Exercises in WGS analysis and the cge tools

EQAsia

Hello,

- My name is Lauge
- Bioinformatician in the research group Global Capacity Building, National Food institute at the Technical University of Denmark
- Email: lahoso@food.dtu.dk

Bioinformatics

- In bioinformatics there are several disciplines which each have their own demands in analysis
  - Metagenomics
  - Whole genome sequencing
  - Expression analysis

- We will focus on Whole genome sequencing (WGS), specifically in bacteria
  - WGS usually have a few requirements
    - The isolate we sequence should be a single organism
    - The isolate should contain a single individual of that population, meaning samples need to be pure (for bacteria a single strain)
    - We must aim to capture “almost” everything in the genome
    - The strain we look at is stable

WGS

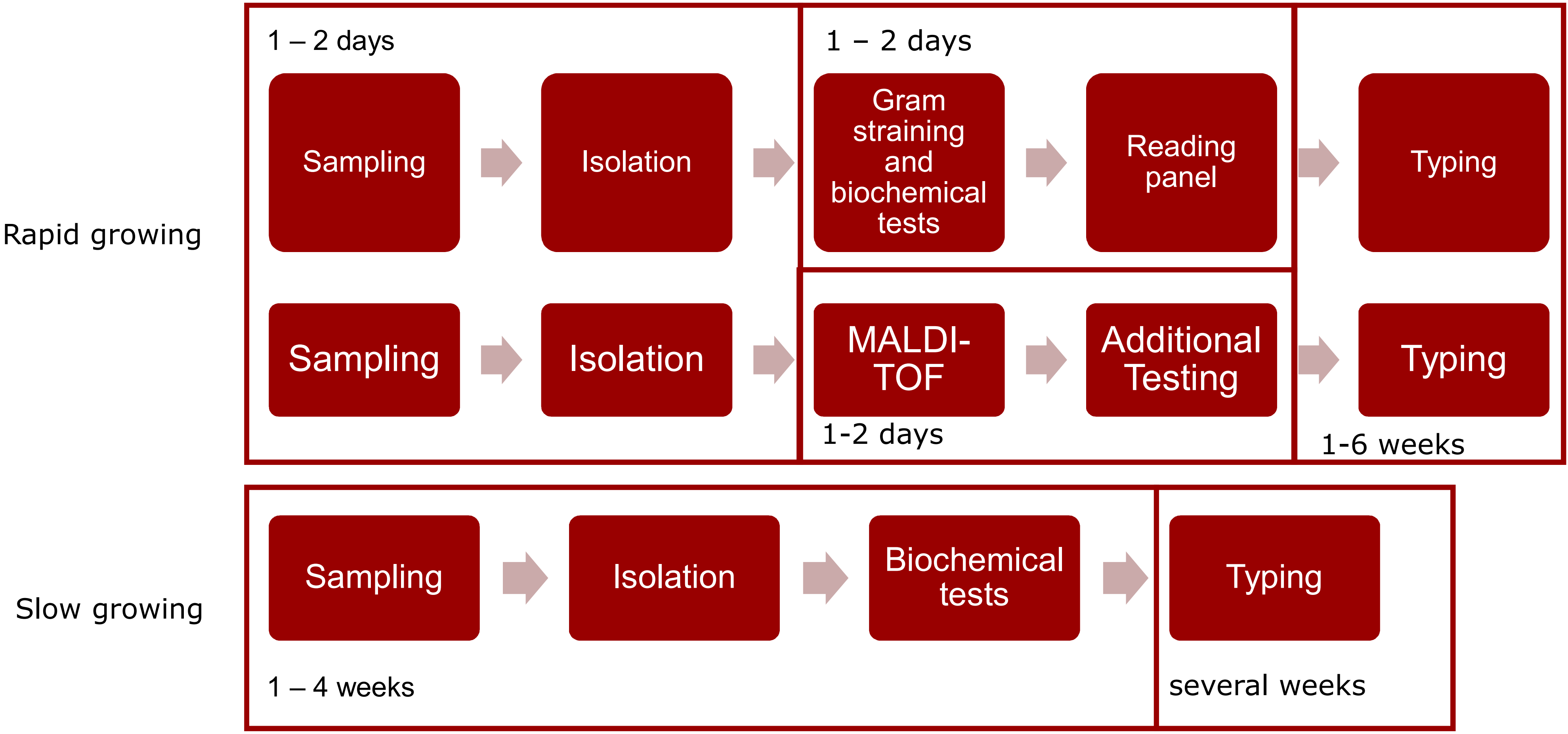
- Sequencing is the process of reading a stretch of DNA, producing a ordered combination of its constituent A, T, G and C

- Whole genome sequencing aims at capturing the entire genetic repertoire
  - All genes of interest if sequenced
  - Additional screening is rapid
  - Facilitates future research

- Ferrer et al. 2014 found a 1% increase in mortality per hour treatment was delayed after sepsis

Ferrer R, Martin-Loeches I, Phillips G, Osborn TM, Townsend S, Dellinger RP, Artigas A, Schorr C, Levy MM. Empiric antibiotic treatment reduces mortality in severe sepsis and septic shock from the first hour: results from a guideline-based performance improvement program. Crit Care Med. 2014 Aug;42(8):1749-55. doi: 10.1097/CCM.0000000000000330. PMID: 24717459.

Overview timeframe



Rapid growing

1 – 2 days

Sampling → Isolation →

1 – 2 days

Gram straining and biochemical tests → Reading panel → Typing

Sampling → Isolation → MALDI-TOF → Additional Testing → Typing

1-2 days

1-6 weeks

Slow growing

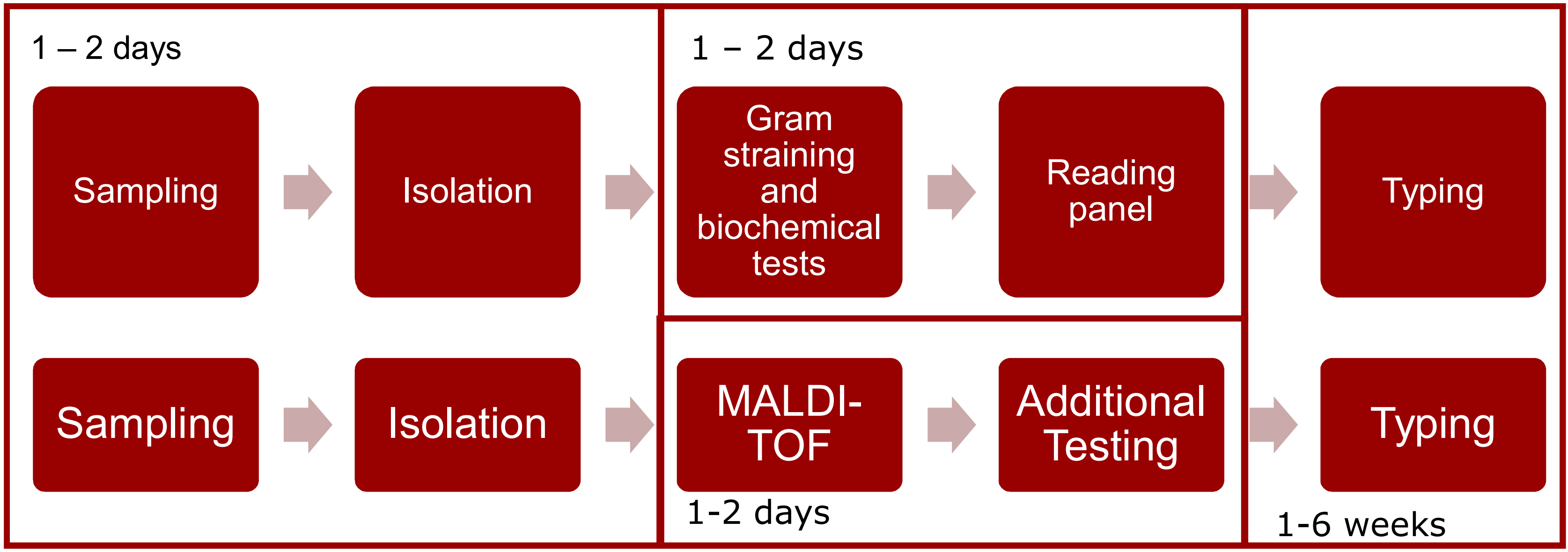
Sampling → Isolation → Biochemical tests → Typing

1 – 4 weeks

several weeks

Overview timeframe

Rapid biochemical methods



1 – 2 days

Sampling → Isolation →

1 – 2 days

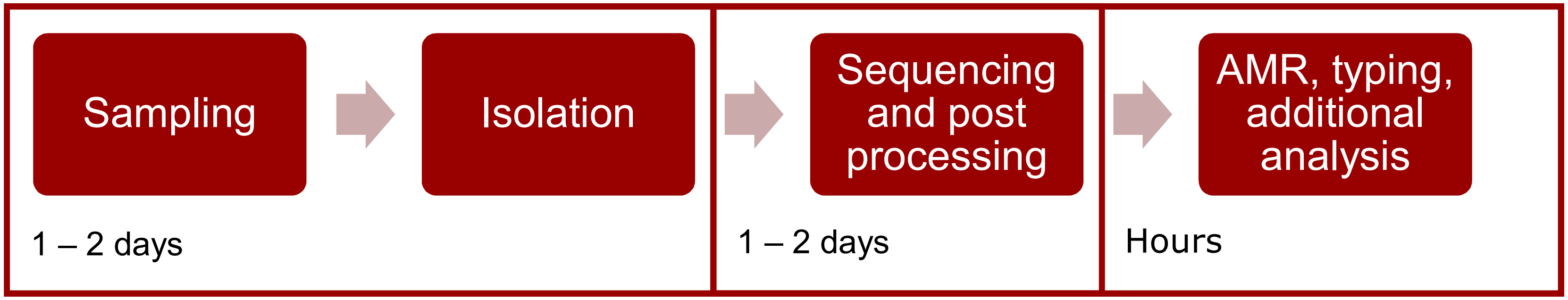
Gram straining and biochemical tests → Reading panel → Typing

Sampling → Isolation → MALDI-TOF → Additional Testing → Typing

1-2 days

1-6 weeks

Whole genome sequencing



Sampling → Isolation → Sequencing and post processing → AMR, typing, additional analysis

1 – 2 days

1 – 2 days

Hours

Good and bad points

Pros	Cons
Captures a lot of information: We aim to capture all the genetic information of your strain	Storage: large amounts of data requires large harddrives
Additional analysis is easy to conduct: including for future research	Costs: machines are expensive and so are reagents (possible less so with new long reads sequencing)
High resolution: We can estimate the phylogenetic relationship between strains at a very in-depth level	CPU power: Programs demand computing power
Relatively fast	Previous knowledge: databases need a solid foundation of knowledge to be precise
Scalable: good if surveillance needs to be expanded	

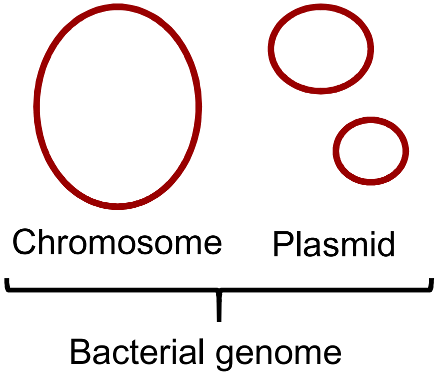
Sequencing, a field in rapid evolution

- Long read sequencing (Oxford Nanopore, Illumina infinity?)
  - Ability to rapidly sequence and analyse data real time
  - Minimal equipment
  - Machines are more affordable
  - Longer reads means assembly is less time and CPU consuming

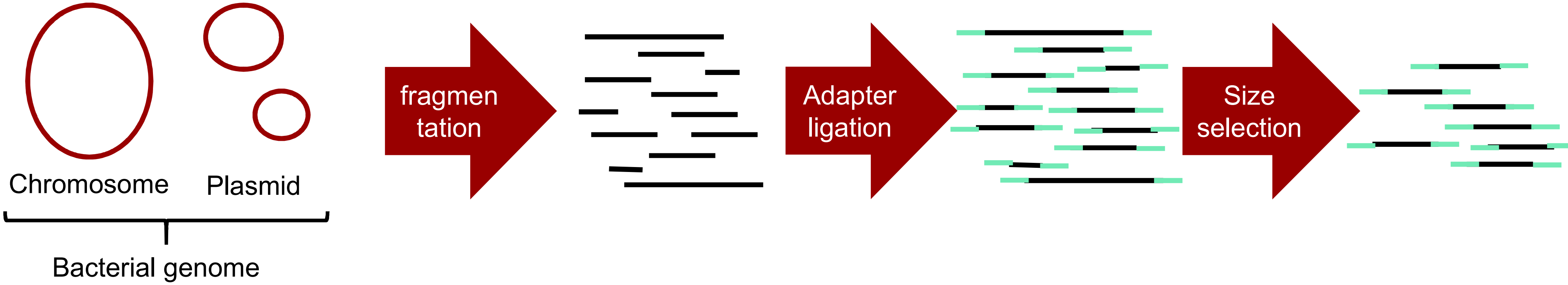
- Facilitates surveillance of mobile genetic elements

- One major challenges consists of high error rate in reads, which have recently been brought down from around 8% to 1%

Summarized library preparation



Summarized library preparation (illumina paired end)

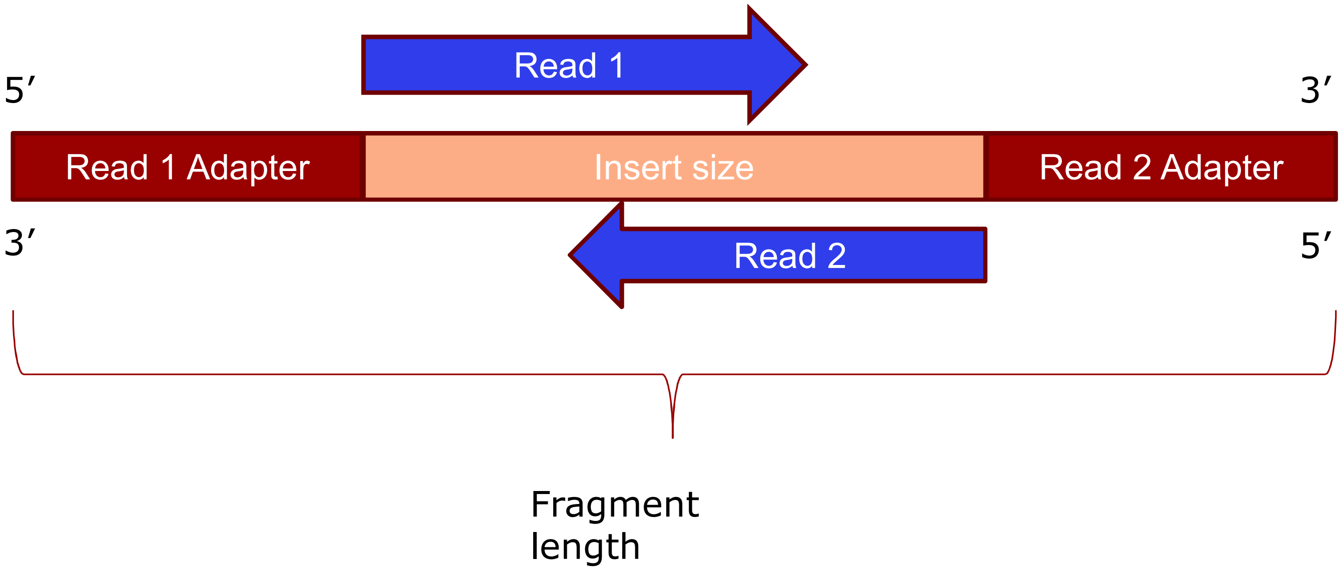


DNA Fragment

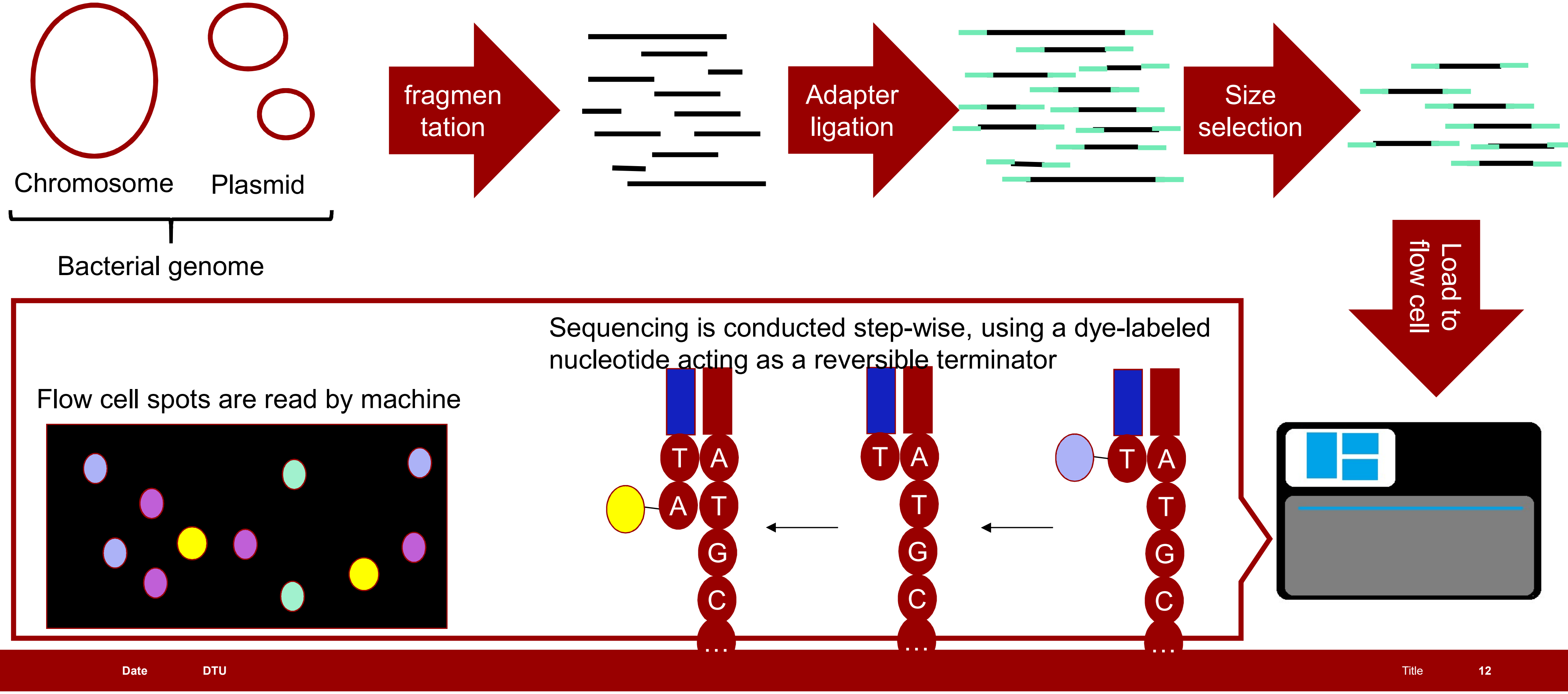
- After size selection you have a range of more similar fragment lengths

- Insert size is the distance between adapters

- A read pair is produced by reading the insert from opposite ends



Summarized library preparation (illumina paired end)



Chromosome

Plasmid

Bacterial genome

fragmen tation

Adapter ligation

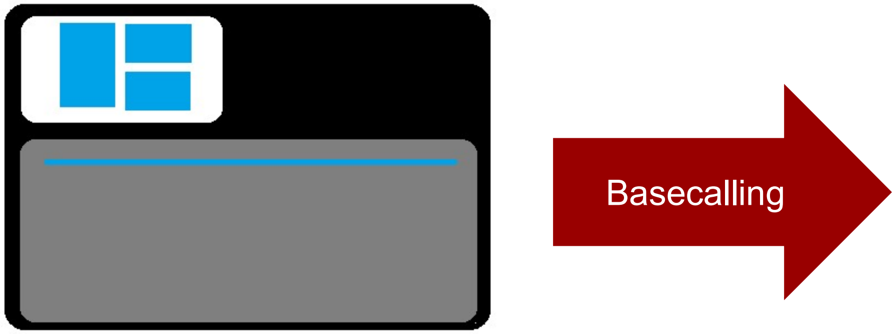
Size selection

Load to flow cell

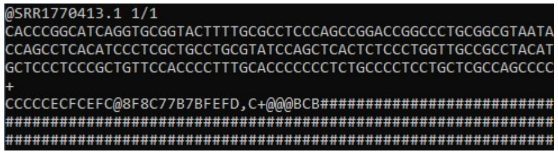
Sequencing is conducted step-wise, using a dye-labeled nucleotide acting as a reversible terminator

Flow cell spots are read by machine

Next generation sequencing data processing



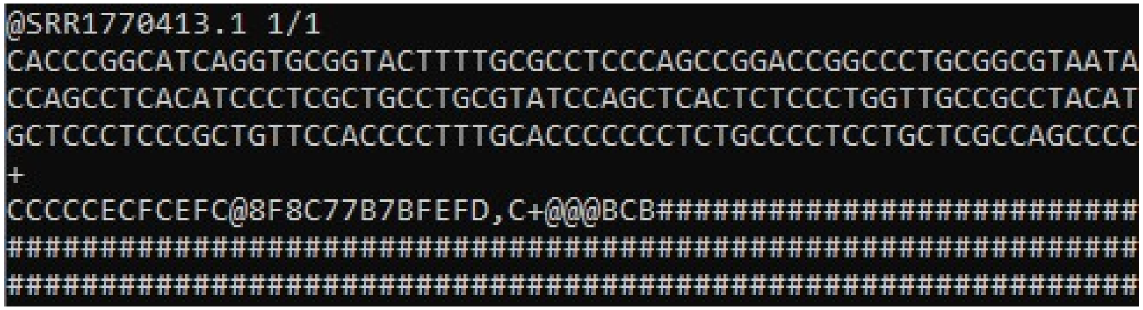
Fastq file containing millions of reads



```
@SRR1770413.1 1/1
CACCCGGCATCAGGTGCGGTACTTTTGCGCCTCCCAGCCGGACCGGCCCTGCGGCGTAATA
CCAGCCTCACATCCCTCGCTGCCTGCGTATCCAGCTCACTCTCCCTGGTTGCCGCCTACAT
GCTCCCTCCCGCTGTTCCACCCCTTTGCACCCCCCCTCTGCCCCTCCTGCTCGCCAGCCCC
+
CCCCCECFCEFC@8F8C77B7BFEFD,C+@@@BCB##########################
#############################################################
#############################################################
```

What is fastq?

- Fastq are the the read files produced by sequencing machines, after base-calling.
- It has a particular format:
  - Header
    - Contains info on the run, depends on machine
    - Unique ID
  - Called bases
    - Sequence
  - Spacer line
    - Spacing
  - Base quality scores
    - Phred-score giving the probability that the base call is incorrect.



```
@SRR1770413.1 1/1
CACCCGGCATCAGGTGCGGTACTTTTGCGCCTCCCAGCCGGACCGGCCCTGCGGCGTAATA
CCAGCCTCACATCCCTCGCTGCCTGCGTATCCAGCTCACTCTCCCTGGTTGCCGCCTACAT
GCTCCCTCCCGCTGTTCCACCCCTTTGCACCCCCCCTCTGCCCCTCCTGCTCGCCAGCCCC
+
CCCCCECFCEFC@8F8C77B7BFEFD,C+@@@BCB##########################
#############################################################
#############################################################
```

Phred scores

- The Phred quality score is a logarithmic score based on the probability that the base call (nucleotide) is incorrect

- Q10 = 1/10 risk of incorrect base
- Q20 = 1/100 risk of incorrect base
- Q30 = 1/1000 risk of incorrect base

- This means that in a sequence of 100 bp at Q20, there will most likely be at least 1 bp called incorrectly

$$Q = -10 \cdot \log_{10}(P)$$

or in terms of probability

$$P = 10^{-\frac{Q}{10}}$$

Where

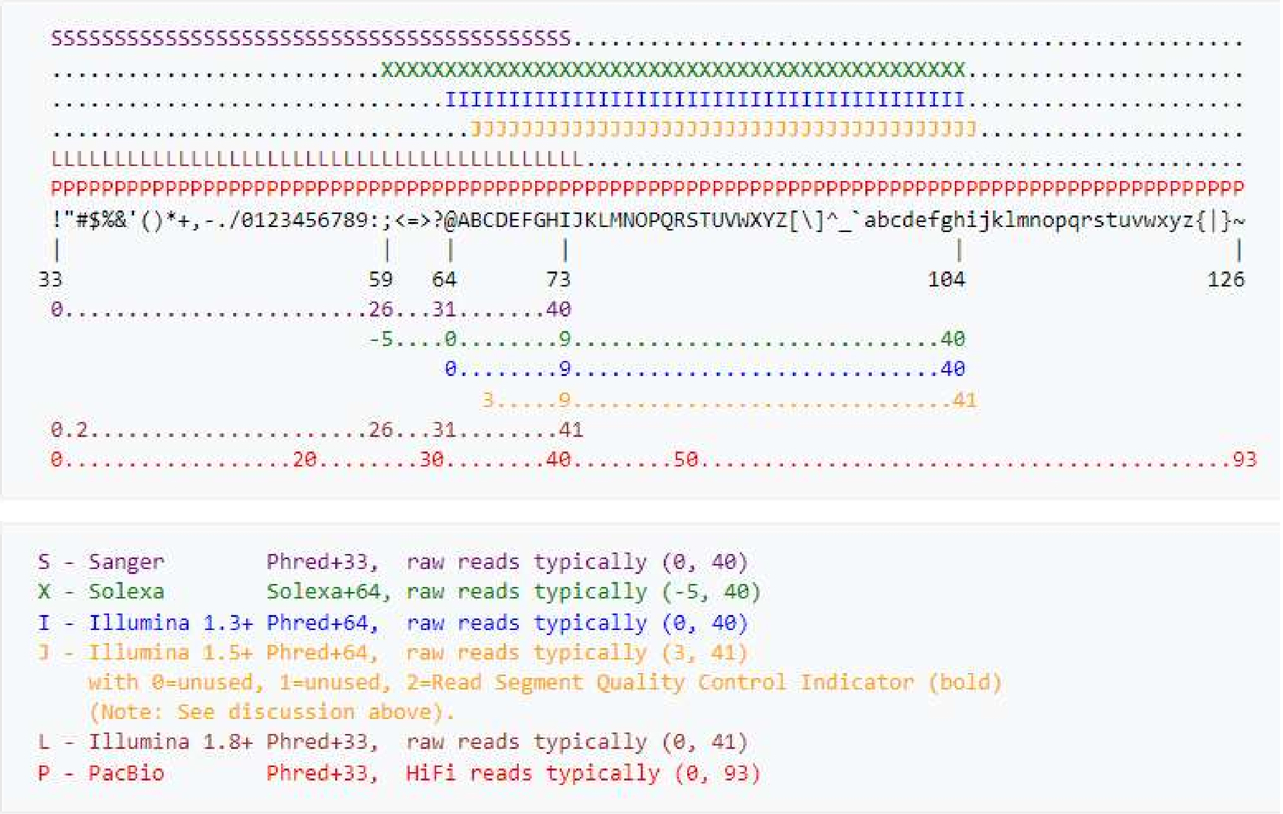
P = probability of incorrect base call

Q = Phred quality score

Phred quality score	Probability of incorrect base call	Probability of being correct
10	0.1	90%
20	0.01	99%
30	0.001	99.9%

Phred scores?

- The Phred quality score given as one of the 127 standard ASCII characters
- The scale is off-set, with different sequencing machines use different scales
- New Illumina machines use the sanger scale
- The base quality score is important in correctly calling Single Nucleotide Polymorphisms (SNP), used in phylogeny and outbreak detection



```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 ................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
 PPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPP
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
 0........................26...31.......40
                          -5....0........9.............................40
                                0........9.............................40
                                   3.....9..............................41
 0.2......................26...31.......41
 0..................20........30........40........50.........................................93
```

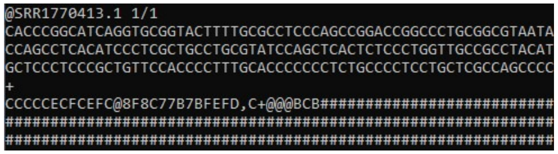
```
S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 41)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
P - PacBio        Phred+33,  HiFi reads typically (0, 93)
```

Phred scales used in different machines, from the FASTQ format entry on wikipedia: FASTQ format - Wikipedia

Next generation sequencing data processing

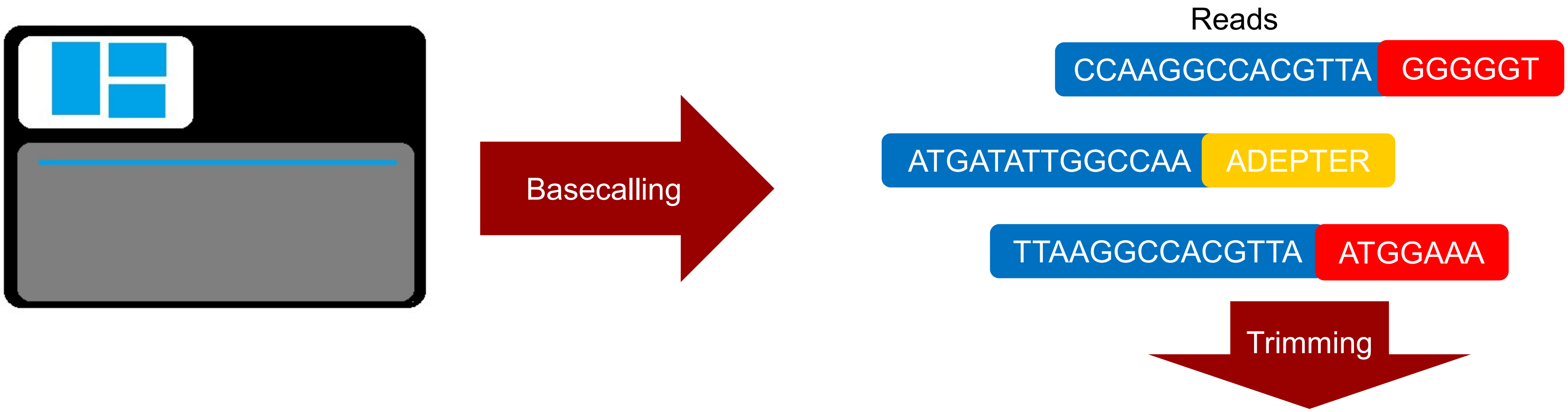
Basecalling

Fastq file containing millions of reads



```
@SRR1770413.1 1/1
CACCCGGCATCAGGTGCGGTACTTTTGCGCCTCCCAGCCGGACCGGCCCTGCGGCGTAATA
CCAGCCTCACATCCCTCGCTGCCTGCGTATCCAGCTCACTCTCCCTGGTTGCCGCCTACAT
GCTCCCTCCCGCTGTTCCACCCCTTTGCACCCCCCCTCTGCCCCTCCTGCTCGCCAGCCCC
+
CCCCCECFCEFC@8F8C77B7BFEFD,C+@@@BCB##########################
#############################################################
#############################################################
```

Next generation sequencing data processing



Basecalling

Reads

CCAAGGCCACGTTA GGGGGT

ATGATATTGGCCAA ADEPTER

TTAAGGCCACGTTA ATGGAAA

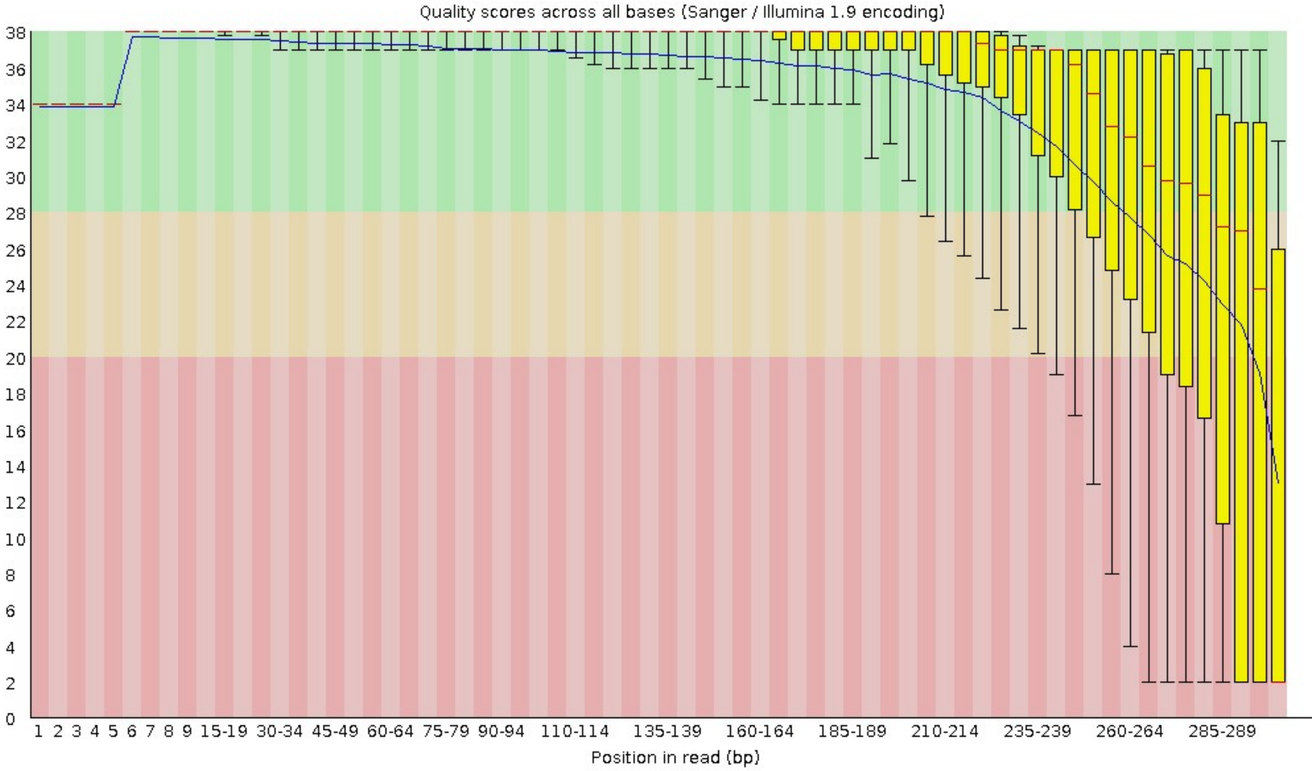
Trimming

What is Trimming

- Adapter sequences are not sequenced at the 5’ end of the read, however we can sequence through the entire fragment and start sequencing the adapter at the 3’ end

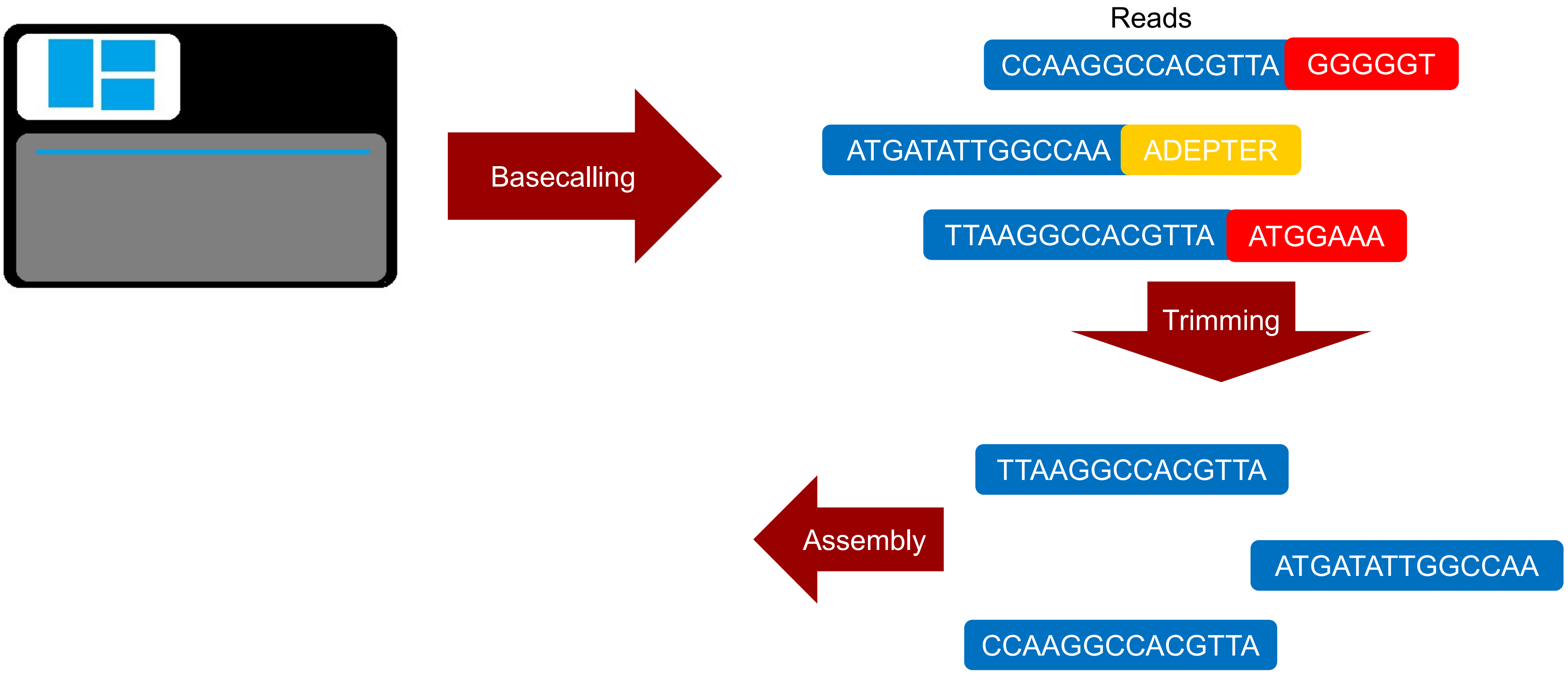
- During sequencing, enzymes start to degrade and errors are more common, for this reason we generally see a lower quality at the 3’ end

- Removing poor quality makes SNP and gene prediction more reliable



Section of output from running fastqc

Next generation sequencing data processing



Reads

CCAAGGCCACGTTA GGGGGT

ATGATATTGGCCAA ADEPTER

TTAAGGCCACGTTA ATGGAAA

Basecalling

Trimming

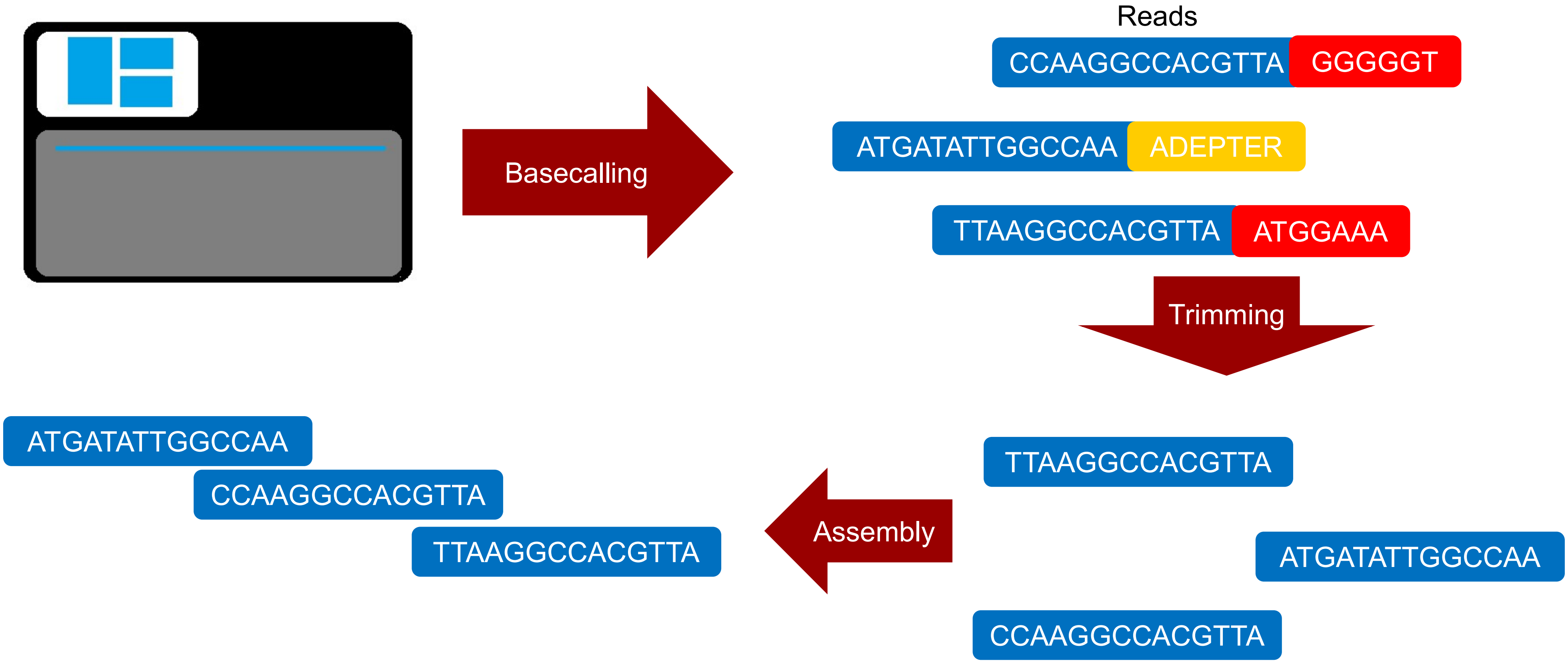
TTAAGGCCACGTTA

ATGATATTGGCCAA

CCAAGGCCACGTTA

Assembly

Next generation sequencing data processing



Basecalling

Reads

CCAAGGCCACGTTA GGGGGT

ATGATATTGGCCAA ADEPTER

TTAAGGCCACGTTA ATGGAAA

Trimming

TTAAGGCCACGTTA

ATGATATTGGCCAA

CCAAGGCCACGTTA

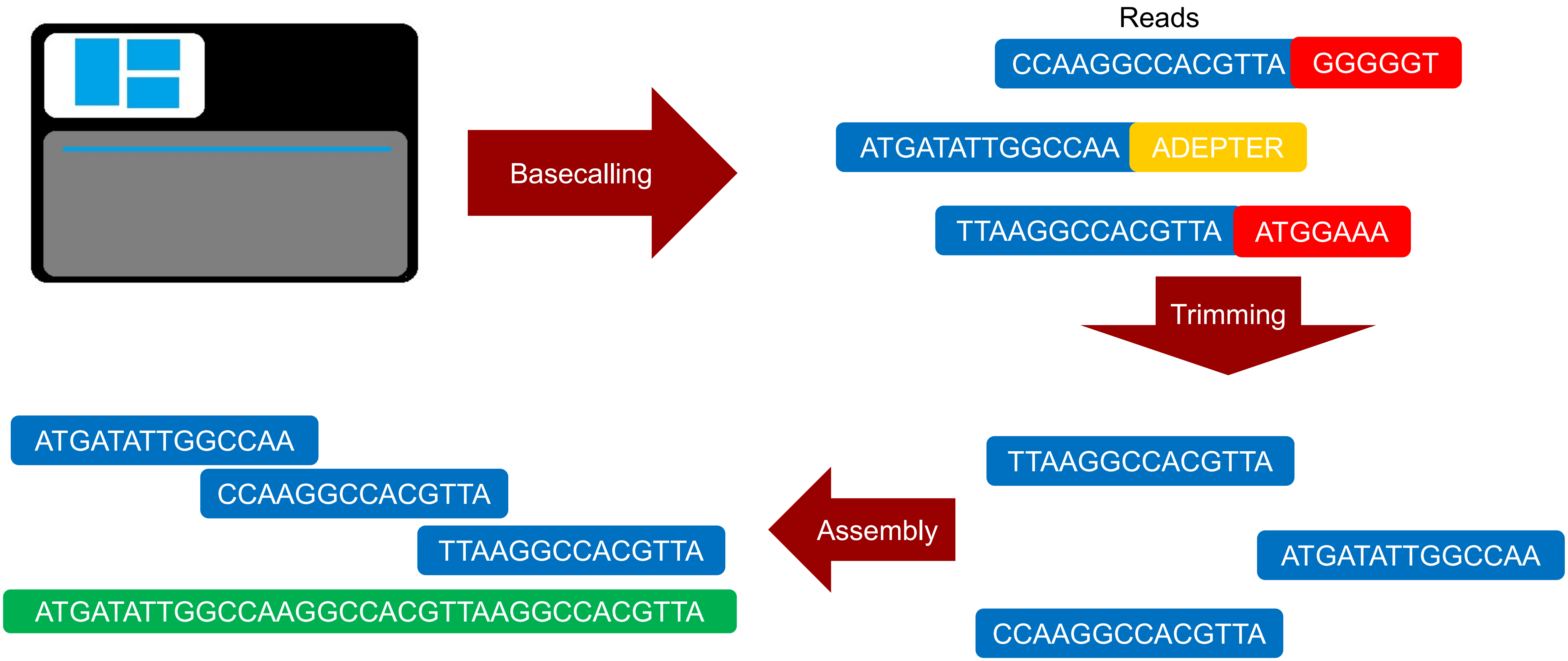
Assembly

ATGATATTGGCCAA

CCAAGGCCACGTTA

TTAAGGCCACGTTA

Next generation sequencing data processing



Reads

CCAAGGCCACGTTA GGGGGT

ATGATATTGGCCAA ADEPTER

TTAAGGCCACGTTA ATGGAAA

Basecalling

Trimming

TTAAGGCCACGTTA

ATGATATTGGCCAA

CCAAGGCCACGTTA

Assembly

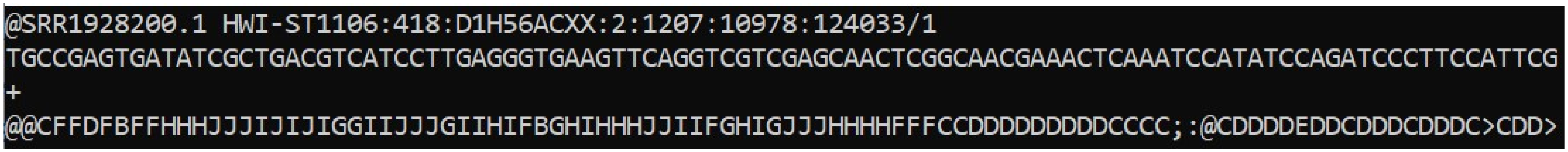
ATGATATTGGCCAA

CCAAGGCCACGTTA

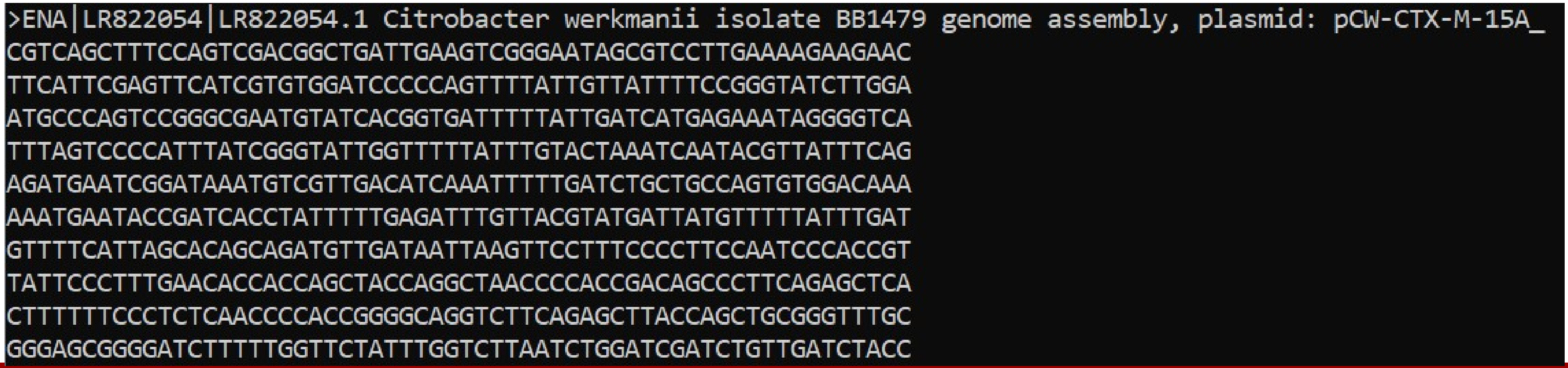
TTAAGGCCACGTTA

ATGATATTGGCCAAGGCCACGTTAAGGCCACGTTA

From fastq to fasta

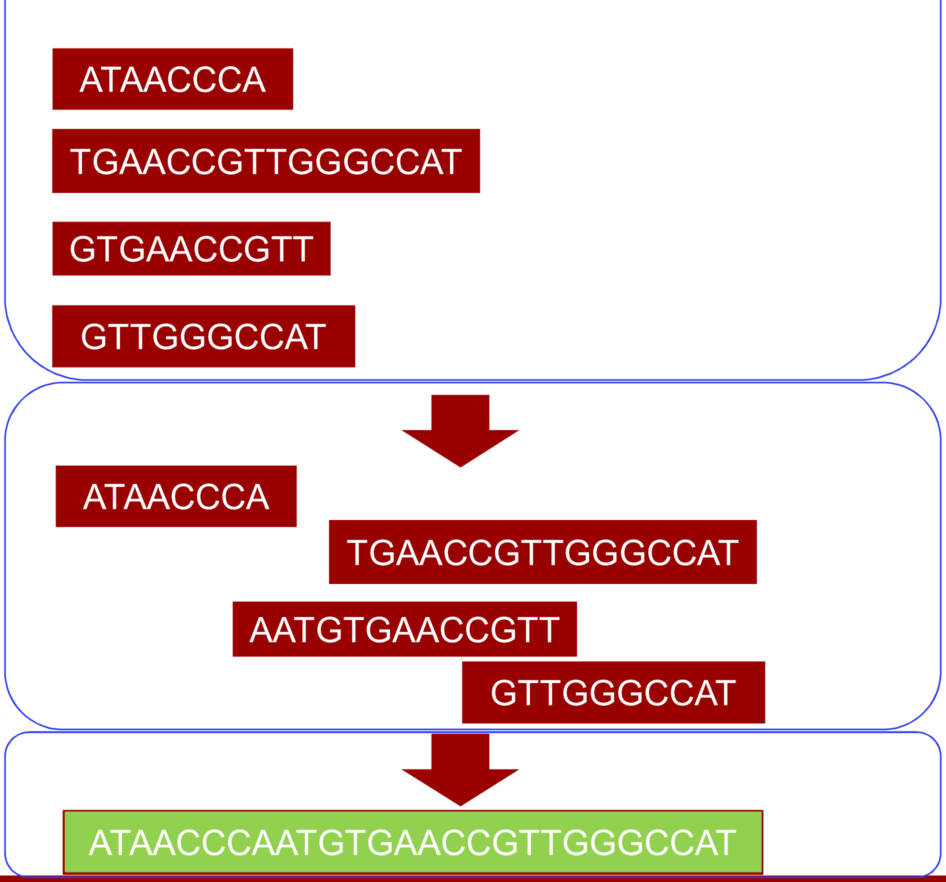


```
@SRR1928200.1 HWI-ST1106:418:D1H56ACXX:2:1207:10978:124033/1
TGCCGAGTGATATCGCTGACGTCATCCTTGAGGGTGAAGTTCAGGTCGTCGAGCAACTCGGCAACGAAACTCAAATCCATATCCAGATCCCTTCCATTCG
+
@@CFFDFBFFHHHJJJIJIJIGGIIJJJGIIHIFBGHIHHHJJIIFGHIGJJJHHHHFFFCCDDDDDDDDDCCCC;:@CDDDDEDDCDDDCDDDC>CDD>
```



```
>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
```

De novo assembly



ATAACCCA

TGAACCGTTGGGCCAT

GTGAACCGTT

GTTGGGCCAT

ATAACCCA

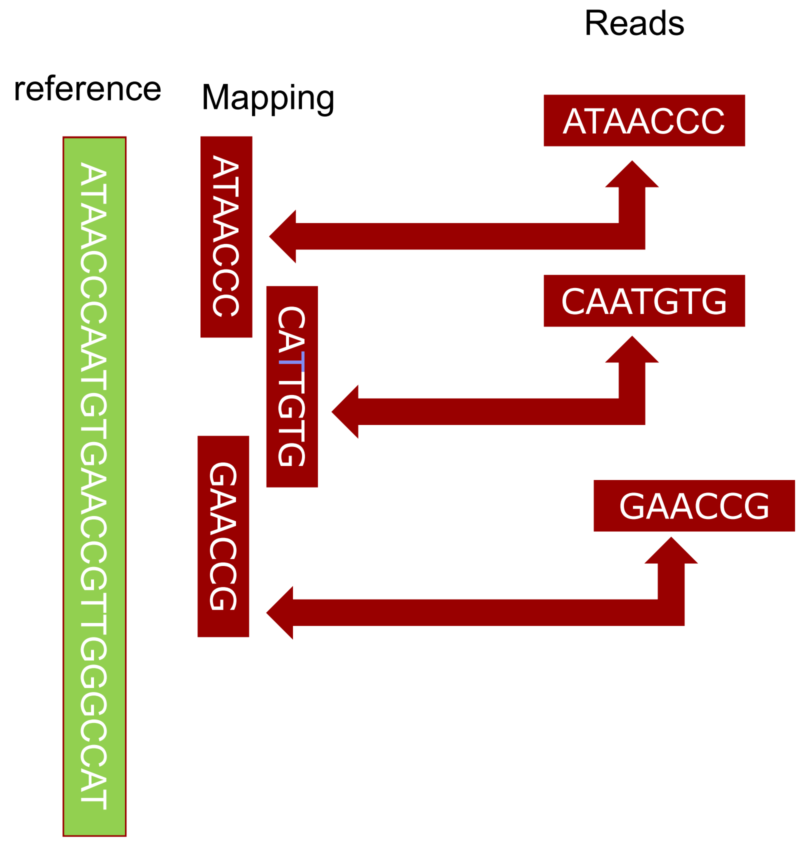
TGAACCGTTGGGCCAT

AATGTGAACCGTT

GTTGGGCCAT

ATAACCCAATGTGAACCGTTGGGCCAT

Mapping



Reads

reference

Mapping

ATAACCC

CAATGTG

GAACCG

ATAACCCAATGTGAACCGTTGGGCCAT

ATAACCC

CATTGTG

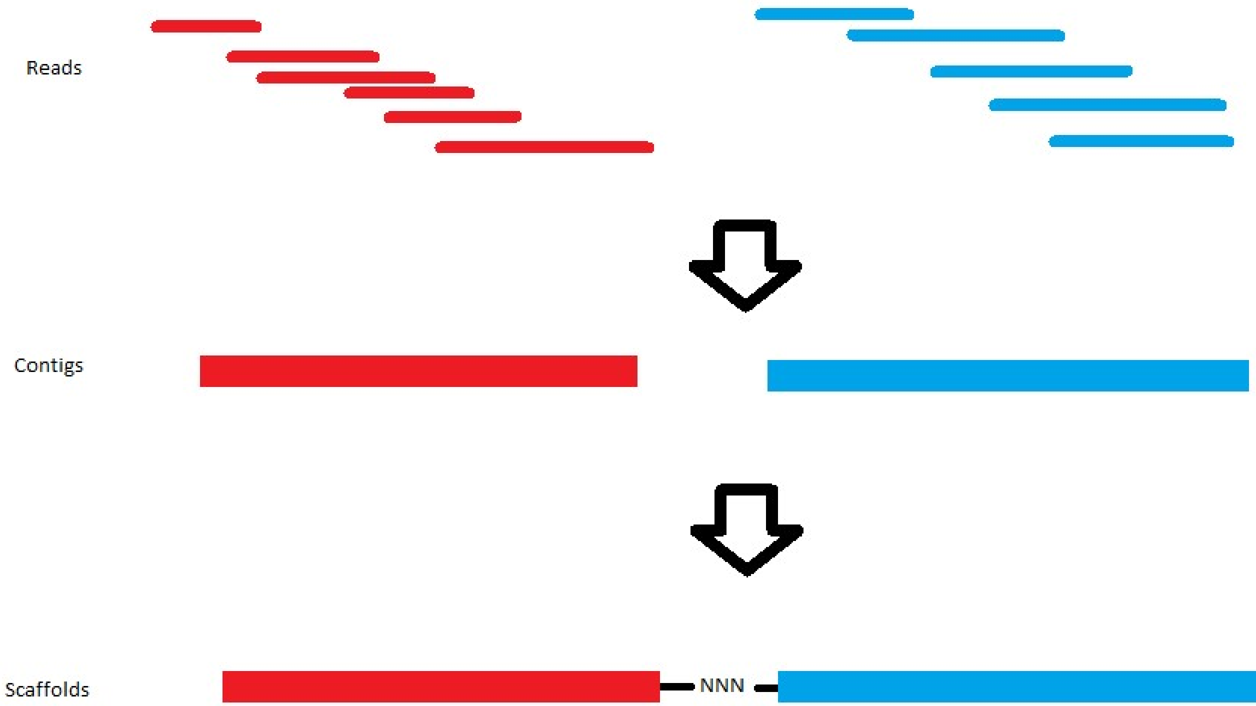
GAACCG

De novo assembly

- Many programs can do assembly, they differentiate by how precise they can construct the assembly, how fast and how computationally heavy their workload
  - spades
  - SOAPdenovo2
  - MEGAHIT
  - Velvet

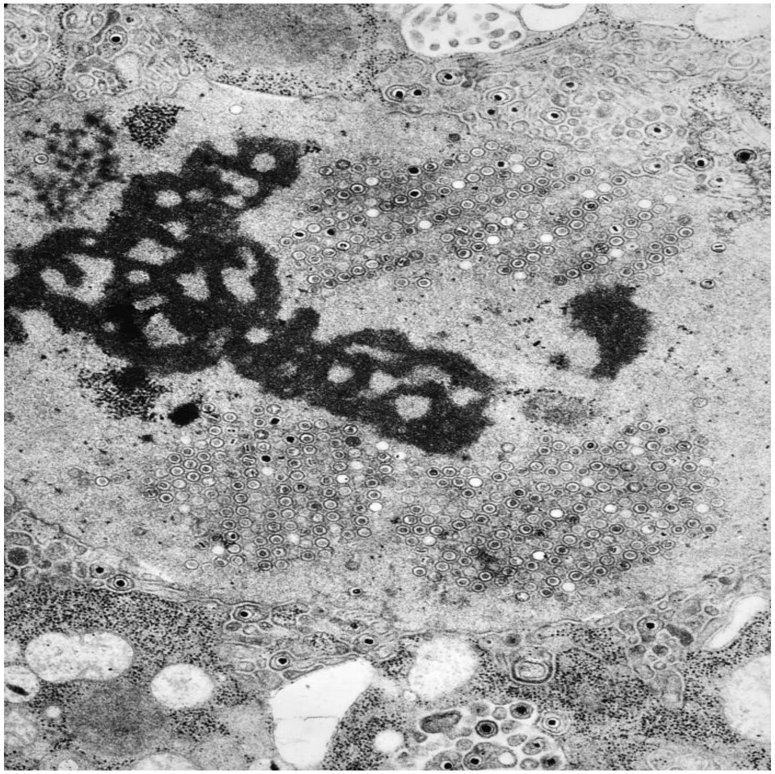
- Reads are assembled into contigs by constructing de bruijn graphs of reads (see Compeau, et al. 2011 for further information on these strategies)

- The assembly should not contain unknown bases (N)



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
- E.g. Salmonella enterica is expected to be 4.4-5.0 Mb, if assembly contains 8Mb, it is like due to contamination



Source: CDC/ Dr. Fred Murphy; Sylvia Whitfield

Assembly statistics – N50

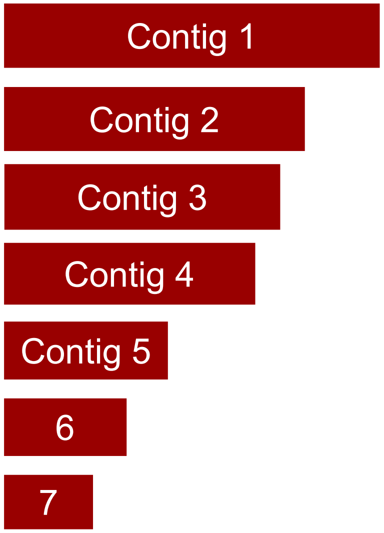
- N50 is found by:
  - Sorting all contigs in assembly from longest to shortest, starting with the longest
  - Adding together the length of the longest contigs until half the assembly is included
  - The length of the last added contig to reach 50% of the assembly is the N50

- 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 1

Contig 2

Contig 3

Contig 4

Contig 5

6

7

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 – 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 thus the sequencing

Ref: 5.000.000bp

N50 is calculated from 5.000.000/2 = 2.500.000

	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	

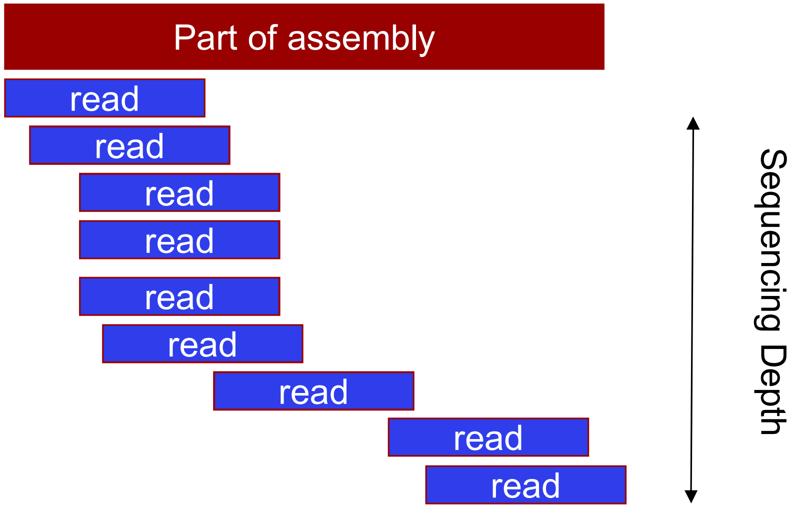
Assembly statistics – Depth (Sequence coverage)

- The number times we cover a 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:



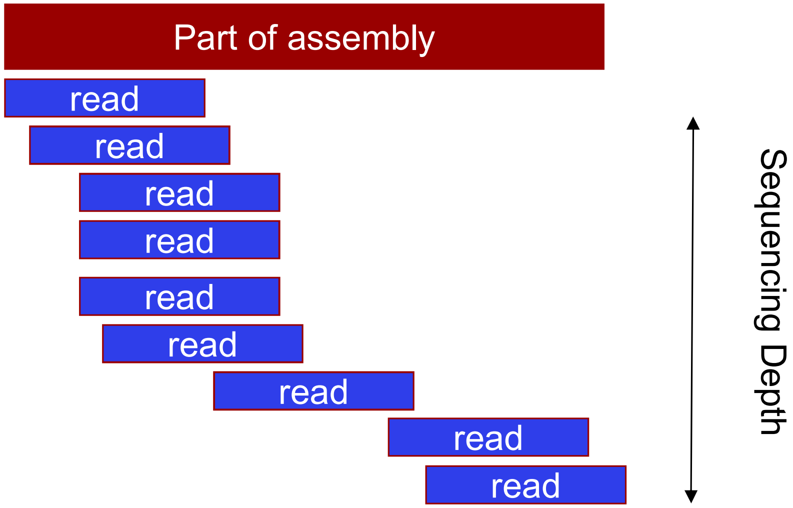
$$sequence\ coverage = \frac{number\ of\ reads\ * average\ read\ length}{Total\ genome\ size}$$

Assembly statistics – Depth (Sequence coverage)

- Let us assume the example on the right is 800bp and the reads are 100bp on average

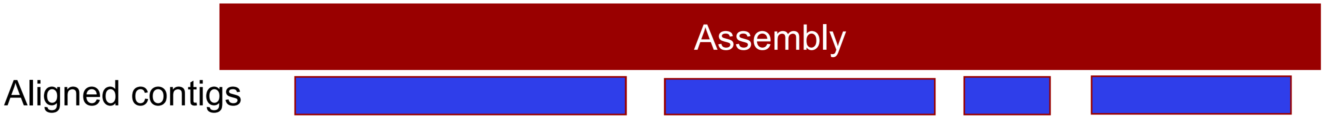
$$sequence\ coverage = \frac{number\ of\ reads\ * average\ read\ length}{Total\ genome\ size}$$

$$sequence\ coverage = \frac{9\ * 100bp}{800bp} = 1.125x$$



Assembly statistics – Physical coverage

- If a closed reference genome is available the physical coverage can likewise be calculated
- The physical coverage is the percentage of the assembly covered by reads
- The percentage should be as high as possible
- Species with low GC content have been known to demand higher sequencing depth to achieve better coverage



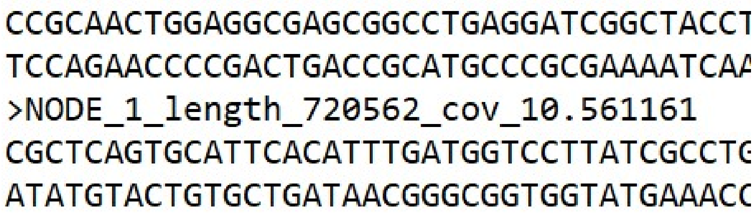
Assembly statistics – number of contigs

- 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



```
CCGCAACTGGAGGCGAGCGGCCTGAGGATCGGCTACCT
TCCAGAACCCCGACTGACCGCATGCCCGCGAAAATCAA
>NODE_1_length_720562_cov_10.561161
CGCTCAGTGCATTCACATTTGATGGTCCTTATCGCCTG
ATATGTACTGTGCTGATAACGGGCGGTGGTATGAAACC
```

Most commonly used QC metrics

- There is no universal threshold for the quality metrics described below, and they can be expected to vary depending on the specific species and strain

- N50
  - Minimal value of 30 000bp suggested
  - Above 100 000bp often obtainable

- Number of contigs
  - Less than 500 suggested
  - Anything above 250 should be evaluated

- Total bp
  - Within accepted range for identified species

What is a kmer?

- A kmer is a substring within a stretch of DNA of length “k”

- When dividing a DNA sequence into kmers, you start with the first k basepairs and then procede by moving one nucleotide at a time

- E.g. let us look at the sequence to the right and divide it into kmers of length 4 (into 4mers)

ATGCATATTG

What is a kmer?

- A kmer is a substring within a stretch of DNA of length “k”

- When dividing a DNA sequence into kmers, you start with the first k basepairs and then procede by moving one nucleotide at a time

- E.g. let us look at the sequence to the right and divide it into kmers of length 4 (into 4mers)

- The first 4mer consist of the first 4 bases

ATGCATATTG

ATGC

What is a kmer?

- A kmer is a substring within a stretch of DNA of length "k"

- When dividing a DNA sequence into kmers, you start with the first k basepairs and then procede by moving one nucleotide at a time

- E.g. let us look at the sequence to the right and divide it into kmers of length 4 (into 4mers)

- The first 4mer consist of the first 4 bases
- We then move one space to the right to identify the next 4mer

```
ATGCATATTG
ATGC
 TGCA
```

What is a kmer?

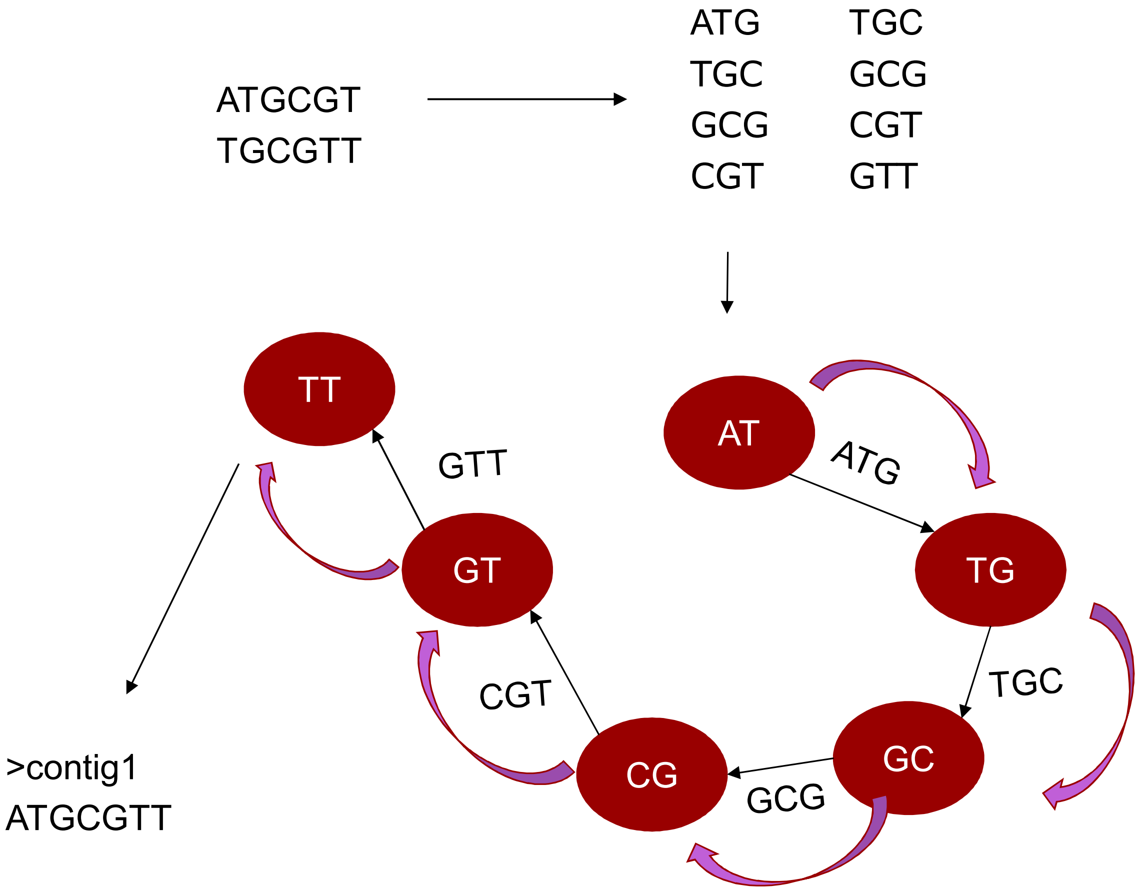
- A kmer is a substring within a stretch of DNA of length “k”
- When dividing a DNA sequence into kmers, you start with the first k basepairs and then procede by moving one nucleotide at a time
- E.g. let us look at the sequence to the right and divide it into kmers of length 4 (into 4mers)
- The first 4mer consist of the first 4 bases
- We then move one space to the right to identify the next 4mer
- We end up with 7 unique 4mers

```
ATGCATATTG
ATGC
 TGCA
  GCAT
   CATA
    ATAT
     TATT
      ATTG
```

But why?

- Kmers are used in multiple settings to make dealing with sequence data more manageable
- In search function like blast
- In assembly (de brujn graphs)
- Basically to do alignment more achievable
- Kmerfinder uses 16mers to align submitted sequences against a database constructed from the overlapping 16kmers starting with ATGAC

De bruijn graph approach to assemble 2 sequences of 6 basepairs



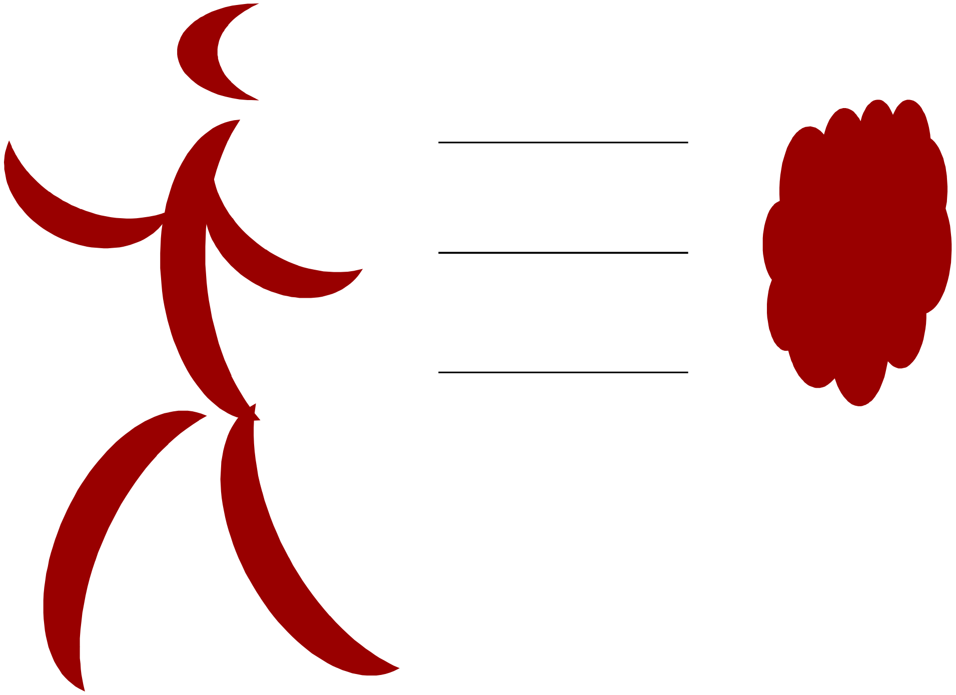
ATGCGT
TGCGTT

ATG TGC GCG CGT

TGC GCG CGT GTT

>contig1
ATGCGTT

Start of exercises!

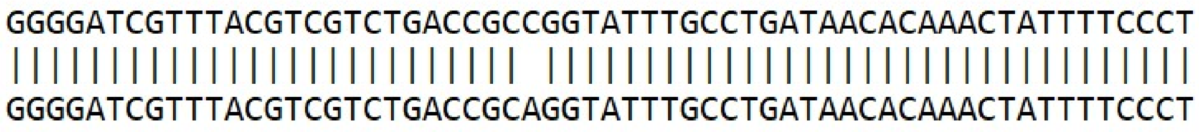


Sequence identity

- Another term we encounter in the cge tools is % identity (ID)

- The identity describes how many bases of the aligned sequences are identical

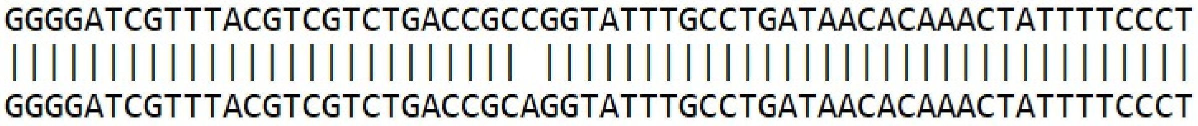
- Given the alignment:



```
GGGGATCGTTTACGTCGTCTGACCGCCGGTATTTGCCTGATAACACAAACTATTTTCCCT
|||||||||||||||||||||||||| |||||||||||||||||||||||||||||||||
GGGGATCGTTTACGTCGTCTGACCGCAGGTATTTGCCTGATAACACAAACTATTTTCCCT
```

Sequence identity

- Another term we encounter in the cge tools is % identity (ID)
- The identity describes how many bases of the aligned sequences are identical
- Given the alignment:



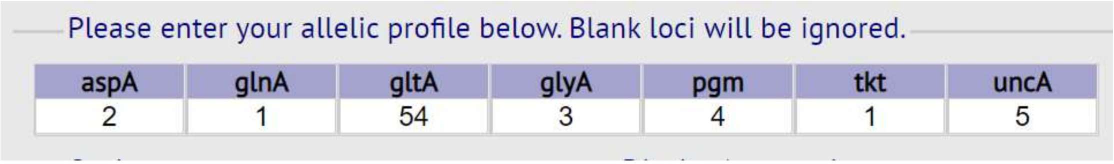
```
GGGGATCGTTTACGTCGTCTGACCGCCGGTATTTGCCTGATAACACAAACTATTTTCCCT
|||||||||||||||||||||||||| |||||||||||||||||||||||||||||||||
GGGGATCGTTTACGTCGTCTGACCGCAGGTATTTGCCTGATAACACAAACTATTTTCCCT
```

- Sequence length 60
- Matches 59
- %ID = 59/60*100% = 98.3%

MLST

- MultiLocus Sequence Typing (MLST), is a scheme of 7 genes specific for a species
- The Unique Allele (DNA sequence) for each of these 7 genes are given a number
- Any time a new allele is discovered, its sequence is given a new number and added to the database
- Each unique combination of alleles are given a number, this is the sequence type

Allele profile for sequence type (ST) 1 in campylobacter jejuni/coli, source: Pubmlst Search by locus combinations (pubmlst.org)



Please enter your allelic profile below. Blank loci will be ignored.

aspA	glnA	gltA	glyA	pgm	tkt	uncA
2	1	54	3	4	1	5

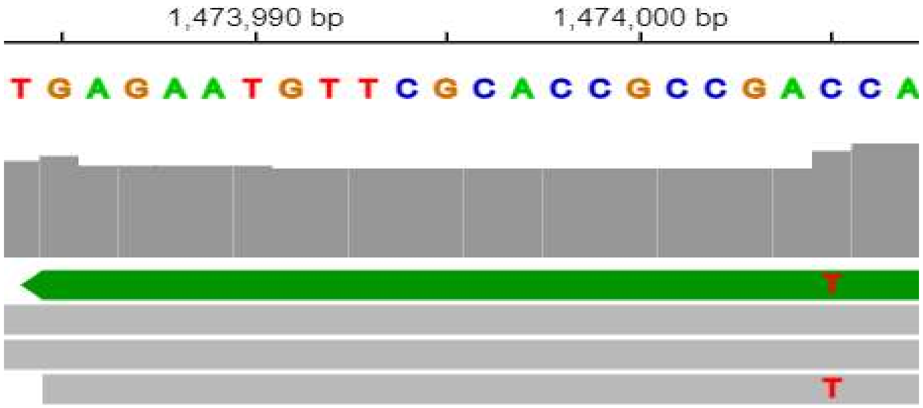
Single nucleotide polymorphism (SNP)

- A SNP is a mutation within a subpopulations of individuals, essentially it is a point mutation which distinguishes two “closely” related strains of the same species

- To separate sequencing error from true SNPs, we need to have:
  - Proper sequencing depth at the position
  - High Q-score

- When we know the amounts of SNP differences we can infer the phylogenic relationship between strains

- High resolution



Section of reads mapped to reference, visualized using integrative genomics viewer, IGV: Integrative Genomics Viewer