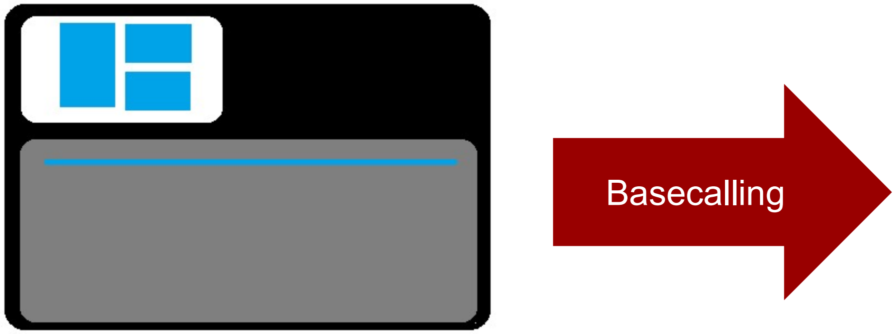
DTU

Inter EURL workshop - 2023

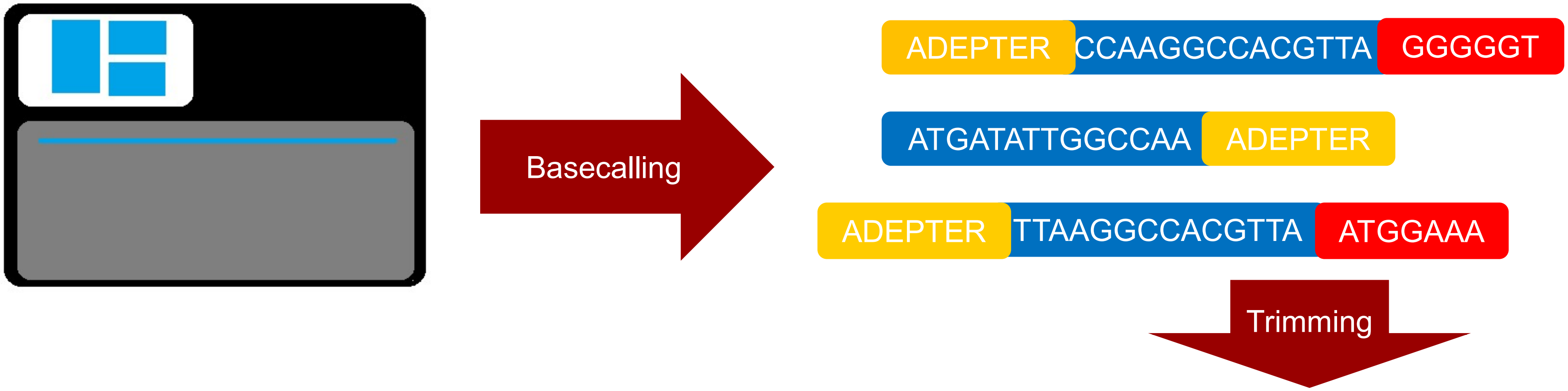
Assembly and assembly statistics

Recap



Basecalling

Recap



Basecalling

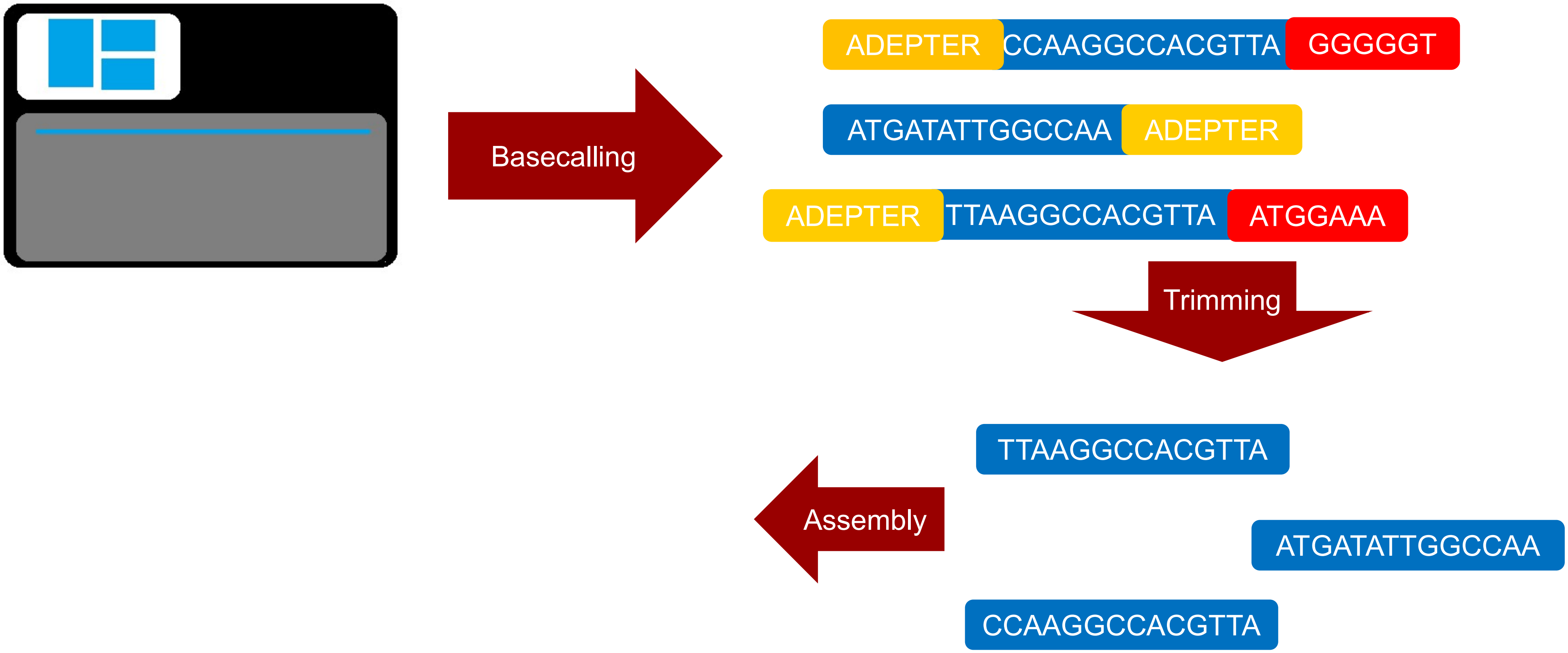
ADEPTER CCAAGGCCACGTTA GGGGGT

ATGATATTGGCCAA ADEPTER

ADEPTER TTAAGGCCACGTTA ATGGAAA

Trimming

Recap



Basecalling

ADEPTER CCAAGGCCACGTTA GGGGGT

ATGATATTGGCCAA ADEPTER

ADEPTER TTAAGGCCACGTTA ATGGAAA

Trimming

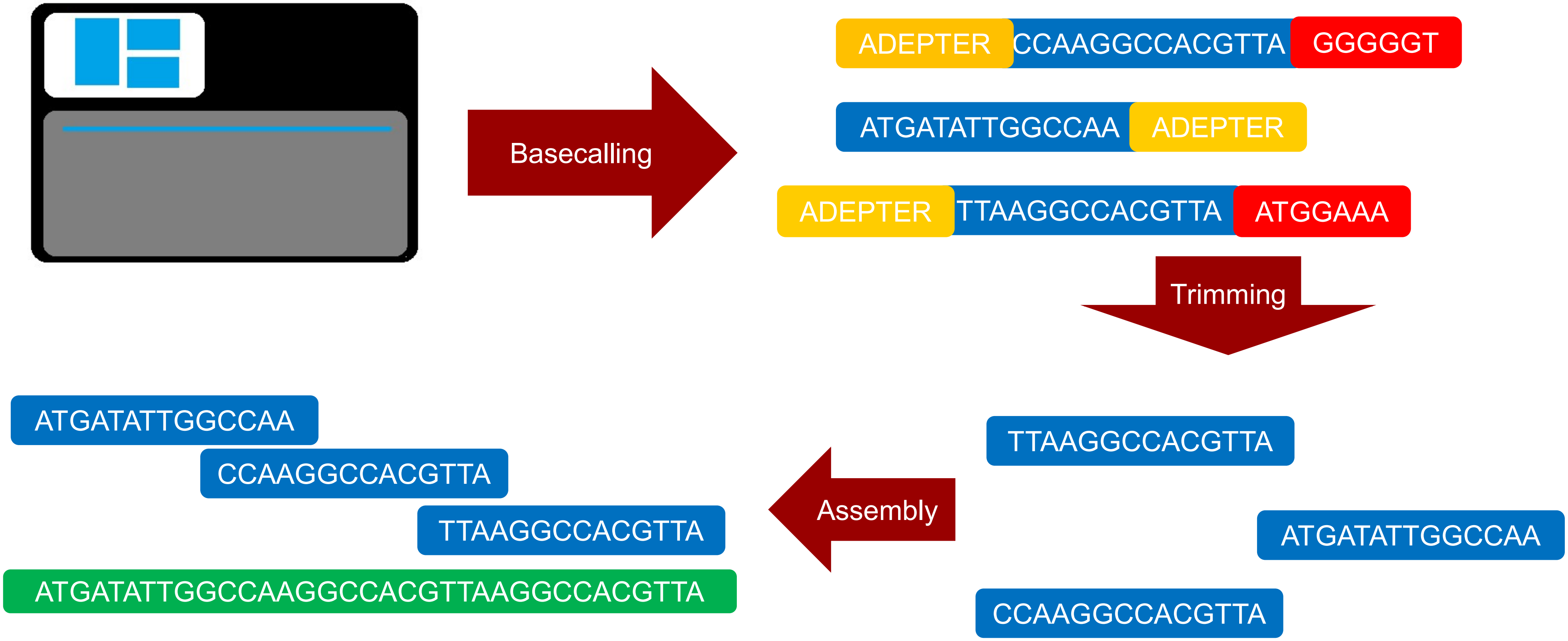
TTAAGGCCACGTTA

ATGATATTGGCCAA

CCAAGGCCACGTTA

Assembly

Recap



Basecalling

ADEPTER CCAAGGCCACGTTA GGGGGT

ATGATATTGGCCAA ADEPTER

ADEPTER TTAAGGCCACGTTA ATGGAAA

Trimming

TTAAGGCCACGTTA

ATGATATTGGCCAA

CCAAGGCCACGTTA

Assembly

ATGATATTGGCCAA

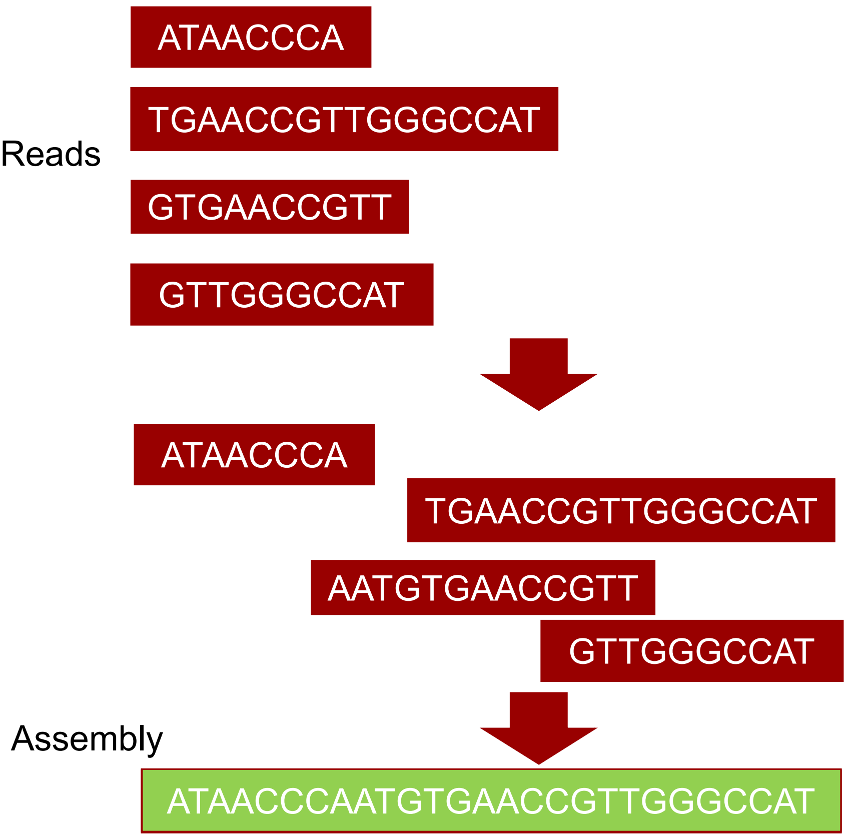
CCAAGGCCACGTTA

TTAAGGCCACGTTA

ATGATATTGGCCAAGGCCACGTTAAGGCCACGTTA

De novo assembly

Genome is unknown and will be constructed from scratch



Reads

ATAACCCA

TGAACCGTTGGGCCAT

GTGAACCGTT

GTTGGGCCAT

ATAACCCA

TGAACCGTTGGGCCAT

AATGTGAACCGTT

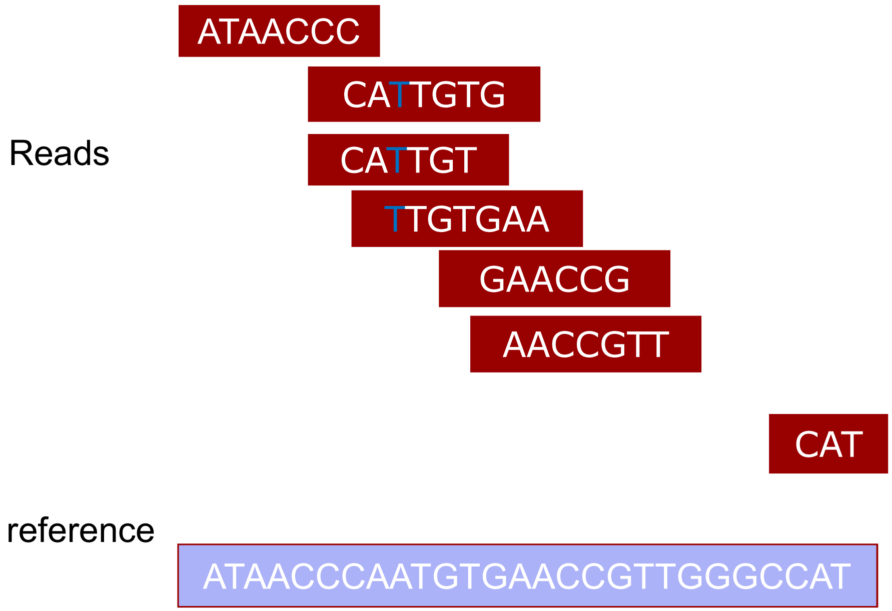
GTTGGGCCAT

Assembly

ATAACCCAATGTGAACCGTTGGGCCAT

Mapping

Genome is expected to be highly similar to previous assembly (the reference genome), we are looking at differences between the two



Reads

ATAACCC

CATTGTG

CATTGT

TTGTGAA

GAACCG

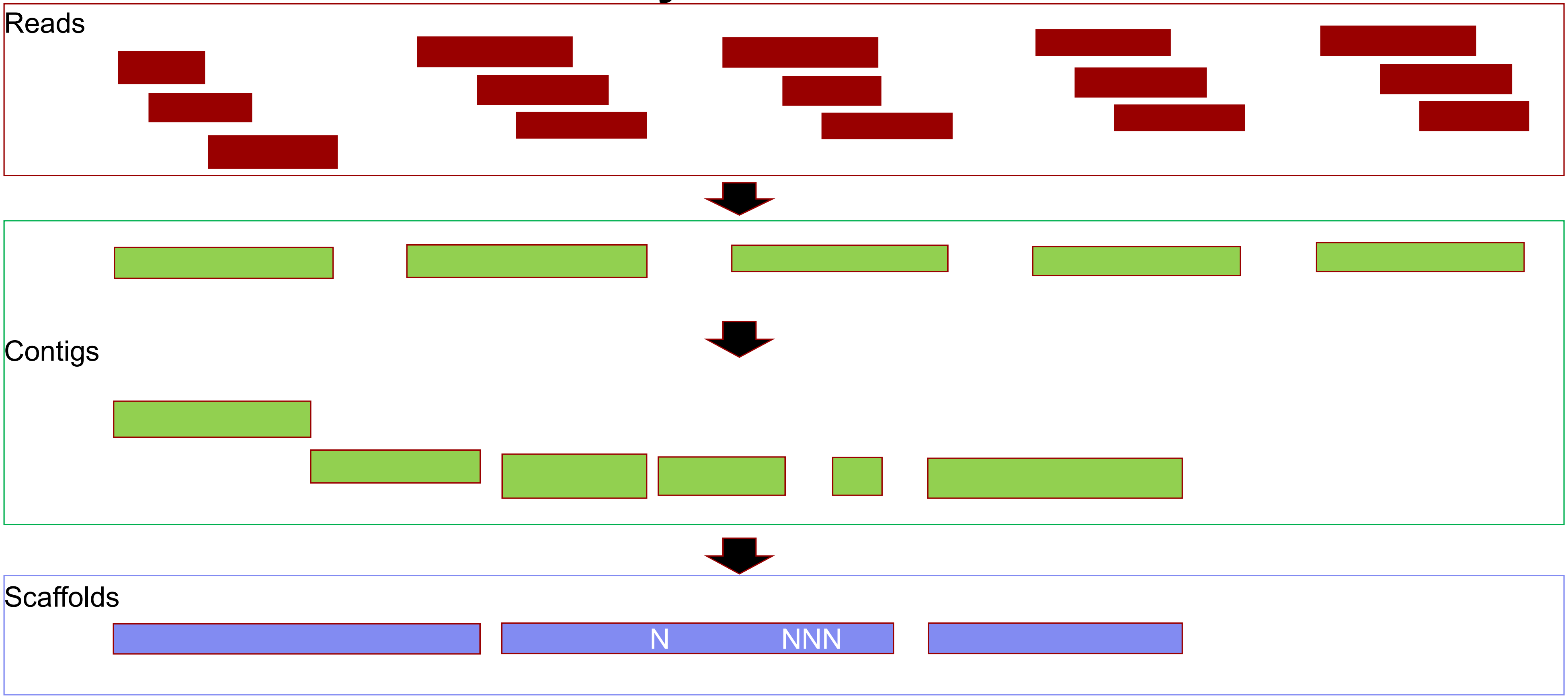
AACCGTT

CAT

reference

ATAACCCAATGTGAACCGTTGGGCCAT

De novo assembly



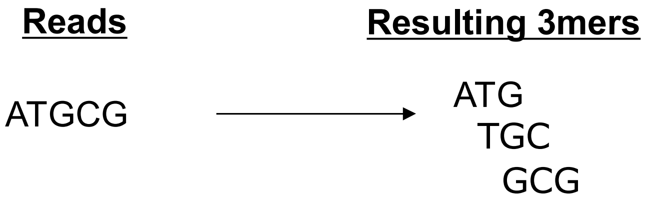
De novo assembly using de Bruijn graphs

- De Bruijn graph is constructed using kmers

- Kmers are obtained by splitting the sequence into overlapping “sub”sequences of length k

- Repeated for more reads

- The most likely genome is constructed by joining all nodes, traveling each edge only ones



Reads

Resulting 3mers

ATGCG

ATG
TGC
GCG

De novo assembly using de Bruijn graphs

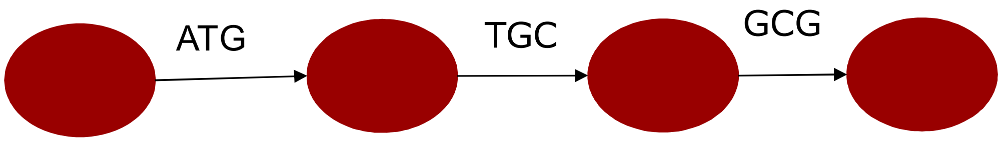
- Example:
  - Reads of 5 bp is split into kmers of length 3 (3mers)
  - De Brujn graph constructed with 3mers as edges

Reads

ATGCG

Resulting 3mers

ATG
TGC
GCG



De novo assembly using de Bruijn graphs

- Example:
  - Reads of 5 bp is split into kmers of length 3 (3mers)
  - De Brujn graph constructed with 3mers as edges
  - Process repeated for new read

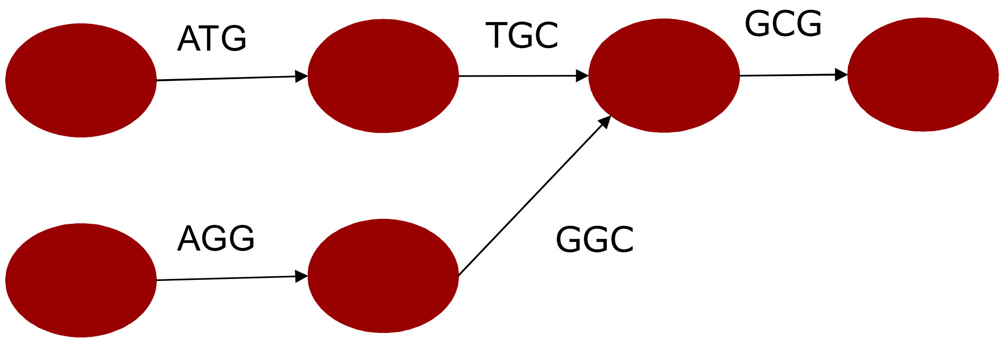
Reads

ATGCG
AGGCG

→

Resulting 3mers

AGG
GGC
GCG



De novo assembly using de Bruijn graphs

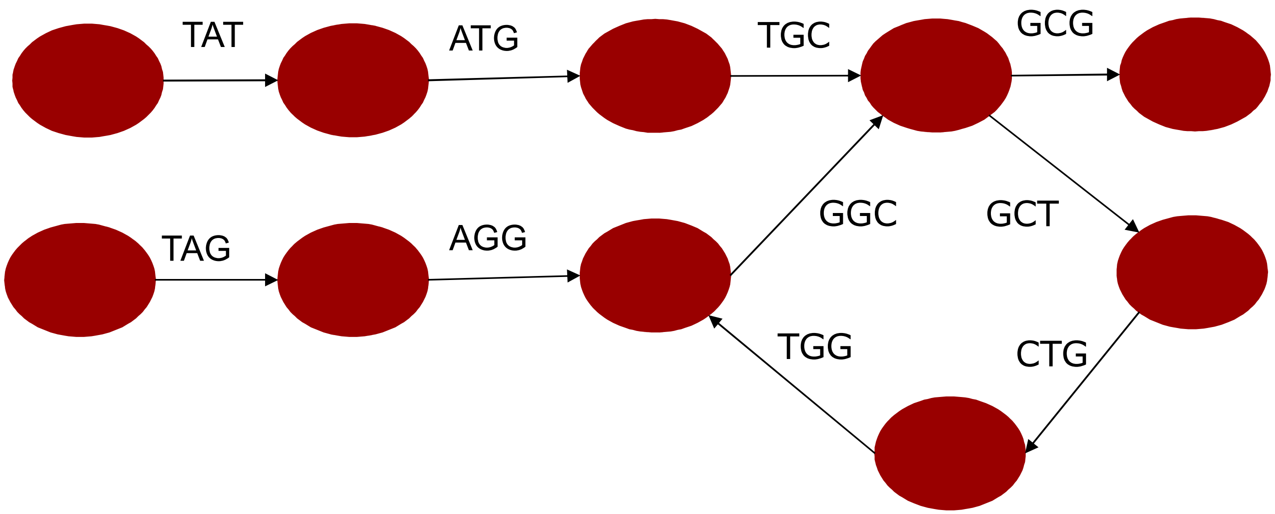
- When all reads have been processed your complete graph is resolved to get contigs
- Different assemblers may vary in how the resolve graphs

Reads

ATGCG
AGGCG
TATGCG
TAGGCG

Resulting 3mers

TAG
AGG
GGC
GCG



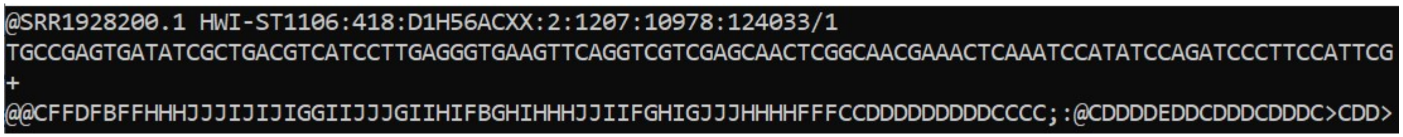
>contig1
TATGCGCG
>contig2
TAGGCG
>contig3
TAGGCTGGC

From fastq to fasta

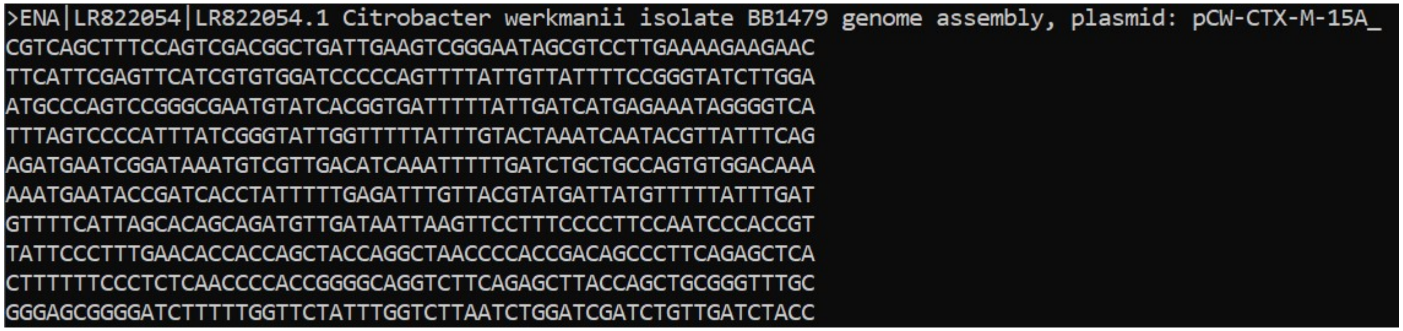
- Read length simplifies graph as longer kmers can be used

- Different assemblers exist
  - SPAdes
  - MEGAHIT
  - Soapdenovo2
  - "skesa"

- Good for different kinds of data, running time, memory, etc



```
@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 and sequencing

Total base pairs in assembly: 5.250.012bp

N50 threshold is 5.250.012/2 = 2.625.006bp

	Contig bp	Summed bp
Contig 1	850.000	850.000
Contig 2	700.000	
Contig 3	600.000	
Contig 4	500.000	
Contig 5	400.000	
6	100.000	
7	50.000	

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 and sequencing

Total base pairs in assembly: 5.250.012bp

N50 threshold is 5.250.012/2 = 2.625.006bp

	Contig bp	Summed bp
Contig 1	850.000	850.000
Contig 2	700.000	1.650.000
Contig 3	600.000	
Contig 4	500.000	
Contig 5	400.000	
6	100.000	
7	50.000	

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 and sequencing

Total base pairs in assembly: 5.250.012bp

N50 threshold is 5.250.012/2 = 2.625.006bp

	Contig bp	Summed bp
Contig 1	850.000	850.000
Contig 2	700.000	1.650.000
Contig 3	600.000	2.250.000
Contig 4	500.000	
Contig 5	400.000	
6	100.000	
7	50.000	

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 and sequencing

Total base pairs in assembly: 5.250.012bp

N50 threshold is 5.250.012/2 = 2.625.006bp

	Contig bp	Summed bp
Contig 1	850.000	850.000
Contig 2	700.000	1.650.000
Contig 3	600.000	2.250.000
Contig 4	500.000	2.750.000
Contig 5	400.000	
6	100.000	
7	50.000	

Assembly statistics – N50

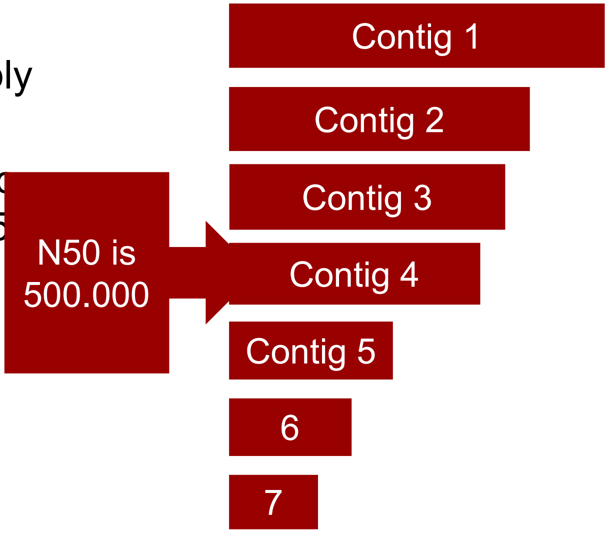
- N50 is found by:
  - Sorting all contigs in assembly from longest to shortest, starting with the longest
  - Adding together the length of the longest contigs until half the assembly is included
  - The length of the last added contig t reach 50% of the assembly is the N5

- 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 and sequencing

Total base pairs in assembly: 5.250.012bp

N50 threshold is 5.250.012/2 = 2.625.006bp



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

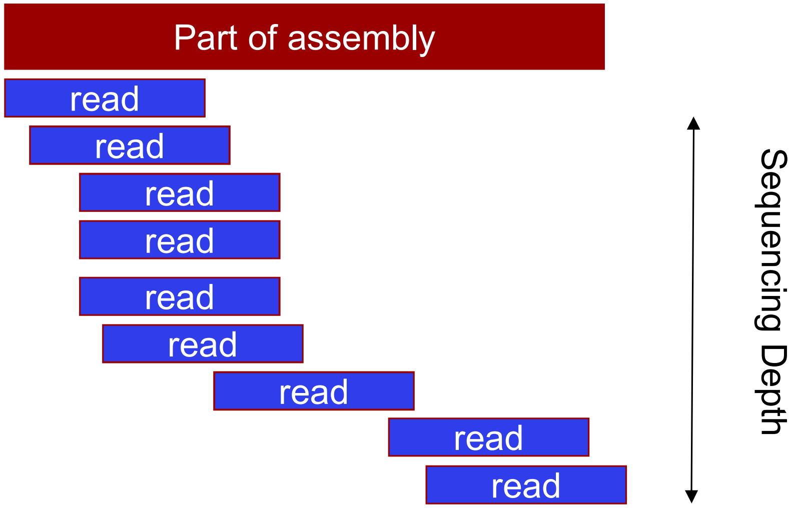
- Often also called coverage

- 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



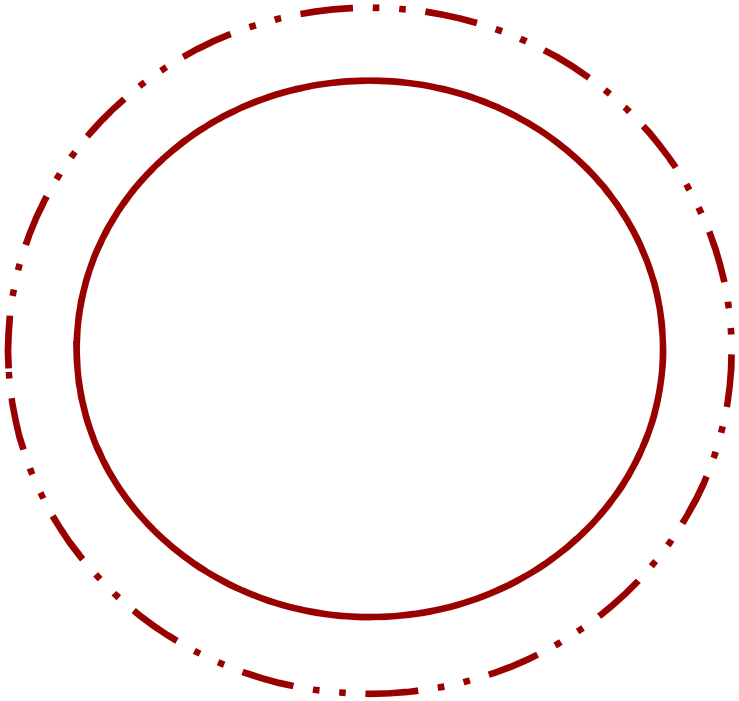
Assembly statistics – number of contigs

- 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 200/500bp 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

- Rule of thumb: within range of sp. or less than 5-10% deviation.

Assembly statistics – (genomic) coverage

- Percentage of the genome covered by reads

- Mainly when reference is available, but can be applied to de novo assemblies

- Setting minimal depth makes metric more reliable as low coverage regions will be sensitive to sequencing error

- Consider using minimal depth e.g. 10x

Thank you

- Lauge Holm Sørensen
  - – You can contact me on lahoso@food.dtu.dtk