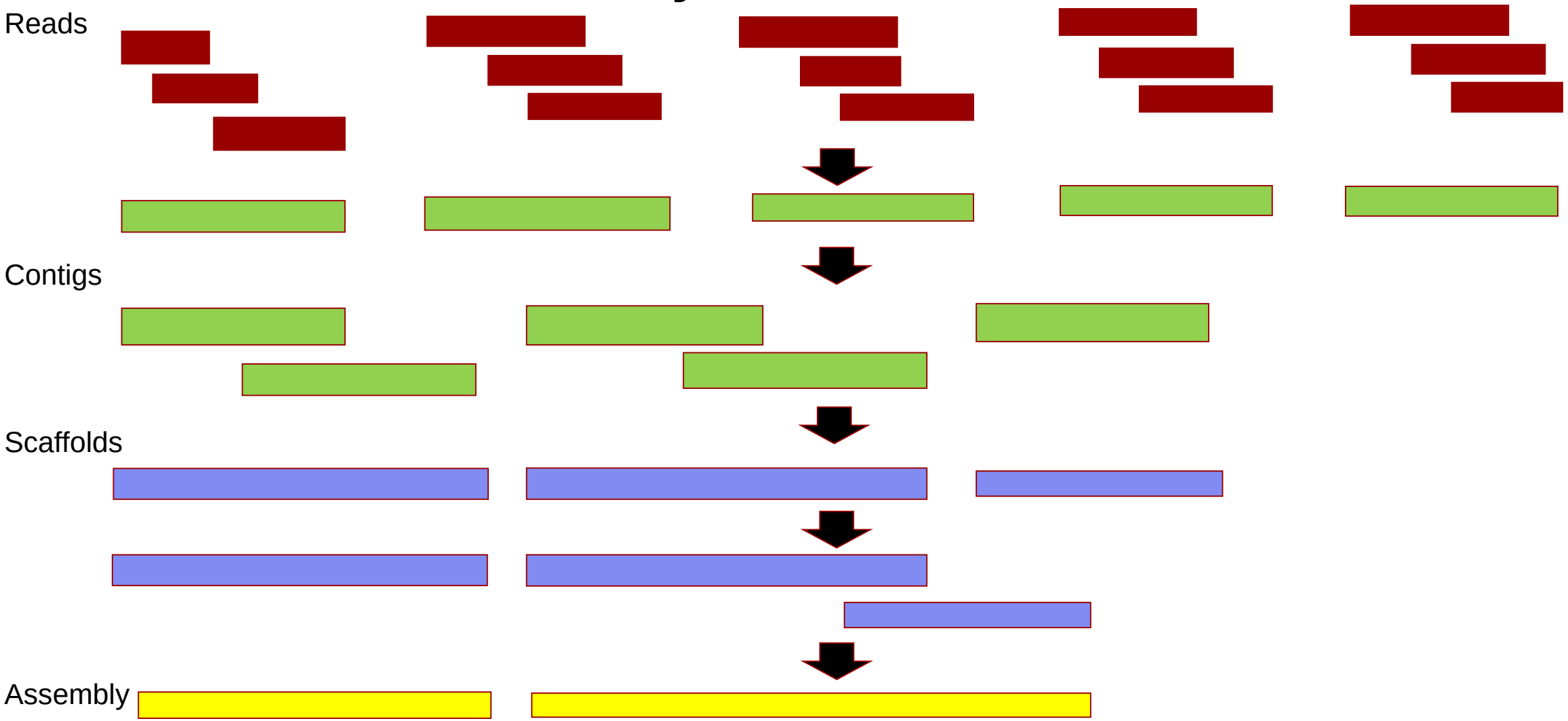
De novo assembly



Mapping

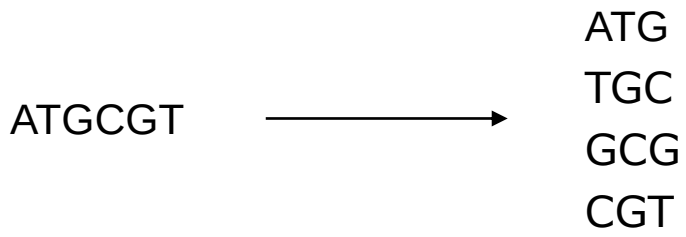


De novo assembly



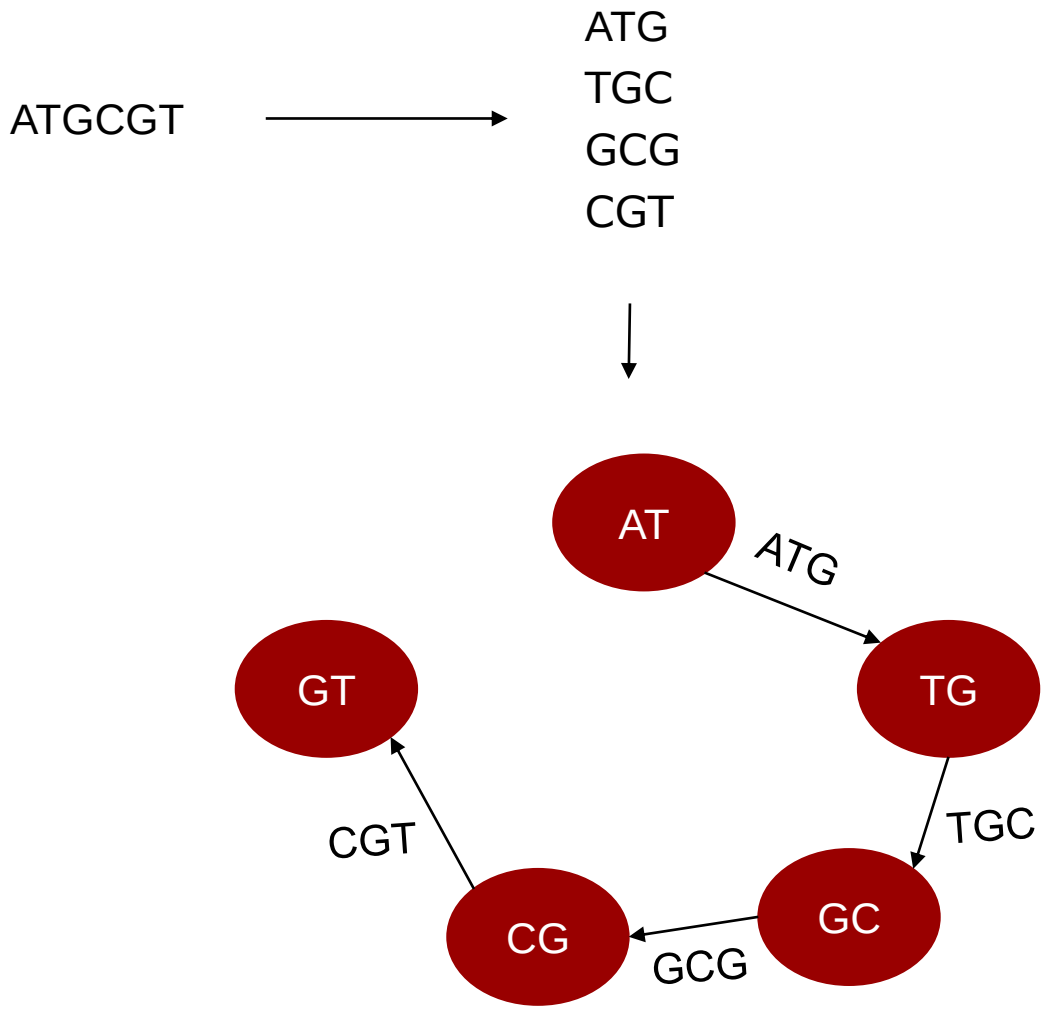
De novo assembly using de Bruijn graphs

- Assemblers work by cutting the read into kmers
- De Bruijn graph is constructed using kmers
- Repeated for more reads
- The most likely genome is constructed by joining all nodes, by traveling each edge only ones



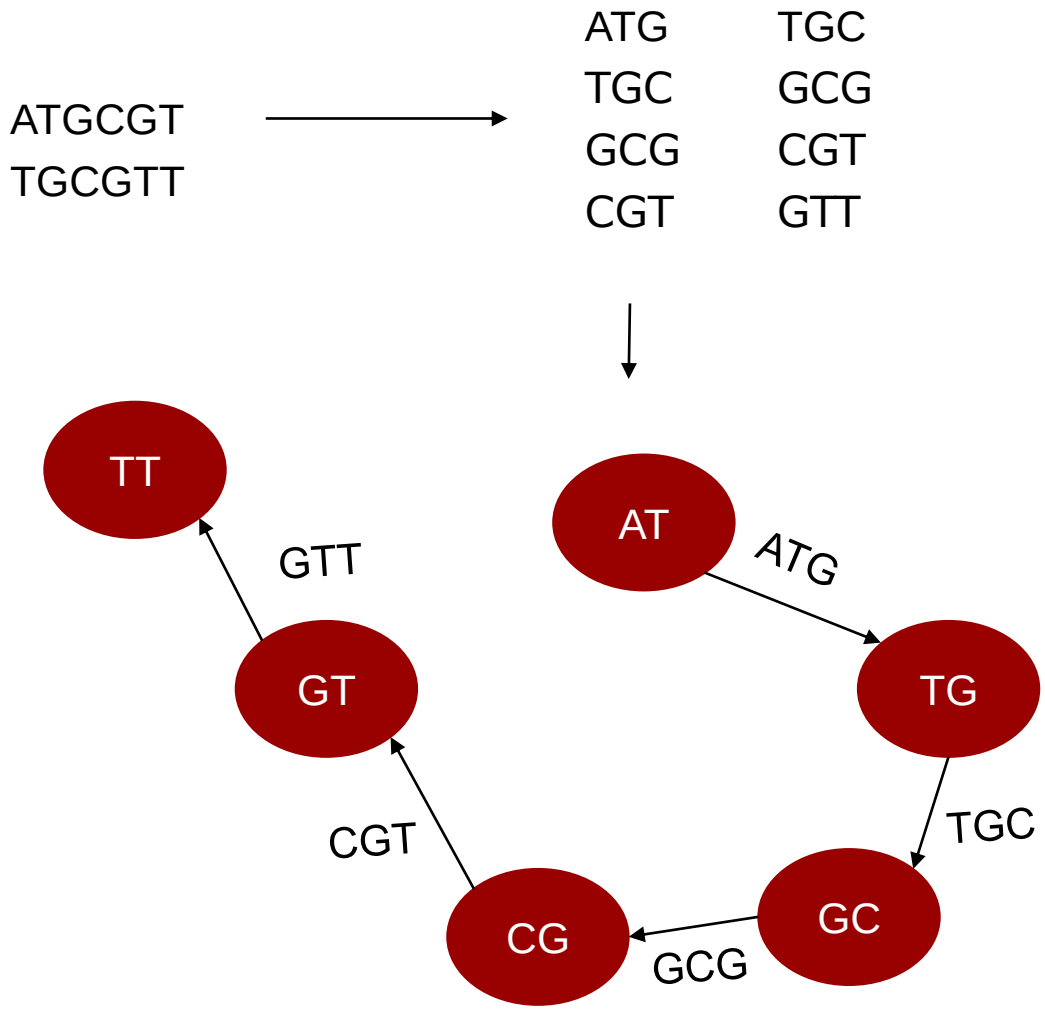
De novo assembly using de Bruijn graphs

- Assemblers work by cutting the read into kmers
- De Bruijn graph is constructed using kmers
- Repeated for more reads
- The most likely genome is constructed by joining all nodes, by traveling each edge only ones



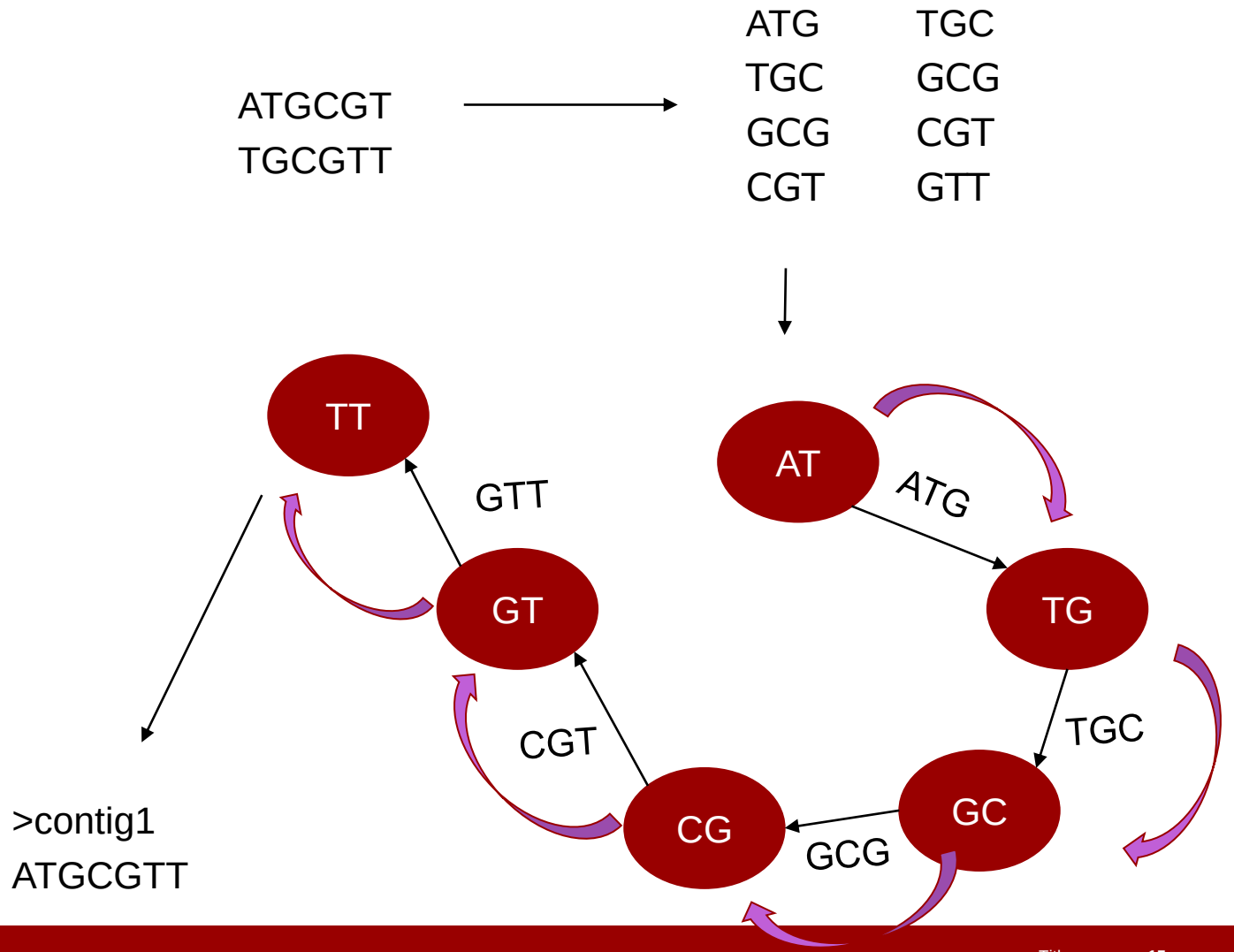
De novo assembly using de Bruijn graphs

- Assemblers work by cutting the read into kmers
- De Bruijn graph is constructed using kmers
- Repeated for more reads
- The most likely genome is constructed by joining all nodes, by traveling each edge only ones



De novo assembly using de Bruijn graphs

- Assemblers work by cutting the read into kmers
- De Bruijn graph is constructed using kmers
- Repeated for more reads
- The most likely genome is constructed by joining all nodes, by traveling each edge only ones



From fastq to fasta

```
@SRR1928200.1 HWI-ST1106:418:D1H56ACXX:2:1207:10978:124033/1
TGCCGAGTGATATCGCTGACGTCATCCTTGAGGGTGAAGTTCAGGTCGTCGAGCAACTCGGCAACGAAACTCAAATCCATATCCAGATCCCTTCCATTCG
+
@@CFFDFBFFHHHJJJIJIJIGGIIJJJGIIHIFBGHIHHHJJIIFGHIGJJJHHHHFFFCCDDDDDDDDDCCCC;:@CDDDDEDDCDDDCDDDC>CDD>
```

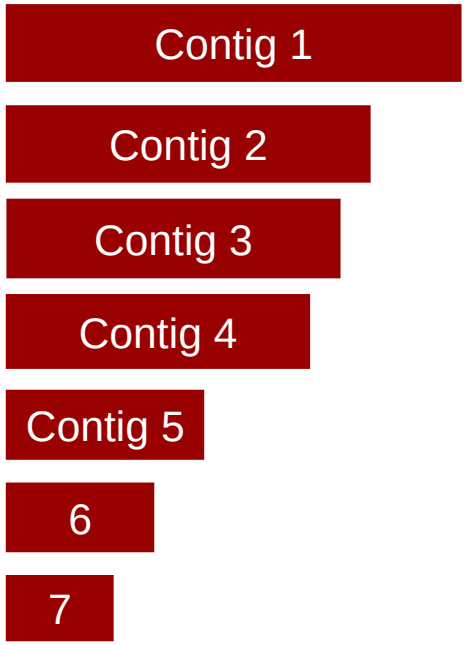
```
>ENA|LR822054|LR822054.1 Citrobacter werkmanii isolate BB1479 genome assembly, plasmid: pCW-CTX-M-15A_
CGTCAGCTTTCCAGTCGACGGCTGATTGAAGTCGGGAATAGCGTCCTTGAAAAGAAGAAC
TTCATTCGAGTTCATCGTGTGGATCCCCCAGTTTTATTGTTATTTTCCGGGTATCTTGGA
ATGCCCAGTCCGGGCGAATGTATCACGGTGATTTTTATTGATCATGAGAAATAGGGGTCA
TTTAGTCCCCATTTATCGGGTATTGGTTTTTATTTGTACTAAATCAATACGTTATTTCAG
AGATGAATCGGATAAATGTCGTTGACATCAAATTTTTGATCTGCTGCCAGTGTGGACAAA
AAATGAATACCGATCACCTATTTTTGAGATTTGTTACGTATGATTATGTTTTTATTTGAT
GTTTTCATTAGCACAGCAGATGTTGATAATTAAGTTCCTTTCCCCTTCCAATCCCACCGT
TATTCCCTTTGAACACCACCAGCTACCAGGCTAACCCCACCGACAGCCCTTCAGAGCTCA
CTTTTTTCCCTCTCAACCCCACCGGGGCAGGTCTTCAGAGCTTACCAGCTGCGGGTTTGC
GGGAGCGGGGATCTTTTTGGTTCTATTTGGTCTTAATCTGGATCGATCTGTTGATCTACC
```

Assembly statistics – N50

- N50 is found by:
  - Sorting all contigs in assembly from longest to shortest, starting with the longest
  - Adding together the length of the longest contigs until half the assembly is included
  - The length of the last added contig to reach 50% of the assembly is the N50
- N50 gives a measure for how much of the assembly is captured in as few contigs as possible
- The higher the N50, the better the assembly, the better the sequencing

Ref: 5.000.000bp

N50 is calculated from 5.000.000/2 = 2.500.000



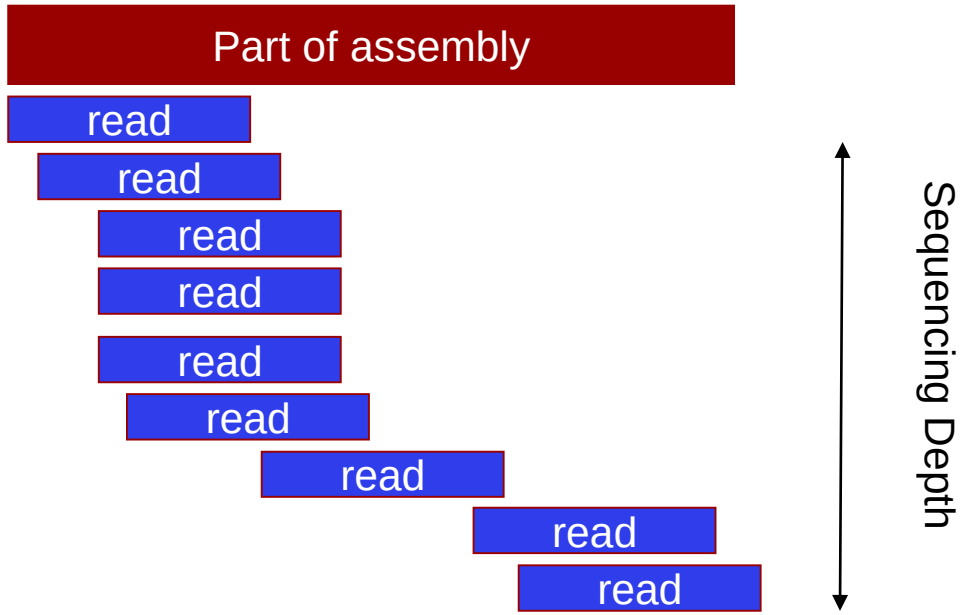
Contig bp	Summed bp
850.000	850.000
700.000	1.650.000
600.000	2.250.000
500.000	2.750.000
400.000	
100.000	
50.000	

Assembly statistics – Depth (Sequence coverage)

- The number of reads that cover a specific part of the assembled genome is called sequencing depth
- Often also called coverage
- The deeper we sequence a part of the genome, the more sure we are about the called bases
- Average coverage would be:

$$sequence\ coverage = \frac{number\ of\ reads\ * average\ read\ length}{Total\ genome\ size}$$

- If a closed reference genome is available the physical coverage can likewise be calculated

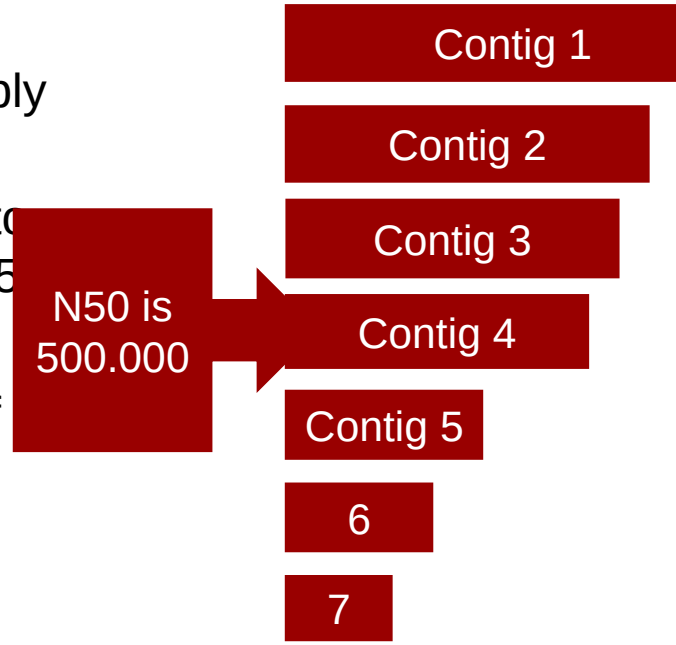


Assembly statistics – N50

- N50 is found by:
  - Sorting all contigs in assembly from longest to shortest, starting with the longest
  - Adding together the length of the longest contigs until half the assembly is included
  - The length of the last added contig to reach 50% of the assembly is the N5
- N50 gives a measure for how much of the assembly is captured in as few contigs as possible
- The higher the N50, the better the assembly, the better the sequencing

Ref: 5.000.000bp

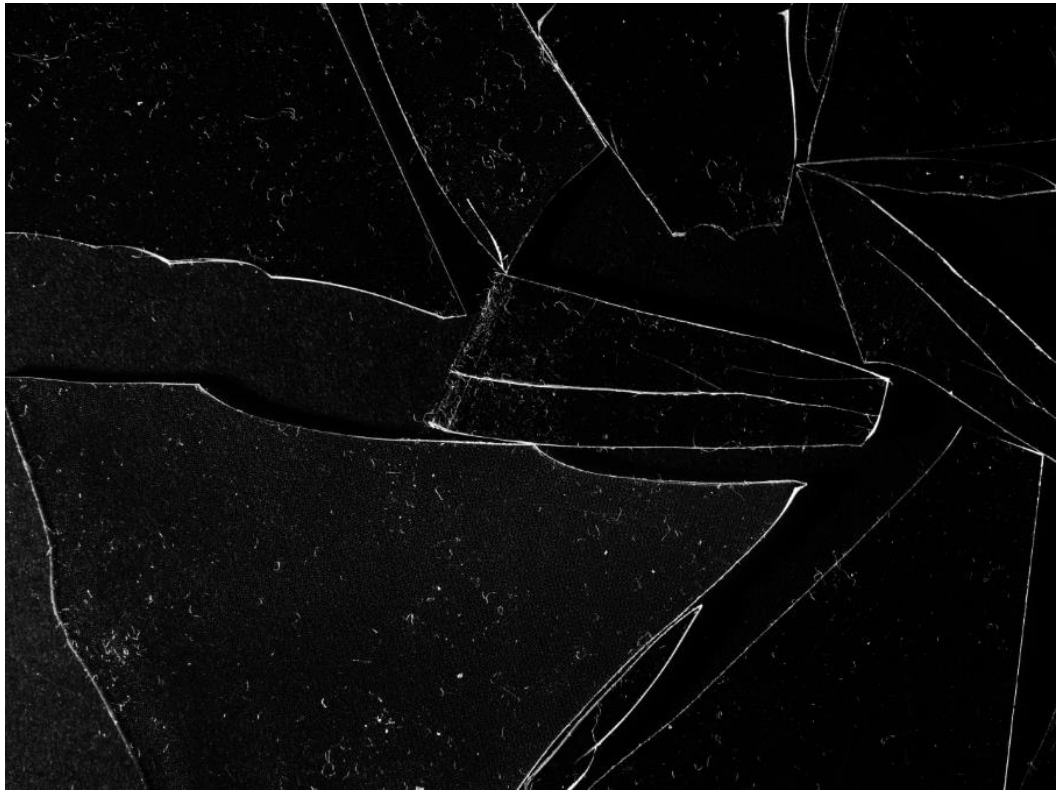
N50 is calculated from 5.000.000/2 = 2.500.000



Contig bp	Summed bp
850.000	850.000
700.000	1.650.000
600.000	2.250.000
500.000	2.750.000
400.000	
100.000	
50.000	

Assembly statistics – number of contigs

- When we assembly we never expect to be able to produce a closed genome (at least not using short read sequencing)
- This is due to several factors including repeated sequences,
- We want the lowest number of contigs possible, as this makes e.g. gene identification and annotation more feasible
- Often, contigs below 200 bp are not counted



Assembly statistics – total base pairs

- Total base pairs are the total length of all contigs in your assembly
- For whole genome sequencing we expect it to be close to the actual size of the genome
- Comparing the total base pairs of an assembly with a reference of the same expected sp. can reveal contamination or misidentification