Module 3
Mapping short reads

Working with pathogen genomes

10th - 14th May 2021

Christine Boinett

Objectives

- Introduce data files required for mapping
- Visualize mapped data in Artemis genome viewer
- Show sequence variation e.g SNPs, INDELS

Popular short read sequencers

	 MiniSeq System	 MiSeq Series	 NextSeq Series	 HiSeq Series	 HiSeq X Series*
Key Methods	Amplicon, targeted RNA, small RNA, and targeted gene panel sequencing.	Small genome, amplicon, and targeted gene panel sequencing.	Everyday exome, transcriptome, and targeted resequencing.	Production-scale genome, exome, transcriptome sequencing, and more.	Population- and production-scale whole-genome sequencing.
Maximum Output	7.5 Gb	15 Gb	120 Gb	1500 Gb	1800 Gb
Maximum Reads per Run	25 million	25 million†	400 million	5 billion	6 billion
Maximum Read Length	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp
Run Time	4–24 hours	4–55 hours	12–30 hours	<1–3.5 days (HiSeq 3000/HiSeq 4000) 7 hours–6 days (HiSeq 2500)	<3 days

Illumina new generation short read sequencers





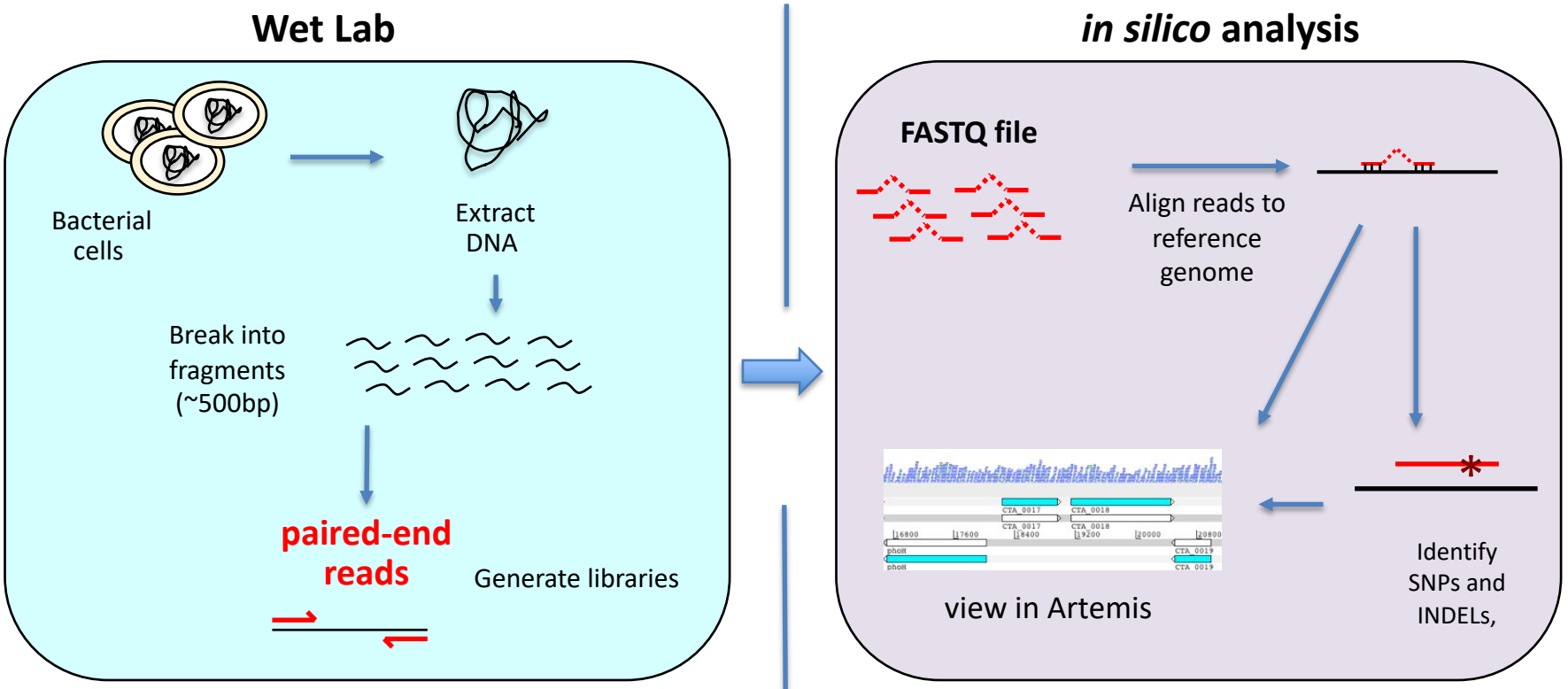
	iSeq 100	MiniSeq	MiSeq Series	NextSeq 550 Series	NextSeq 1000 & 2000
Popular Applications & Methods	Key Application	Key Application	Key Application	Key Application	Key Application
Large Whole-Genome Sequencing (human, plant, animal)					
Small Whole-Genome Sequencing (microbe, virus)	●	●	●	●	●
Exome & Large Panel Sequencing (enrichment-based)				●	●
Run Time	9.5–19 hrs	4–24 hours	4–55 hours	12–30 hours	11-48 hours
Maximum Output	1.2 Gb	7.5 Gb	15 Gb	120 Gb	330 Gb*
Maximum Reads Per Run	4 million	25 million	25 million †	400 million	1.1 billion*
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp

https://emea.illumina.com/systems/sequencing-platforms.html

What is mapping?

Mapping allows the locations of genes and other genetic features to be identified based on a reference genome whose sequence is known and annotated.

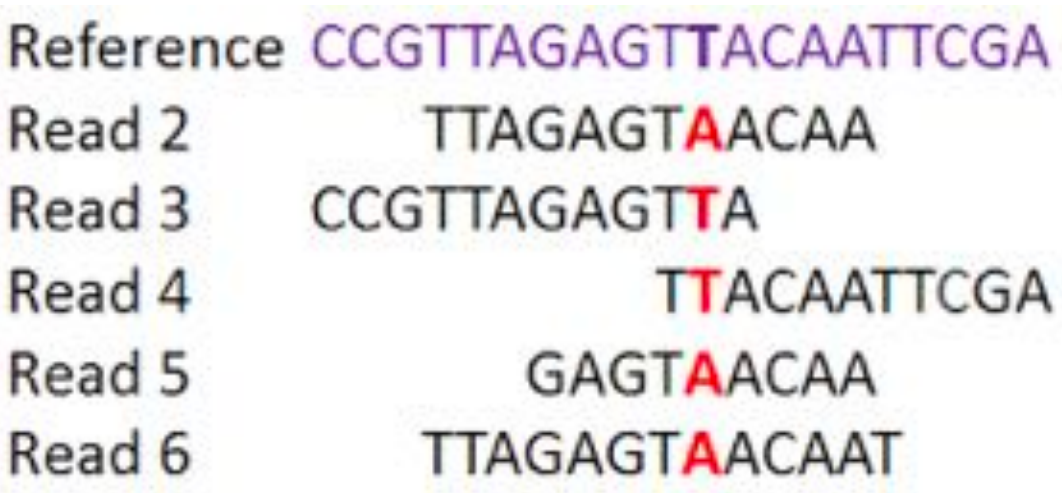
Workflow: generating sequencing reads and *in silico* analysis



DNA extraction and DNA library generation

Mapping of reads to a known reference to determine genetic differences

Single Nucleotide Polymorphisms (SNPs)



https://aschuerch.github.io/MolecularEpidemiology_AnalysisWGS/09-SNPphylo/index.html

INDELS

https://www.nature.com/articles/s41598-018-23978-z

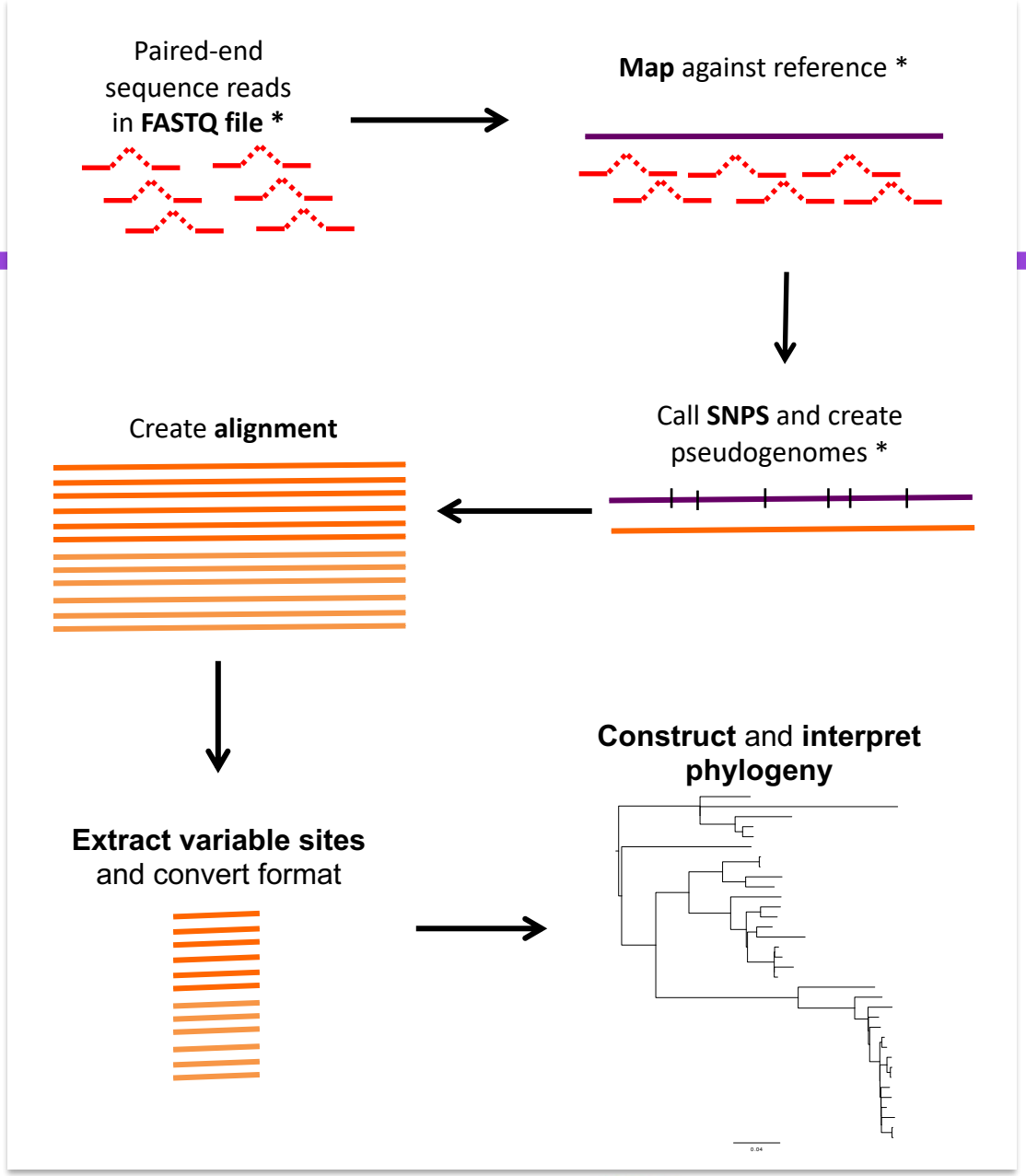
Why do we choose the mapping approach?

- We can capture information given the availability of a suitable reference genome:
  - Single Nucleotide Polymorphisms (SNPs),
  - insertions and deletions (indels)
  - Copy Number Variants (CNVs) between variants of the same bacteria.
  - Presence / absence of genes (AMR)

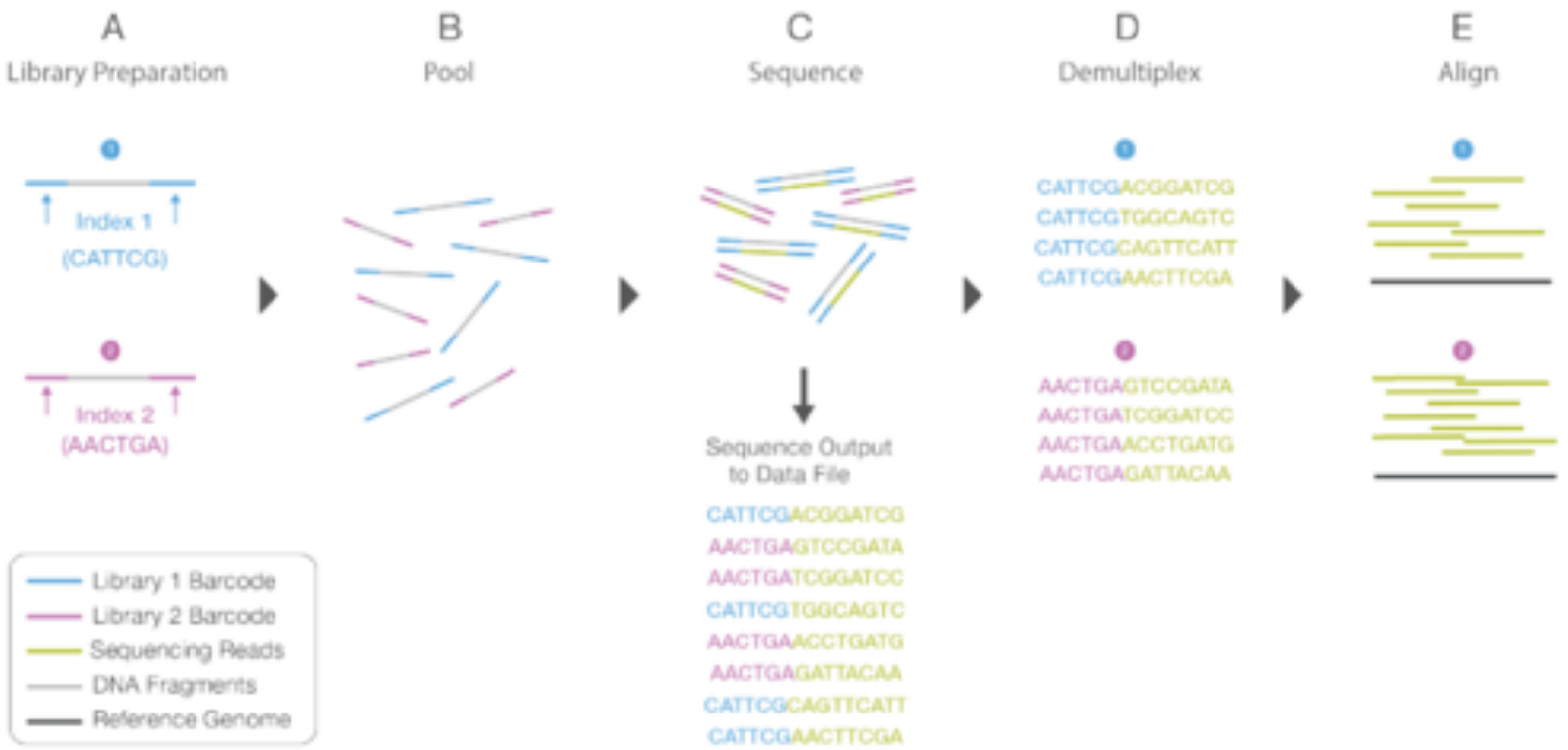
- As sequences diverge from the reference, mapping becomes progressively less effective

Workflow:

FastQ files (raw data) to phylogenetic trees to infer genetic relationships



Illumina sequencing reads - fastq



https://emea.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

Sequence output to Demultiplex

FASTQ file

```
@IL22_1561:1:1:1771:914
TTTTTTCAATACAATTTCATCTCTAATTTCAATTCAGCCTAA
+
>>>>7>>>>>>>>>>>>>>>>>>:>5>>>96>+>>7;>>>>
@IL22_1561:1:1:1222:1105
ATGATATCATCTACACGGTTTAAAAATTCTGGACGGCTGATC
+
>>;>>>>>>>;>>>>>55>>>>>>>>>>>>+2>>8>>>>>57
@IL22_1561:1:1:1438:709
ACACCGATGACAATAATTGTTCCAATATCTGTAACAAAGCTA
+
<<<<<2<<2<<<<<<<<<)'<<<<<<<<<<+<;;7<)<<2<<
@IL22_1561:1:1:1671:1462
TGCAAGAACATTAGACAACGTATCTTCAATCGTTTATACAAG
+
>>7>:>:>>7>>7/:>>>7/.7>>>>>57>>0>>>>>>7>>>
@IL22_1561:1:1:1168:891
AATGGAAATCAAACTATAACTTCAACACTAAATGAGAAGCTA
+
>>>:7>>>>>>>>>>>>>>>>>>>>>>>>>>0>>>>>>>>
@IL22_1561:1:1:1097:1281
TAAAAGTATAACTTTCCATCACCAAAGTCAGAATTTACATCG
+
>>>>>>>>>>>>>>>>>>>>>>>8>8>>>>>>>>>>>>>>>
```

TAG	Strain
CGTGAT	HGSA942
ACATCG	3HK
GCCTAA	CHI59
TGGTCA	CHI61
CACTGT	BK2491
ATTGGC	TUR9
GATCTG	HU25
TCAAGT	HU109
CTGATC	HSJ216
AAGCTA	FFP103
GTAGCC	ICP5062
TACAAG	GRE4

wellcome connecting science

Fastq format

```
1 @SEQ_ID
2 GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
3 +
4 !''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description (like a FASTA title line).

Line 2 is the raw sequence letters.

Line 3 begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again.

Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

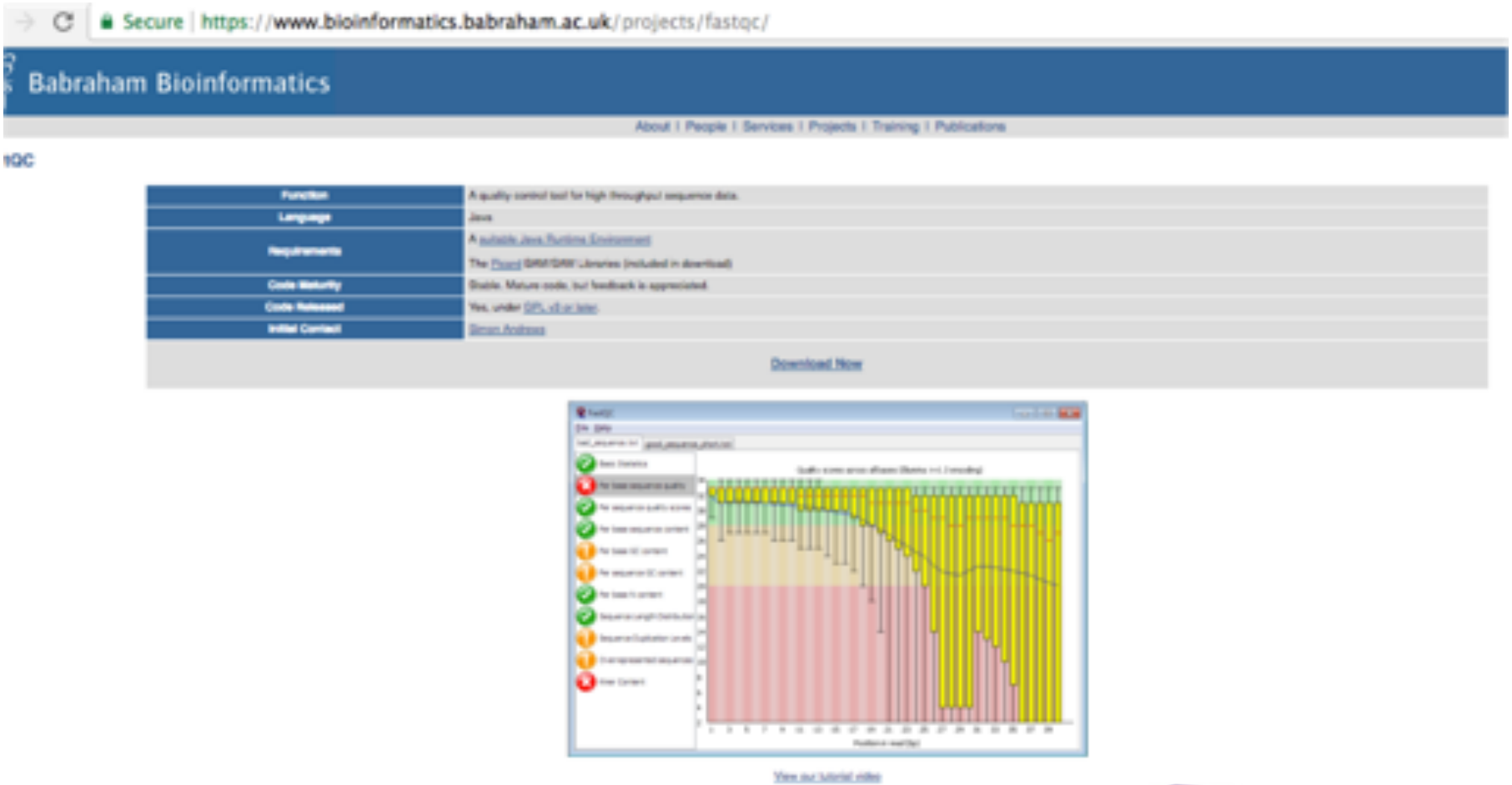
Fastq quality score/Phred score

$$Q = -10 \log_{10} P \quad \Longrightarrow \quad P = 10^{\frac{-Q}{10}}$$

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

The quality (Q), also called phred score, is the probability (P) that the corresponding basecall is incorrect.

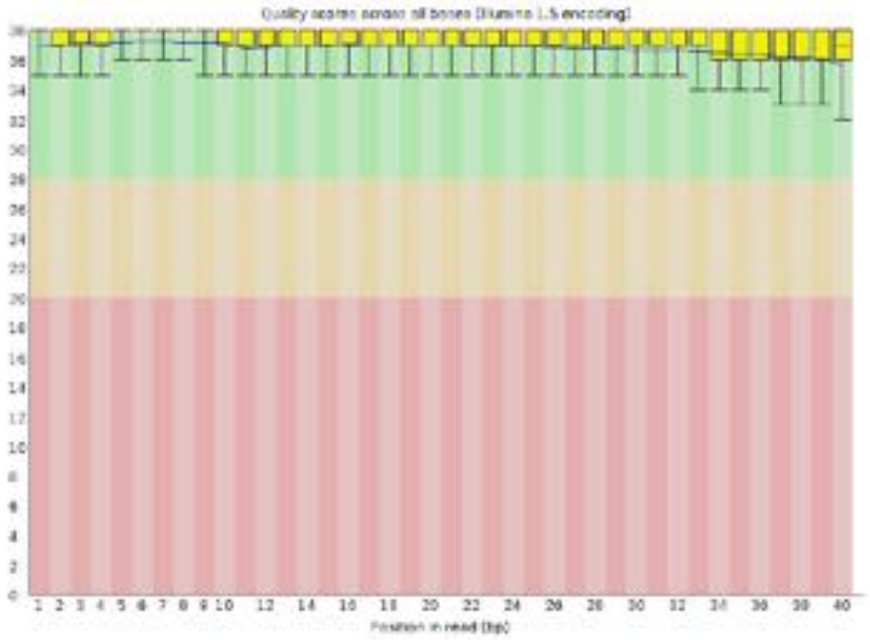
Fastq Quality Check made easy!



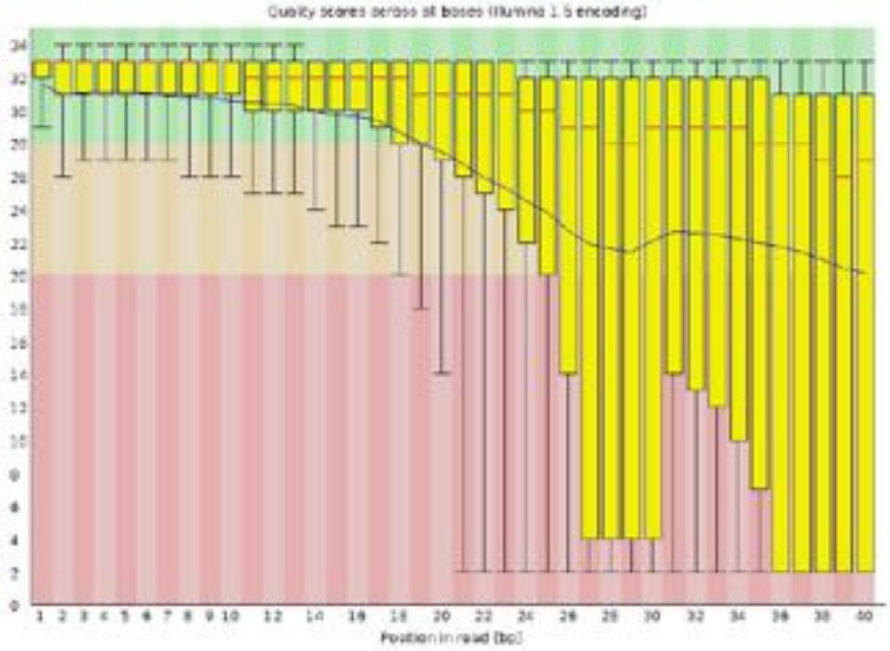
wellcome connecting science

Good vs Bad Fastq files

Good



Bad



Why do we choose the mapping approach?

- We can capture information given the availability of a suitable reference genome:
  - Single Nucleotide Polymorphisms (SNPs),
  - insertions and deletions (indels)
  - Copy Number Variants (CNVs) between variants of the same bacteria.
  - Presence / absence of genes (AMR)

- As sequences diverge from the reference, mapping becomes progressively less effective

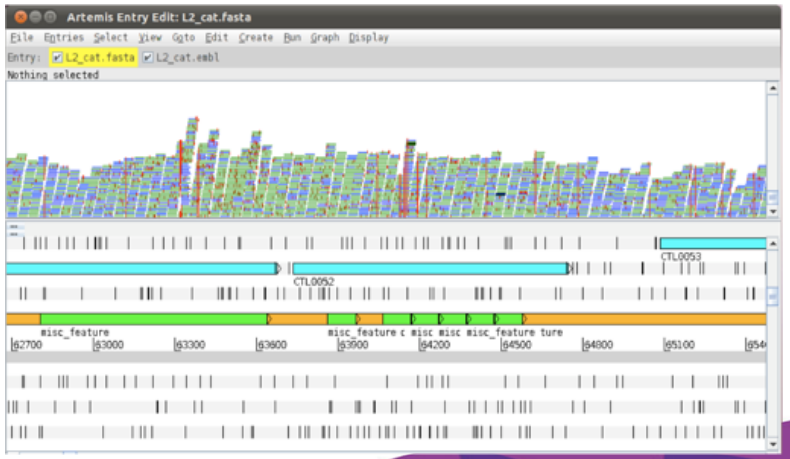
Mapping Illumina sequence data

Isolate - Fastq files

@IL24_5151:3:1:1553:916#9/1
NAAACTACTTACACCCACTCAGGACACCAGGGACATCATT
GCTGACGCCACGGCCTCACAGTGCTGAGCTGATGAT
+
$705291596=>>>>>>=>=>=>==>>>==>535:6=>=>>>>=>==
5:;318656:===991/1,-0,0015204.1
@IL24_5151:3:1:2173:904#9/1
NTTTTAACCGTACTTTCACCAGGATTATCGCAGGCGGATTC
CTGGTGATTAATTTCAAAAAATAGCGTTTAATCCA
+
$948883999>==>>>>>=>>=>>9>>>>=>===>=>==>>:;:::==
=>=55:88>==>9:0;;:===>>==>>>5
@IL24_5151:3:1:2948:912#9/1
NCCACCAGACACTGTCCGCAACCCCGGTAAGGGGCCAAC
GTTAGAACATCAAACATTAAAGGGTGGTATTTCAAGG
+

Reference – in fasta format

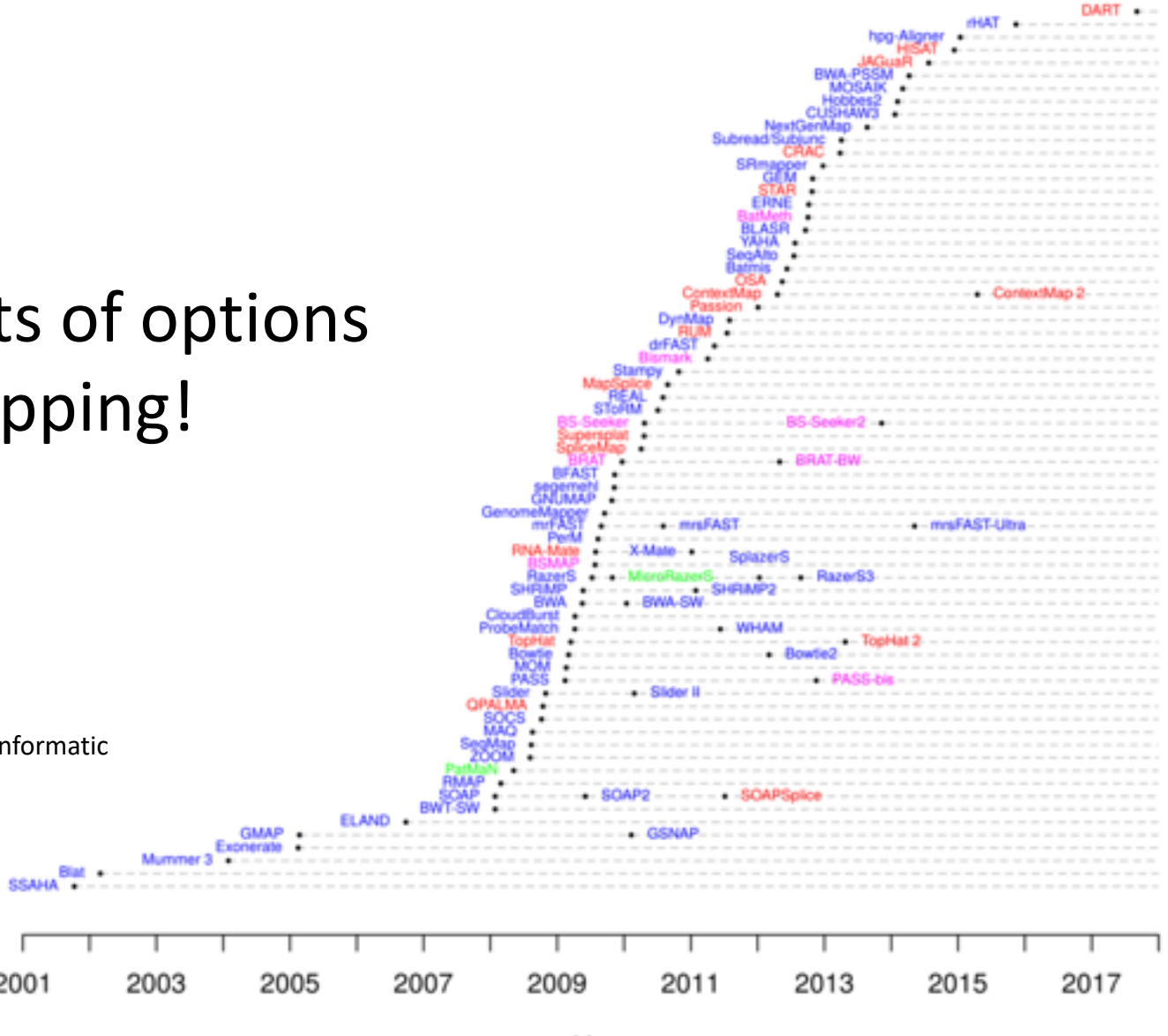
>reference sequence
ATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAGCGGCAGCGGGAAGTAGTTT
TACTTTGCCGGCGAGCGGCGGACGGGTGAGTAATGTCTGGGAAACTGCCTGATGGAGG
GATAACTACTGGAAACGGTAGCTAATACCGCATGACCTCGTAAGAGCAAAGTGGGGGAC
TTCGGGCCTCACGCCATCGGATGTGCCCAGATGGGATTAGCTAGTAGGTGGGGTAATGG
TCACCTAGGCGACGATCCCTAGCTGGTCTGAGAGGATGACCAGCCACACTGGAACTGAG
CACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCAAGC
GATGCAGCCATGCCGCGTGTGTGAAGAAGGCCTTCGGGTTGTAAAGCACTTTCAGCGAG
AGGAAGGCAGTCGTGTTAATAGCACGATTGATTGACGTTACTCGCAGAAGAAGCACCGGC



wellcome connecting science

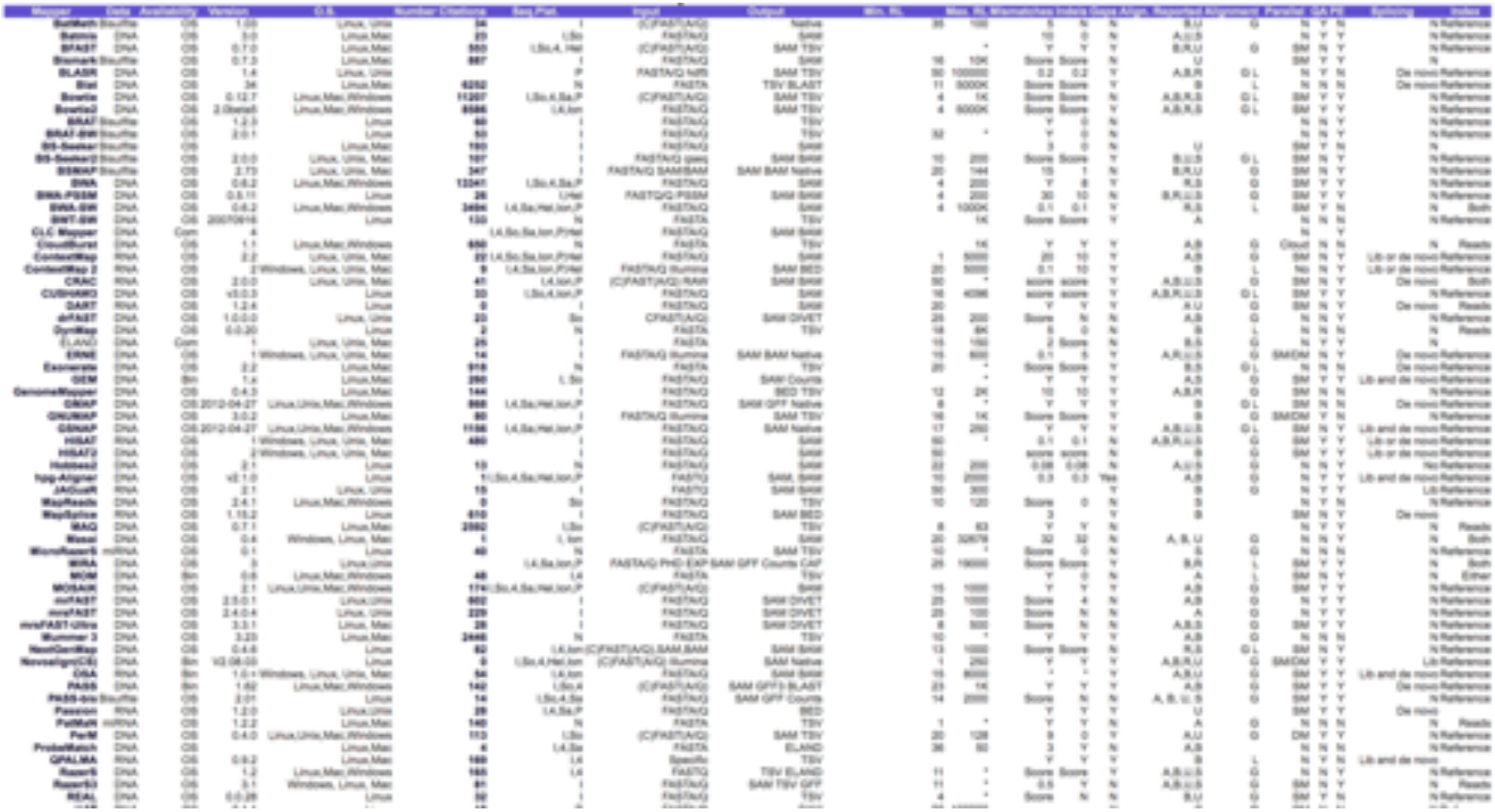
There are lots of options for mapping!

https://academic.oup.com/bioinformatics/article/28/24/3169/245777



wellcome connecting science

Comparison of different mappers



https://academic.oup.com/bioinformatics/article/28/24/3169/245777

Good general aligners

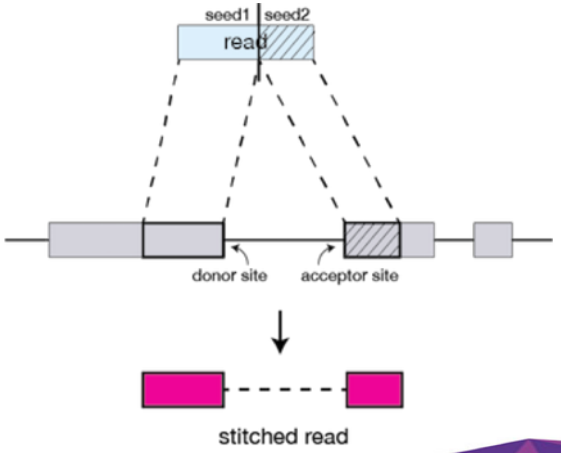
★ bwa
bowtie2
minimap2

→ Fast, sensitive and easy to use!

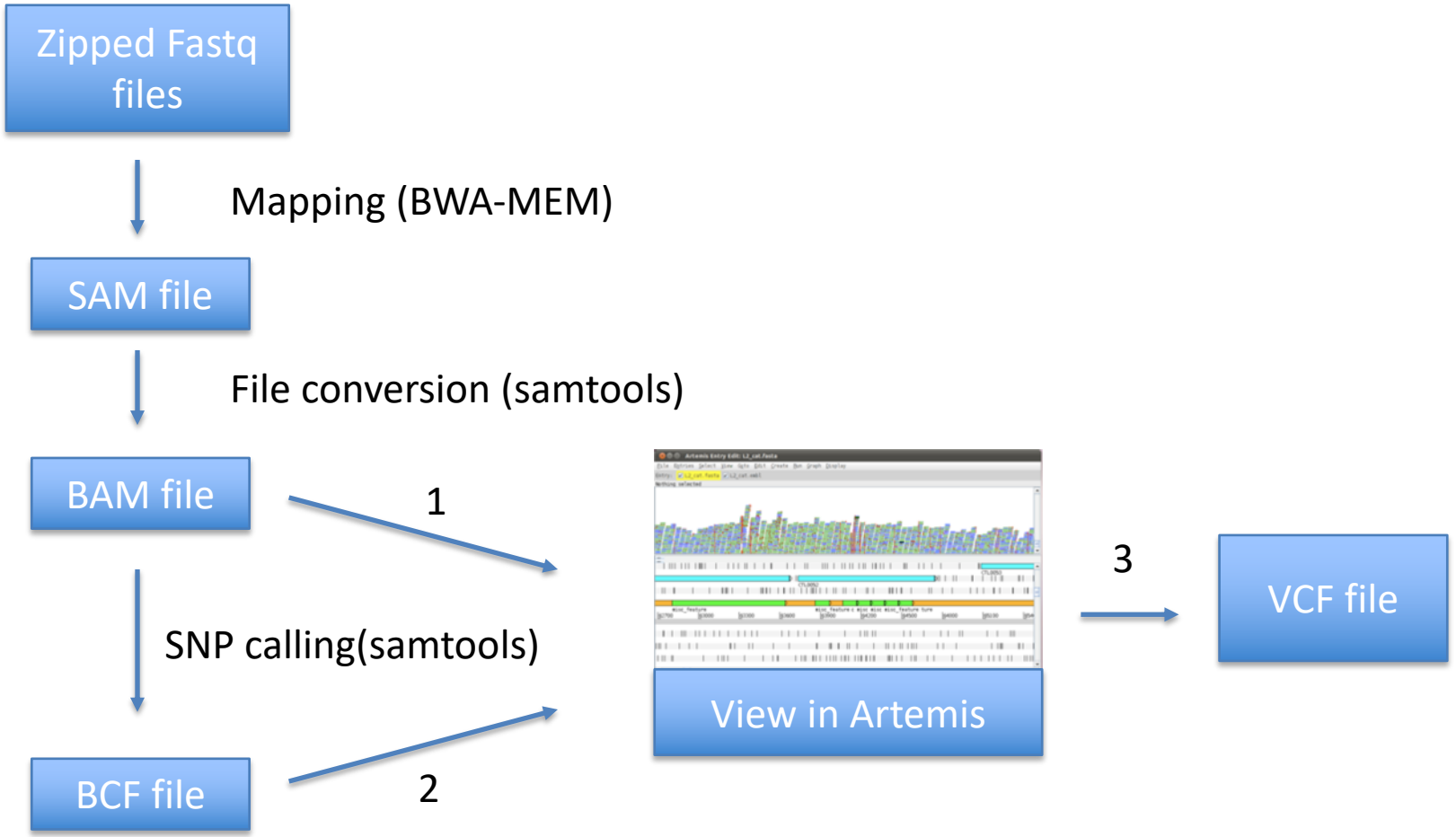
Splice-aware aligners for RNA-seq

STAR

★ HISAT2

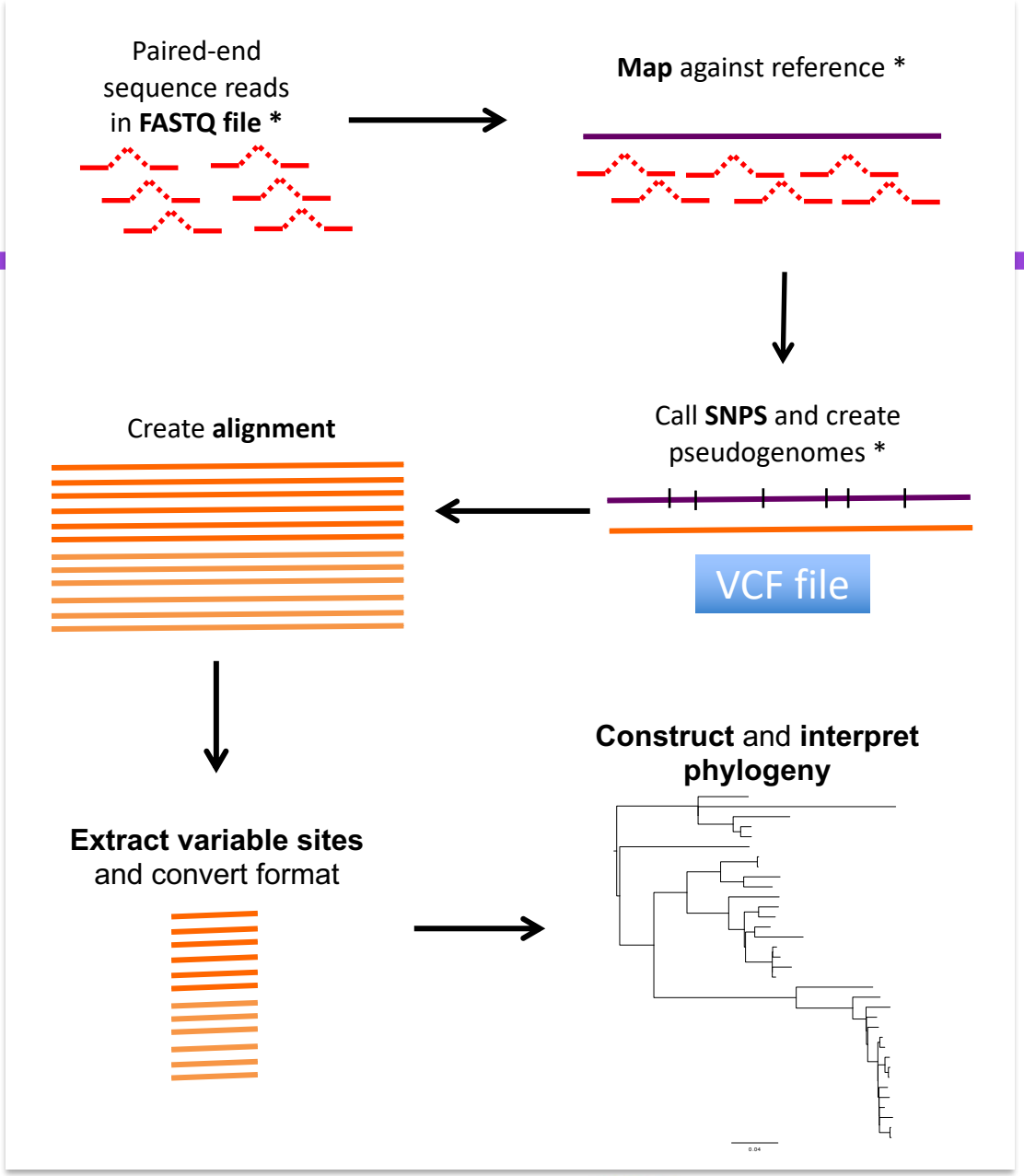


Module 3: Mapping sequence reads workflow



Workflow

From **Mapping** to **Phylogenetic trees**: process to infer genetic relationships between strains



Choosing your reference is important!

You must think carefully about what reference genome you want to use!

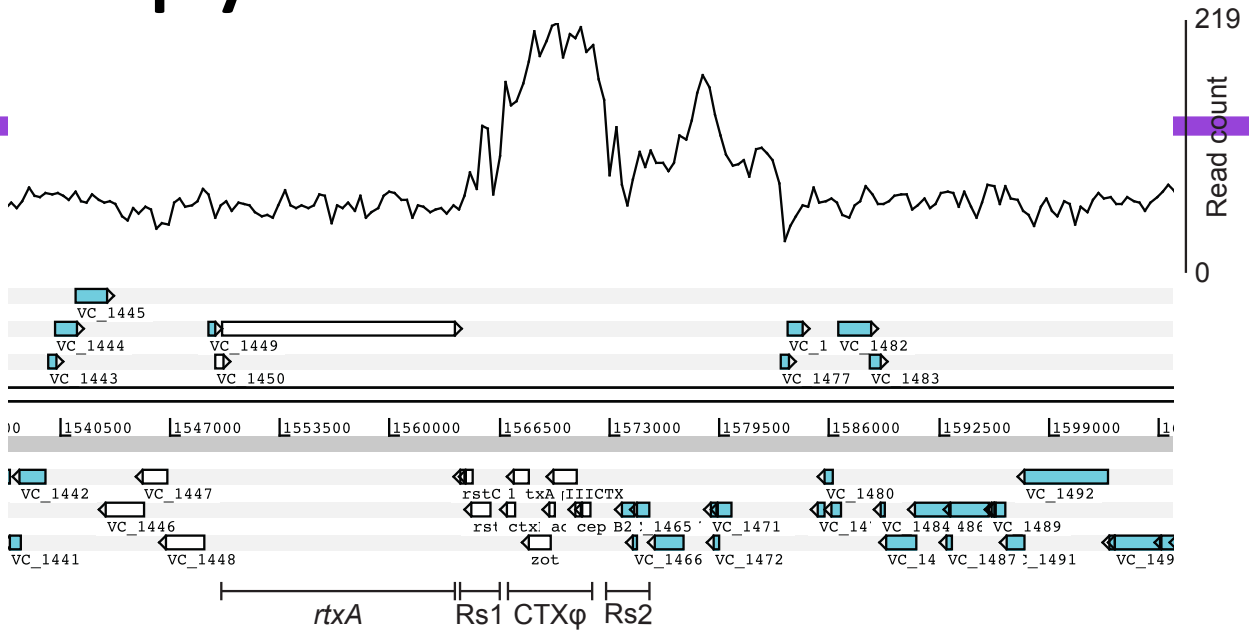
You won't find things in your sample that are not in the reference!

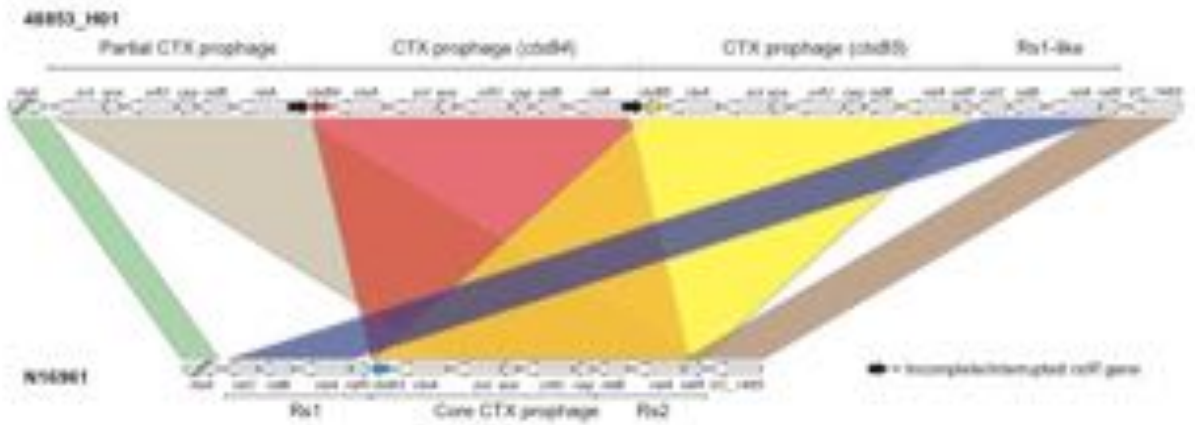
Why do we choose the mapping approach?

- We can capture information given the availability of a suitable reference genome:
  - Single Nucleotide Polymorphisms (SNPs),
  - insertions and deletions (indels)
  - Copy Number Variants (CNVs) between variants of the same bacteria.
  - Presence / absence of genes (AMR)

- As sequences diverge from the reference, mapping becomes progressively less effective

Copy number variation





wellcome connecting science

Gene presence/absence: AMR

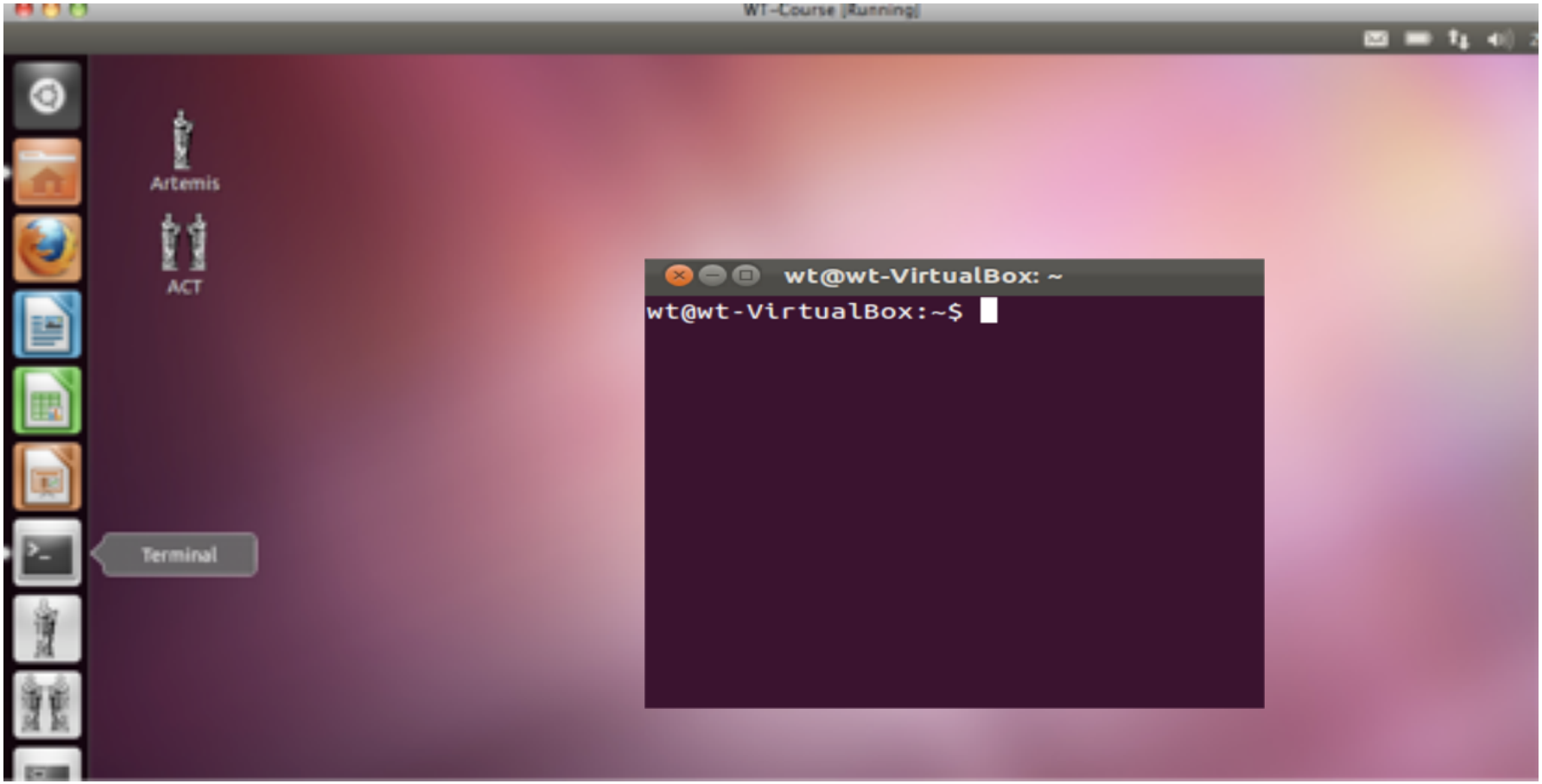
- Reference choice is important!
- Absence/Deletions is easier to spot

- To identify insertions is a little tricky.

Gene insertions/novel genes

- In this instance you must investigate:
  - Metadata (phenotype)
  - Map to a different reference
  - If AMR/Virulence – map to a database
  - Assembly

The exercise:

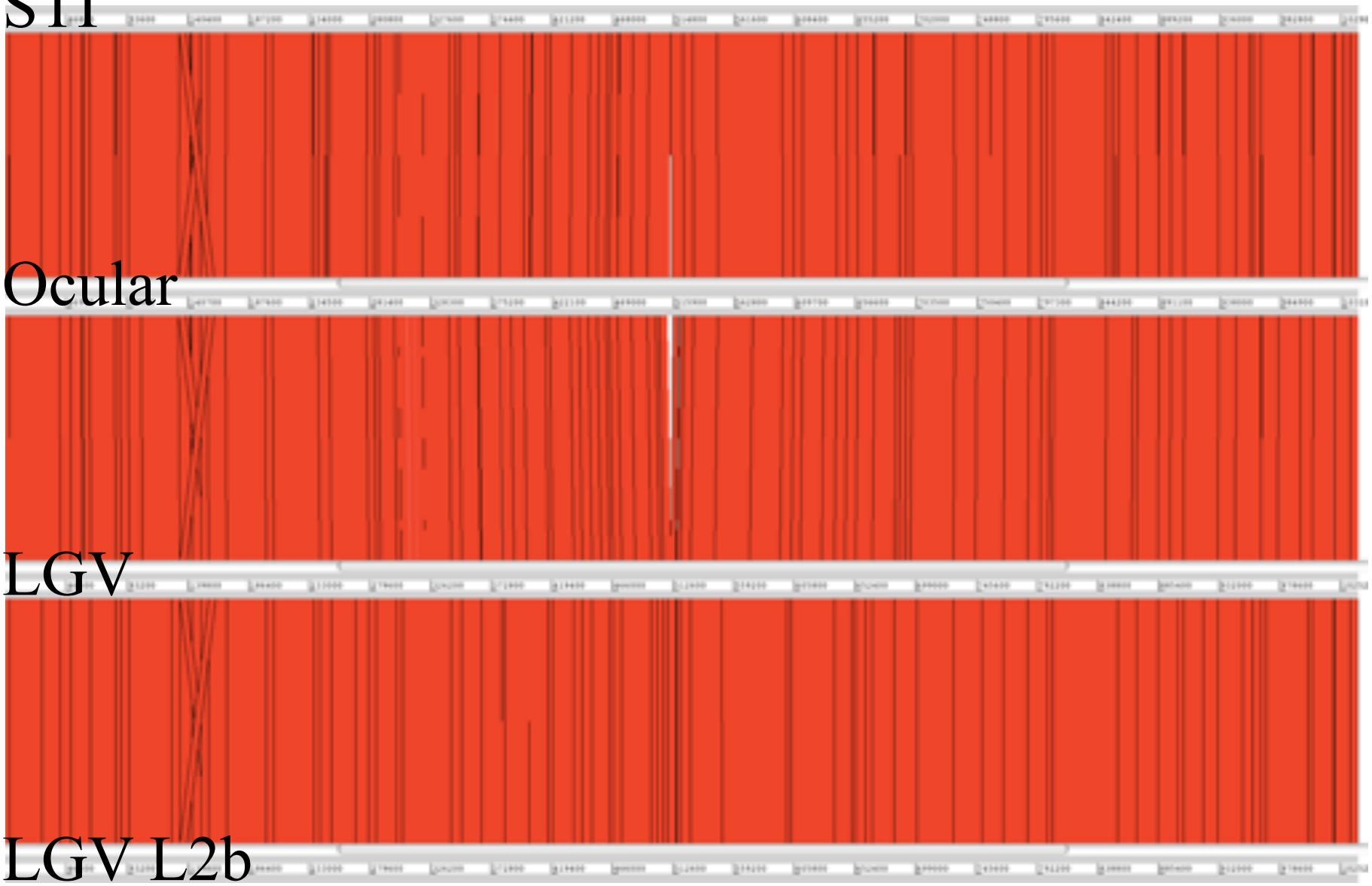


Chlamydia trachomatis

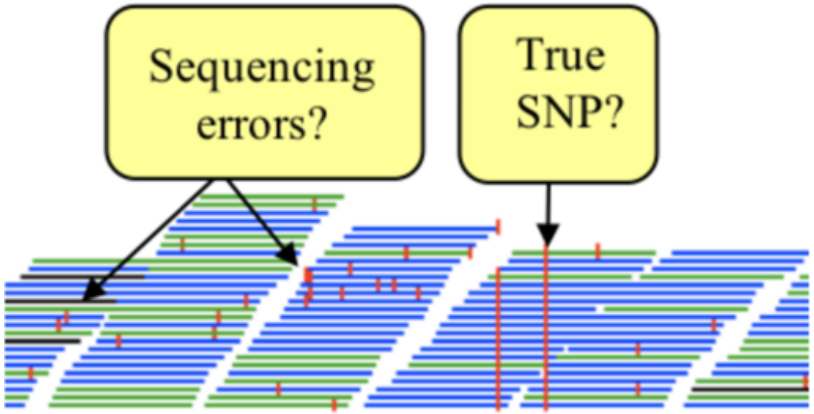
Classification by tissue tropism

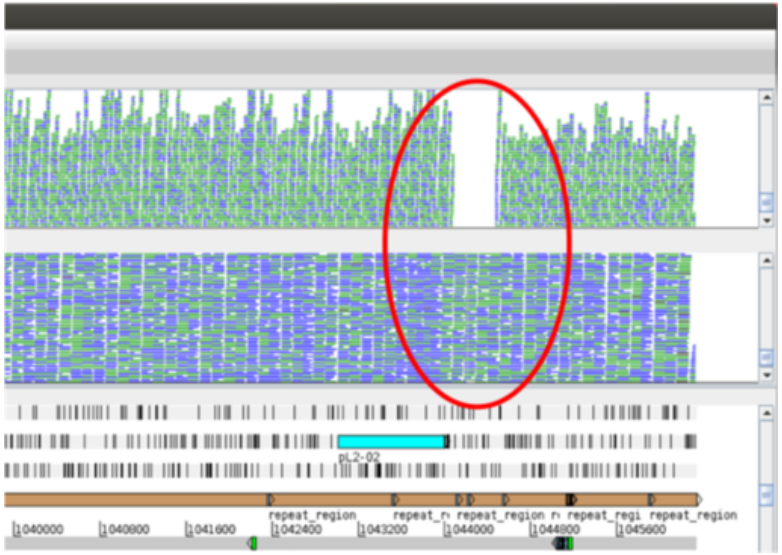
https://www.futuremedicine.com/doi/full/10.2217/fmb.13.80

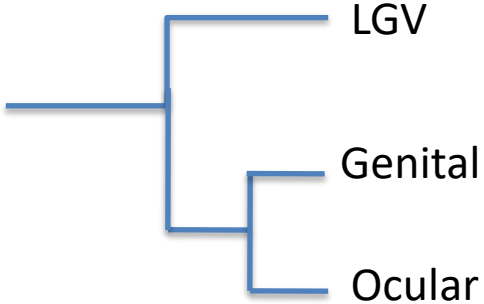
Whole Genome alignments. How do you distinguish between the strains?



SNPs and presence/absence of genes

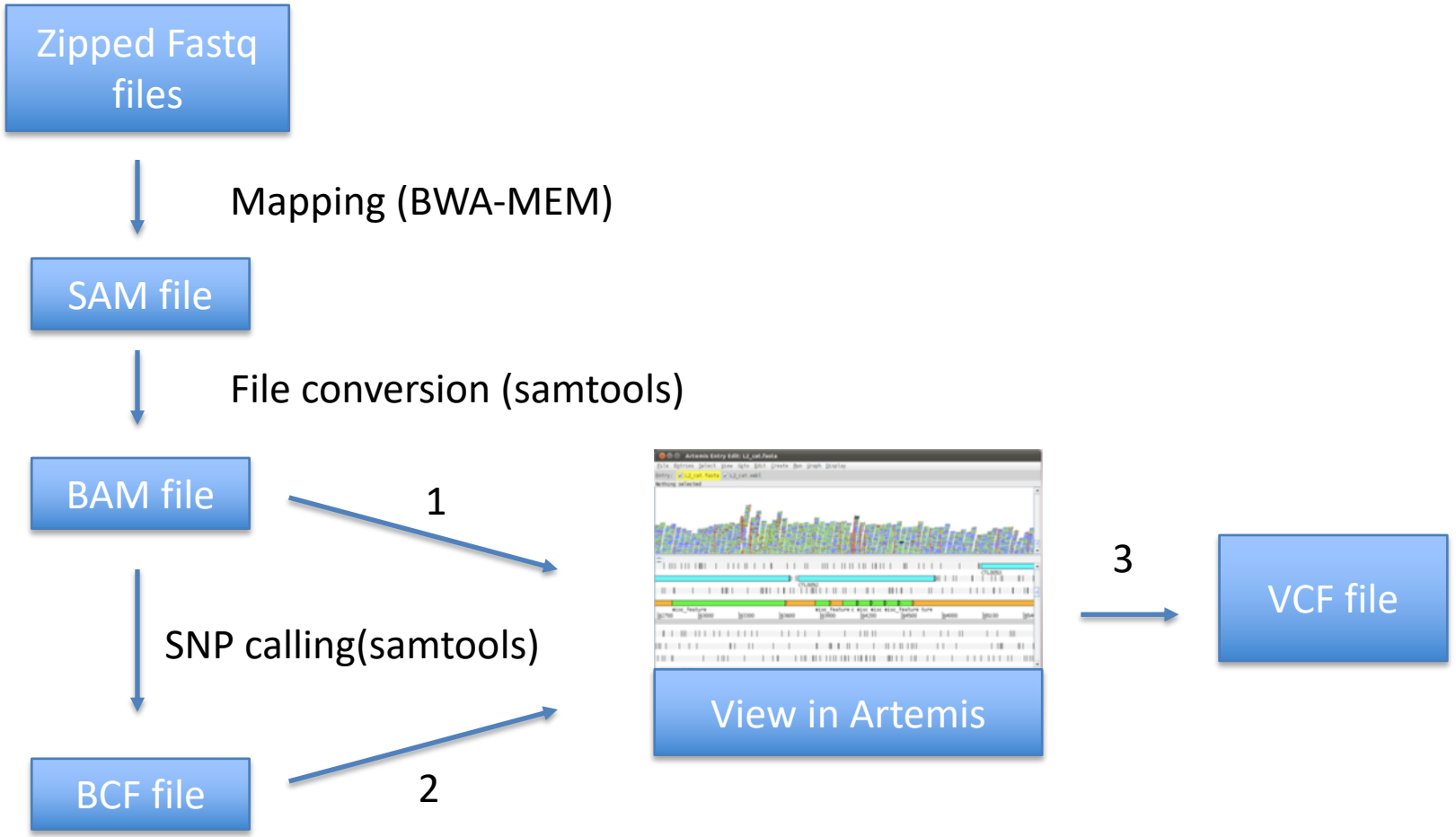






wellcome connecting science

Module 3: Mapping sequence reads workflow



Wrap-up

From **Mapping** to **Phylogenetic trees**: process to infer genetic relationships between strains