

# *Module 3*

## Mapping short reads

Working with pathogen genomes

10<sup>th</sup> - 14<sup>th</sup> 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

					
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 <sup>†</sup>	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 <sup>†</sup>	400 million	1.1 billion*
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp

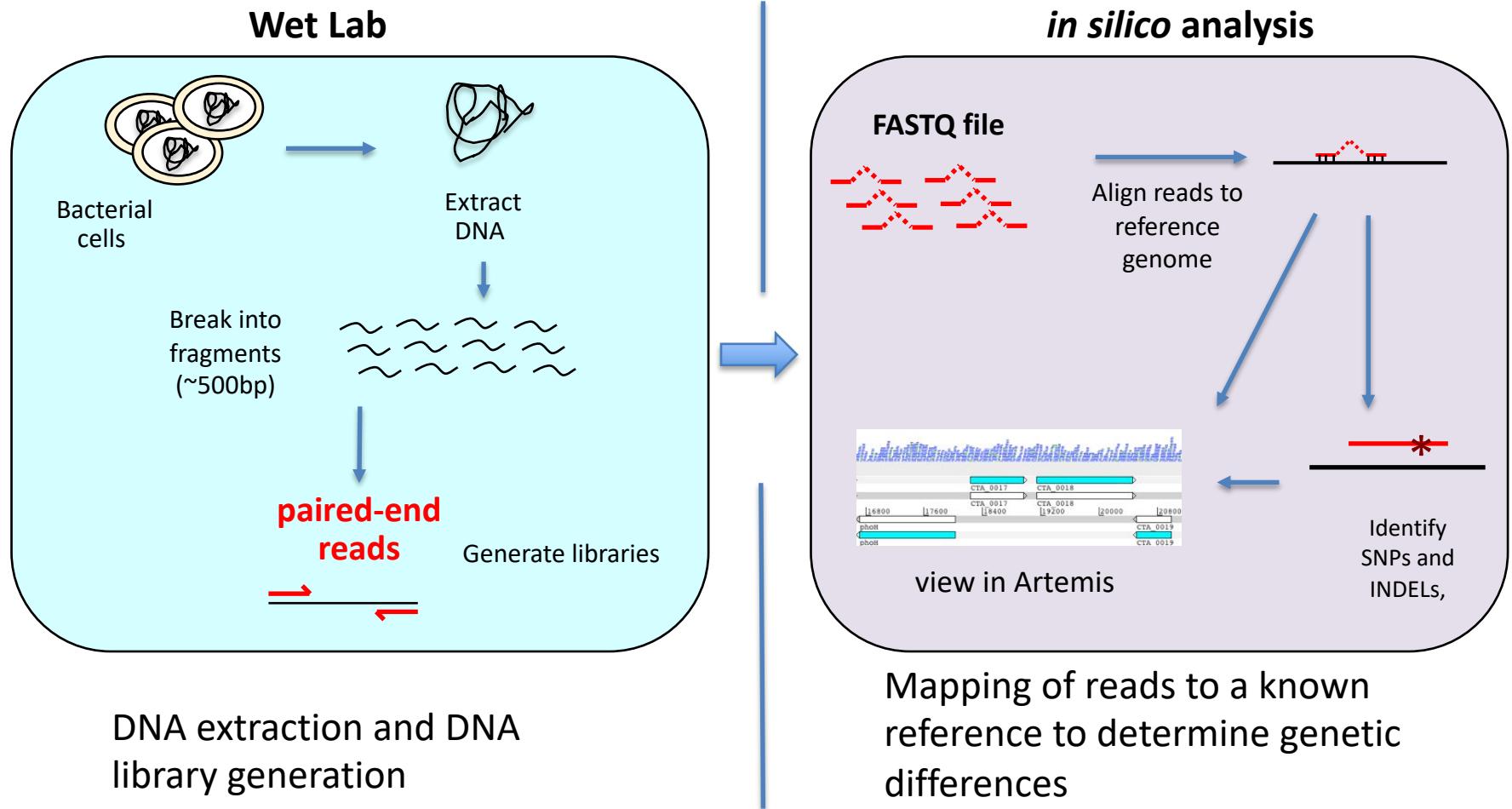
<https://emea.illumina.com/systems/sequencing-platforms.html>

# 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

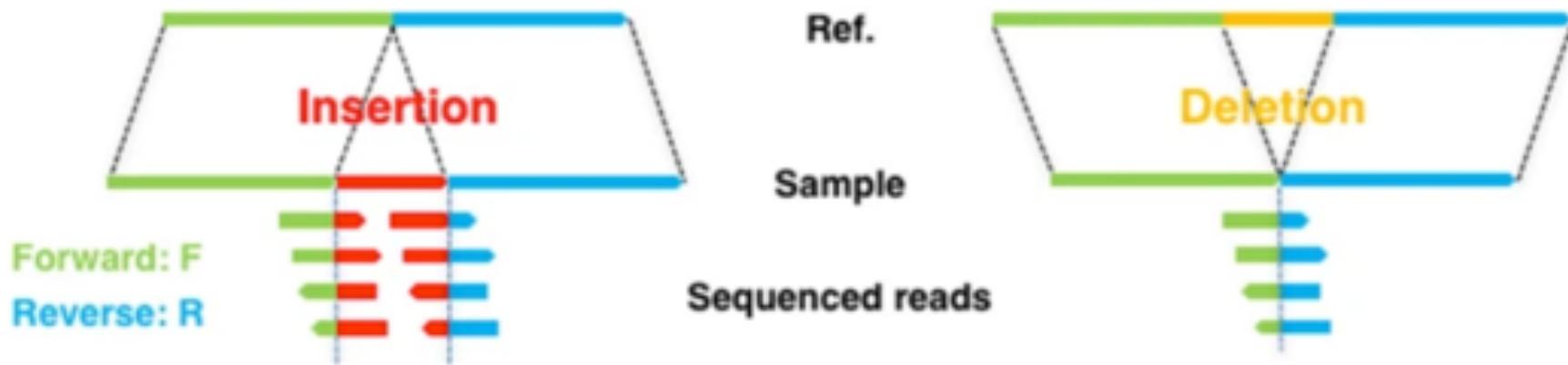


# Single Nucleotide Polymorphisms (SNPs)

Reference	CCGTTAGAGTTACAATTCTGA
Read 2	TTAGAGT <b>A</b> ACAA
Read 3	CCGTTAGAGT <b>T</b> A
Read 4	<b>T</b> TACAATTCTGA
Read 5	GAGT <b>A</b> ACAA
Read 6	TTAGAGT <b>A</b> ACAAT

[https://aschuerch.github.io/MolecularEpidemiology\\_AnalysisWGS/09-SNPphylo/index.html](https://aschuerch.github.io/MolecularEpidemiology_AnalysisWGS/09-SNPphylo/index.html)

# INDELS



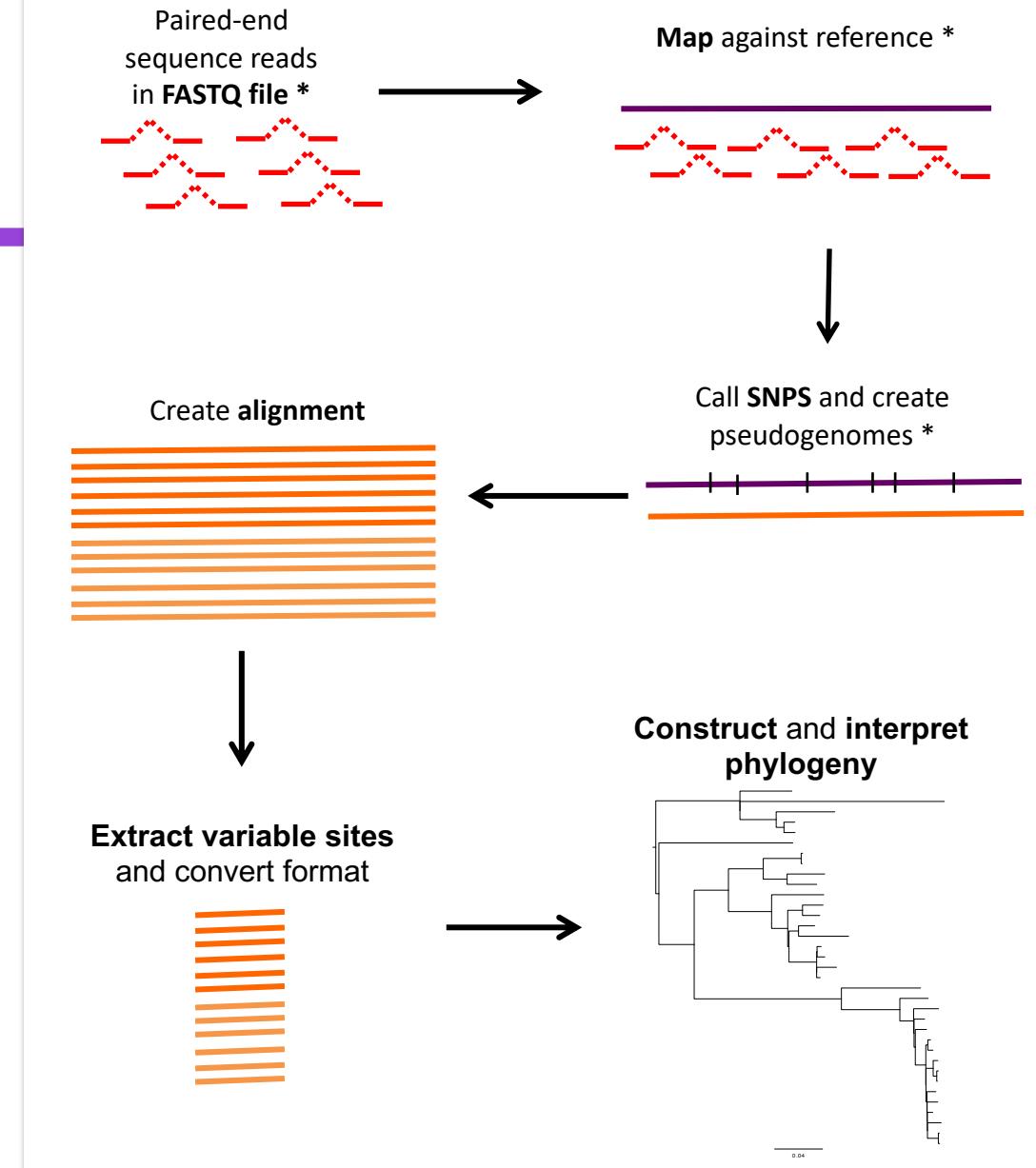
<https://www.nature.com/articles/s41598-018-23978-z>

# Why do we choose the mapping approach?

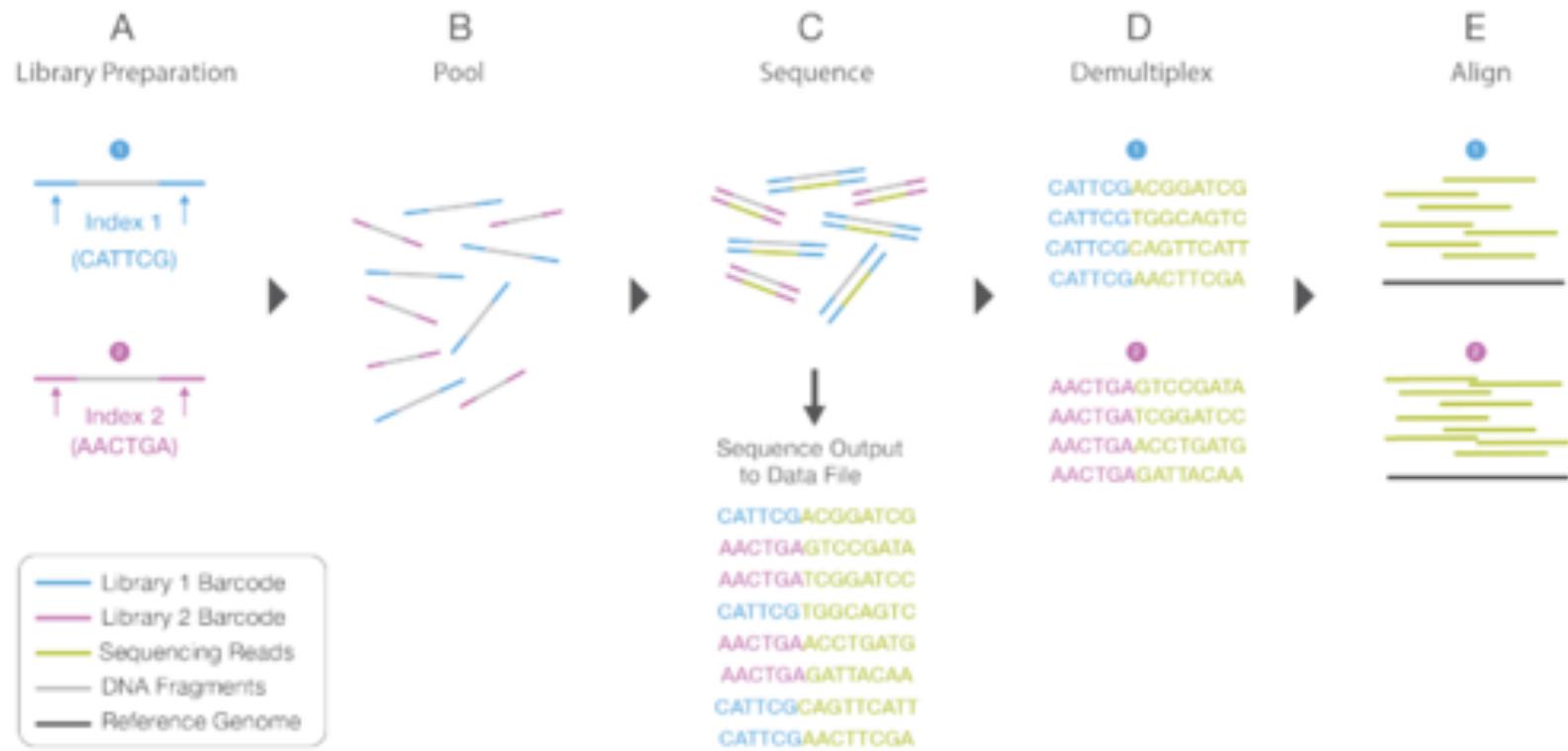
- We can capture information given the availability of a suitable reference genome:
  - Single Nucleotide Polymorphisms (SNPs),
  - insertions and deletions (indels)
  - Copy Number Variants (CNVs) between variants of the same bacteria.
  - Presence / absence of genes (AMR)
- As sequences diverge from 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](https://emea.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf)

# Sequence output to Demultiplex

# FASTQ file

# Fastq format

```
1 @SEQ_ID
2 GATTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTGTTCAACTCACAGTT
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!

Secure | https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

## Babraham Bioinformatics

About | People | Services | Projects | Training | Publications

### FastQC

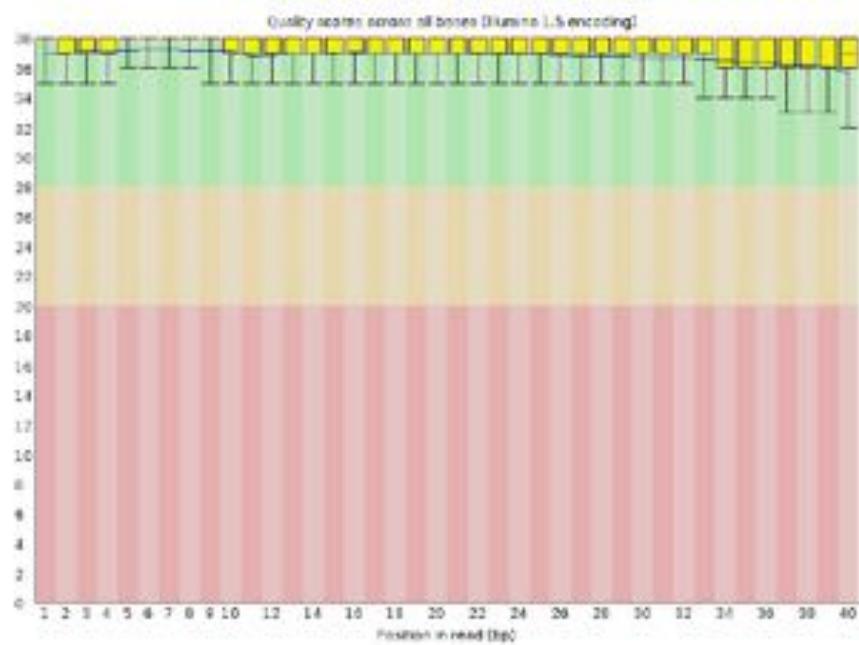
Function	A quality control tool for high throughput sequence data.
Language	Java
Requirements	A suitable Java Runtime Environment
Code Maturity	Stable. Mature code, but feedback is appreciated.
Code Released	Yes, under <a href="#">GFDL v1.2 license</a> .
Initial Contact	<a href="#">Source Analysis</a>

[Download Now](#)

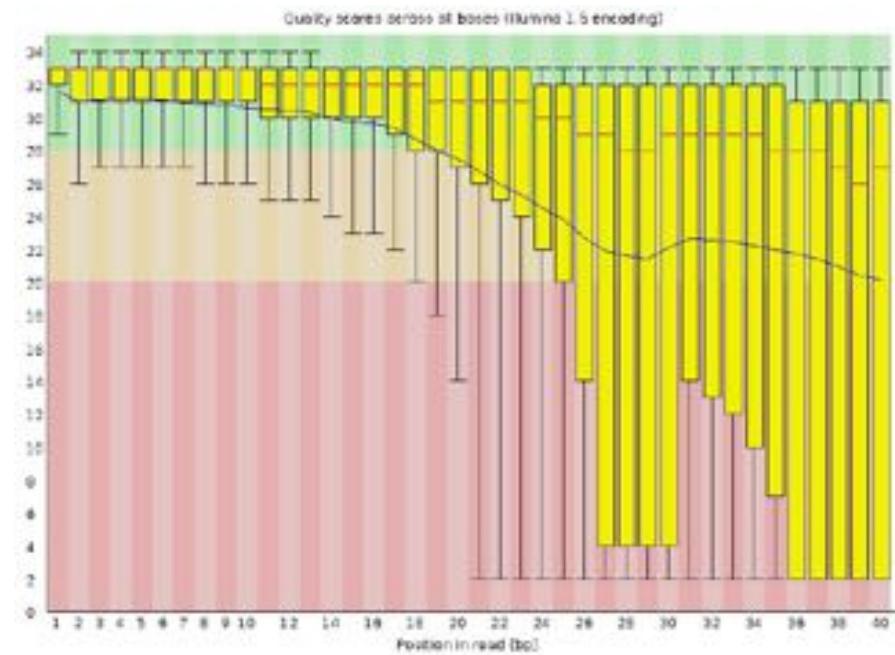


# Good vs Bad Fastq files

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
NAAACTACTACACCCACTCAGGACACCAGGGACATCATT
GCTGACGCCACGGCCTCACAGTGCTGAGCTGATGAT
+
$705291596=>>>=>=>=>=>=>535:6=>=>>=>=
5;;318656:==991/1,-0,0015204.1
@IL24_5151:3:1:2173:904#9/1
NTTTAACCGTACCTCACCAAGGATTATCGCAGGCGGATTC
CTGGTGATTAATTCAAAAATAGCGTTAATCCA
+
$948883999>=>>>=>>=>9>>=>=>=>=>:>:::==
=>55:88>=>9:0;:==>=>=>>
@IL24_5151:3:1:2948:912#9/1
NCCACCAGACACTGTCCGCAACCCCGGTAAAGGGGCAAC
GTTAGAACATCAAACATTAAAGGGTGGTATTCAAGG
+
```



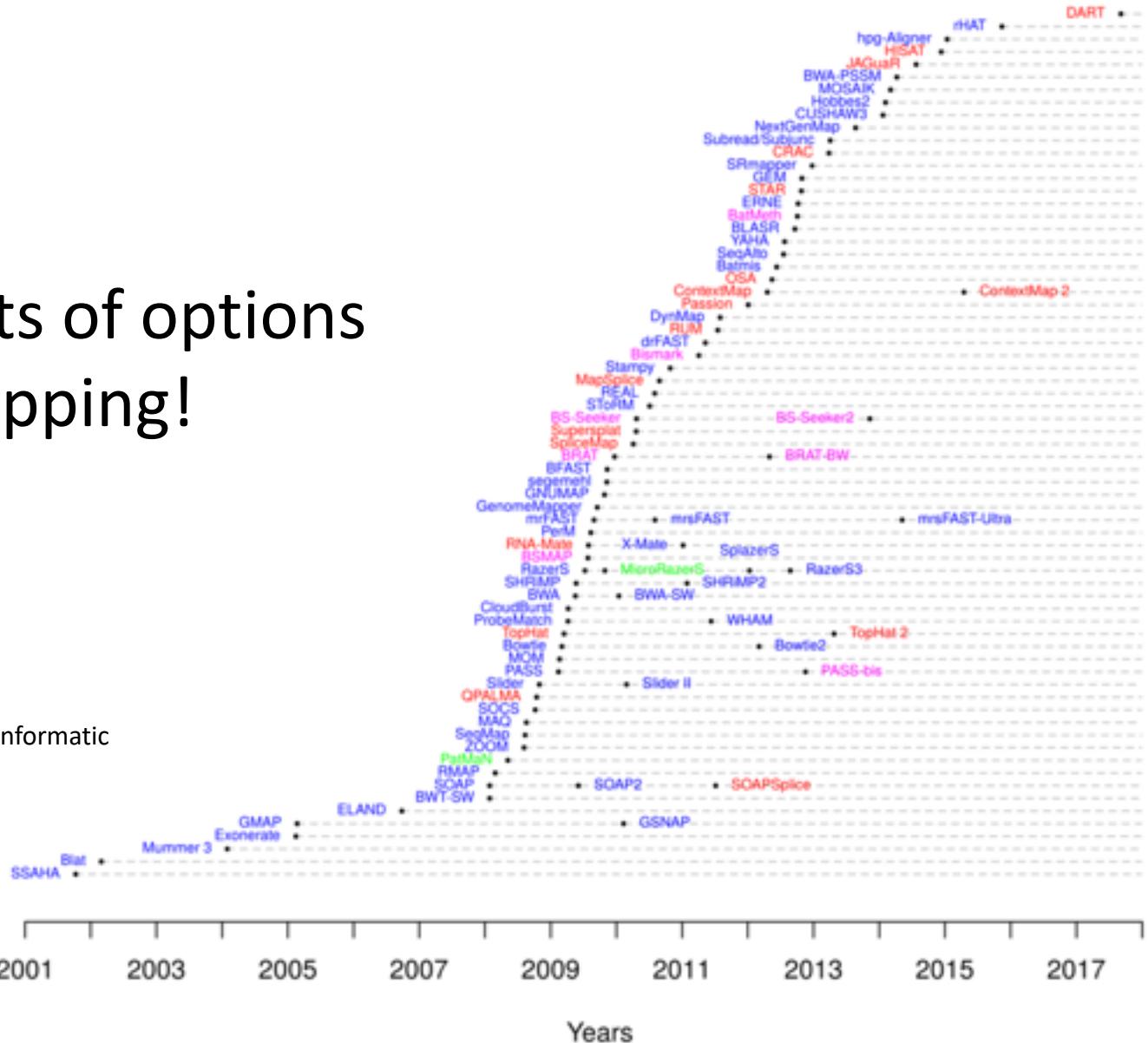
## Reference – in fasta format

```
>reference sequence
ATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAGCGGCAGCGGGAAAGTAGTTT
TACTTTGCCGGCGAGCGGCGGACGGGTGAGTAATGCTCTGGGAAACTGCCTGATGGAGG
GATAACTACTGAAACGGTAGCTAACCGCATGACCTCGTAAGAGCAAAGTGGGGAC
TTCGGGCCTCACGCCATCGGATGTGCCAGATGGGATTAGCTAGTAGGTGGGTAATGG
TCACCTAGGCAGCAGATCCCTAGCTGGTCTGAGAGGGATGACCAGCCACACTGGAACGTGAG
CACGGTCAGACTCCTACGGGAGGCAGCAGTGGGAATATTGCACAATGGCGCAAGC
GATGCAGCCATGCCCGTGTGAAGAAGGCCTCGGGTTGAAAGCACTTCAGCGAG
AGGAAGGCAGTCGTGTTAATAGCAGATTGACGTTACTCGCAGAAGAAGCACCGGC
```



# There are lots of options for mapping!

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



# Comparison of different mappers

Mapper	Code	Availability	Version	OS	Number of cores	Seq/Pos	Input	Output	Min. Pos.	Max. Pos.	Min. Score	Max. Score	Score Range	Scored Alignment	Scored Alignment	Version	OS	Supporting	Index				
Bamtools	Bamtools	0.6	1.00	Linux, Unix	50	-	(C)FASTA(AQ)	Name	25	100	0	0	N	A,B	0	N	Y	N	No Reference				
Bamtools	DNA	0.6	0.0	Linux, Mac	50	1.00	FASTA(Q)	SAM	-	-	0	0	N	A,B	0	N	Y	T	No Reference				
BLAST	DNA	0.6	0.70	Linux, Mac	500	1.00,A,Pos	(C)FASTA(AQ)	SAM TSV	16	104	0.2	0.2	N	A,B	0	N	SM	Y	No Reference				
BLASR	DNA	0.6	1.0	Linux, Unix	50	-	(C)FASTA(NTR)	SAM TSV	50	1000000	0.2	0.2	N	A,B	0	N	SM	Y	No Reference				
Blast	DNA	0.6	0.0	Linux, Mac	4200	N	FASTA(Q)	TSV BLAST	11	8000K	Score	Score	N	A,B	0	N	SM	Y	De novo Reference				
Bowtie	DNA	0.6	0.12.7	Linux, Mac, Windows	11207	1.00,A,Pos,P	(C)FASTA(AQ)	SAM TSV	4	1K	Score	Score	N	A,B	0	N	SM	Y	De novo Reference				
Bowtie2	DNA	0.6	2.3.0	Linux, Mac, Windows	8886	1.00	FASTA(Q)	SAM TSV	4	800K	Score	Score	N	A,B	0	N	SM	Y	No Reference				
BWA	DNA	0.6	1.2.3	Linux	60	-	FASTA(Q)	TSV	-	-	0	0	N	A,B	0	N	SM	Y	No Reference				
BWA-SW	DNA	0.6	2.0.1	Linux	50	-	FASTA(Q)	TSV	-	-	0	0	N	A,B	0	N	SM	Y	No Reference				
BWA-SWard	DNA	0.6	2.0.0	Linux, Unix, Mac	187	-	FASTA(Q) (pos)	SAM	10	200	Score	Score	N	A,B	0	N	SM	Y	No Reference				
BWA-SWard	DNA	0.6	2.1.0	Linux, Unix, Mac	347	-	FASTA(Q) SAMBAM	SAM BAM Native	20	144	0	1	N	A,B	0	N	SM	Y	No Reference				
BWA-SW	DNA	0.6	0.4.2	Linux, Mac, Windows	12541	1.00,A,Pos,P	FASTA(Q)	SAM	4	200	0.0	0.0	N	A,B	0	N	SM	Y	No Reference				
BWA-FISSER	DNA	0.6	0.8.11	Linux	26	1.00	FASTA(Q) SAMBAM	SAM	4	200	0.0	0.0	N	A,B	0	N	SM	Y	No Reference				
BWA-SW	DNA	0.6	0.4.2	Linux, Mac, Windows	2486	1.00,A,Pos,P	FASTA(Q)	SAM	4	1000K	0.0	0.1	N	A,B	0	N	SM	Y	No Reference				
BWA-SW	DNA	0.6	2.0070916	Linux	120	N	FASTA(Q)	TSV	-	-	0	0	N	A	0	N	SM	Y	No Reference				
CLC Mapper	DNA	Com	4	-	1.00	1.00,A,Pos,Pos,P	FASTA(Q)	SAM	10	100	Score	Score	N	A,B	0	N	SM	Y	No Reference				
CloudMapper	DNA	0.6	1.1	Linux, Mac, Windows	400	-	FASTA(Q)	TSV	-	-	0	0	N	A,B	0	N	Cloud	Y	No Results				
CloudMapper	DNA	0.6	0.2	Linux, Unix, Mac	27	1.00,A,Pos,Pos,P	FASTA(Q)	TSV	1	5000	0.0	0.0	N	A,B	0	N	Cloud	Y	No Results				
CloudMapper	DNA	0.6	2.0	Windows, Linux, Unix, Mac	5	1.00,A,Pos,Pos,P	FASTA(Q) SAMBAM	SAM	10	200	0.0	0.0	N	A,B	0	N	Cloud	Y	No Results				
CRAC	DNA	0.6	2.0.0	Linux, Unix, Mac	47	1.00,A,Pos,P	(C)FASTA(AQ) RAW	SAM	10	1000	Score	Score	N	A,B	0	N	SM	Y	No Reference				
CURAGO	DNA	0.6	0.0.0	Linux	20	1.00,A,Pos,P	FASTA(Q)	SAM	10	1000	Score	Score	N	A,B	0	N	SM	Y	No Reference				
DART	DNA	0.6	1.2.8	Linux	5	-	FASTA(Q)	SAM	10	1000	Score	Score	N	A,B	0	N	SM	Y	No Reference				
dFAST	DNA	0.6	1.0.0.0	Linux, Unix	20	N	CFPFASTA(AQ)	SAM	20	200	Score	Score	N	A,B	0	N	SM	Y	No Reference				
DyNalox	DNA	0.6	0.0.20	Linux	5	-	FASTA(Q)	TSV	10	1000	Score	Score	N	A,B	0	N	SM	Y	No Reference				
ELAND	DNA	Com	1	1.00	Linux, Unix, Mac	26	-	FASTA(Q)	SAM	10	1000	Score	Score	N	A,B	0	N	SM	Y	No Reference			
ERNE	DNA	0.6	1.00	Windows, Linux, Unix, Mac	14	-	FASTA(Q) Summa	SAM	10	1000	Score	Score	N	A,B	0	N	SM	Y	No Reference				
EzMapper	DNA	0.6	0.2	Linux, Unix, Mac	918	N	FASTA(Q)	TSV	10	1000	Score	Score	N	A,B	0	N	SM	Y	No Reference				
FASTQ	DNA	0.6	1.0	Linux, Mac	280	1.00	FASTA(Q)	SAM	10	1000	Score	Score	N	A,B	0	N	SM	Y	No Reference				
GenomeMapper	DNA	0.6	0.4.3	Linux, Mac	144	-	FASTA(Q)	SAM	10	1000	Score	Score	N	A,B	0	N	SM	Y	No Reference				
GMAP	DNA	0.6	2012-04-27	Linux, Unix, Mac, Windows	988	1.00,A,Saluken,P	FASTA(Q)	SAM	10	1000	Score	Score	N	A,B	0	N	SM	Y	No Reference				
GRUMMP	DNA	0.6	3.0.2	Linux, Mac	80	-	FASTA(Q) Summa	SAM TSV	10	14	Score	Score	N	A,B	0	N	SM	Y	No Reference				
GRUMPF	DNA	0.6	2012-04-27	Linux, Unix, Mac, Windows	1136	1.00,A,Saluken,P	FASTA(Q)	SAM	10	200	Score	Score	N	A,B	0	N	SM	Y	No Reference				
HISAT	DNA	0.6	1.00	Windows, Linux, Unix, Mac	480	-	FASTA(Q)	Summa	10	1000	Score	Score	N	A,B	0	N	SM	Y	No Reference				
HISAT2	DNA	0.6	2.00	Windows, Linux, Unix, Mac	60	-	FASTA(Q)	Summa	10	1000	Score	Score	N	A,B	0	N	SM	Y	No Reference				
Holoseed	DNA	0.6	0.1	Linux	19	-	FASTA(Q)	SAM	10	1000	Score	Score	N	A,B	0	N	SM	Y	No Reference				
Hop-Aligner	DNA	0.6	>1.0	Linux	11	1.00,A,Saluken,P	FASTA(Q)	SAM	10	1000	Score	Score	N	A,B	0	N	SM	Y	No Reference				
JAligner	DNA	0.6	0.21	Linux, Unix	15	-	FASTA(Q)	SAM	10	1000	Score	Score	N	A,B	0	N	SM	Y	No Reference				
MapReader	DNA	0.6	2.4.1	Linux, Mac, Windows	5	0	FASTA(Q)	TSV	10	120	Score	Score	N	A,B	0	N	SM	Y	No Reference				
MapReader	DNA	0.6	1.15.2	Linux	410	-	FASTA(Q)	SAM	10	1000	Score	Score	N	A,B	0	N	SM	Y	No Reference				
MAQ	DNA	0.6	0.7.1	Linux, Mac	2860	1.00	(C)FASTA(AQ)	TSV	8	82	0	0	N	A,B	0	N	SM	Y	No Reference				
Massif	DNA	0.6	0.6	Windows, Linux, Mac	5	1, 1m	FASTA(Q)	SAM	20	32079	Score	Score	N	A,B	0	N	SM	Y	No Reference				
WormMapper	DNA	0.6	0.1	Linux, Unix	40	N	FASTA(Q)	SAM	10	1000	Score	Score	N	A,B	0	N	SM	Y	No Reference				
WIRL	DNA	0.6	0	Linux, Unix	1,A,Saluken,P	(C)FASTA(PH) EXP SAM GFF Counts CAP	SAM	25	10000	Score	Score	N	A,B	0	N	SM	Y	No Reference					
WOMB	DNA	0.6	0.1	Linux, Mac, Windows	48	1,00,A	FASTA(Q)	TSV	10	1000	Score	Score	N	A,B	0	N	SM	Y	No Reference				
WOMB	DNA	0.6	0.1	Linux, Unix, Mac, Windows	119	1.00,A,Saluken,P	(C)FASTA(AQ)	SAM	10	1000	Score	Score	N	A,B	0	N	SM	Y	No Reference				
mrFAST	DNA	0.6	2.0.0.1	Linux, Unix	900	-	FASTA(Q)	SAM	20	1000	Score	Score	N	A,B	0	N	SM	Y	No Reference				
mrFAST	DNA	0.6	2.4.0.6	Linux, Unix	239	-	FASTA(Q)	SAM	20	1000	Score	Score	N	A,B	0	N	SM	Y	No Reference				
Summer 3	DNA	0.6	3.2.3	Linux, Mac	25	-	FASTA(Q)	SAM	10	1000	Score	Score	N	A,B	0	N	SM	Y	No Reference				
NextGenMap	DNA	0.6	0.4.6	Linux	62	N	FASTA(Q)	TSV	10	1	Score	Score	N	A,B	0	N	SM	Y	No Reference				
RealignGCR	DNA	Bin	10.0.0	Windows, Linux, Unix, Mac	6	1.00,A,Pos,P	(C)FASTA(AQ) SAM	SAM	10	1000	Score	Score	N	A,B	0	N	SM	Y	No Reference				
ZIGA	DNA	Bin	1.0.0	Windows, Linux, Unix, Mac	64	1.00,A	FASTA(Q)	SAM	10	1000	Score	Score	N	A,B	0	N	SM	Y	No Reference				
PASS	DNA	Bin	1.00	Linux, Mac, Windows	142	1.00,A	(C)FASTA(AQ)	SAM	20	1000	Score	Score	N	A,B	0	N	SM	Y	No Reference				
PRISM-BS	DNA	0.6	0.91	Linux	14	1.00,A,Pos,P	FASTA(Q)	SAM	10	2000	Score	Score	N	A,B	0	N	SM	Y	No Reference				
Passeon	DNA	0.6	1.2.0	Linux, Unix	28	1,A,Sal,P	FASTA(Q)	SAM	10	1000	Score	Score	N	A,B	0	N	SM	Y	No Reference				
PutMut	DNA	0.6	1.2.2	Linux, Mac	140	N	FASTA(Q)	TSV	1	+	Score	Score	N	A,B	0	N	SM	Y	No Reference				
Perf	DNA	0.6	0.4.0	Linux, Unix, Mac, Windows	152	1.00	(C)FASTA(AQ)	SAM	20	1000	Score	Score	N	A,B	0	N	SM	Y	No Reference				
ProteinMatch	DNA	0.6	0	Linux, Mac	6	1,A,Sal	FASTA(Q)	ELAND	30	50	Score	Score	N	A,B	0	N	SM	Y	No Reference				
QPALMA	DNA	0.6	0.0.2	Linux, Mac	169	1,0	Specific	TSV	-	-	Score	Score	N	A,B	0	N	SM	Y	No Reference				
Raservi	DNA	0.6	1.2	Linux, Mac, Windows	165	1,0	FASTA(Q)	SAM	10	1000	TSV	TSV	11	+	Score	Score	N	A,B	0	N	SM	Y	No Reference
Raservi	DNA	0.6	0.1	Windows, Linux, Mac	81	1	FASTA(Q)	SAM	10	1000	TSV	TSV	11	+	Score	Score	N	A,B	0	N	SM	Y	No Reference
REAL	DNA	0.6	0.0.09	Linux, Mac	99	1	FASTA(Q)	TSV	4	+	Score	Score	N	A,B	0	N	SM	Y	No Reference				

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

# Good general aligners

★ bwa

bowtie2

minimap2

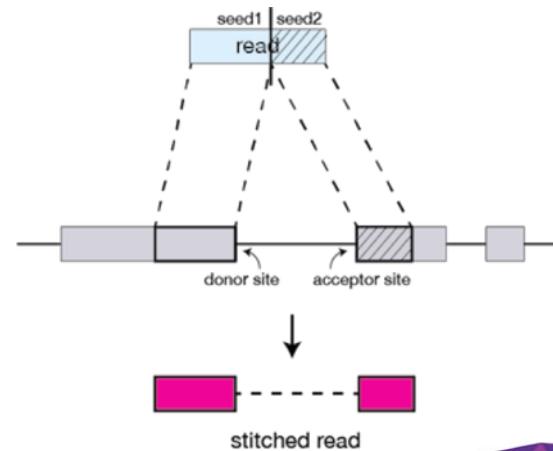


Fast, sensitive and  
easy to use!

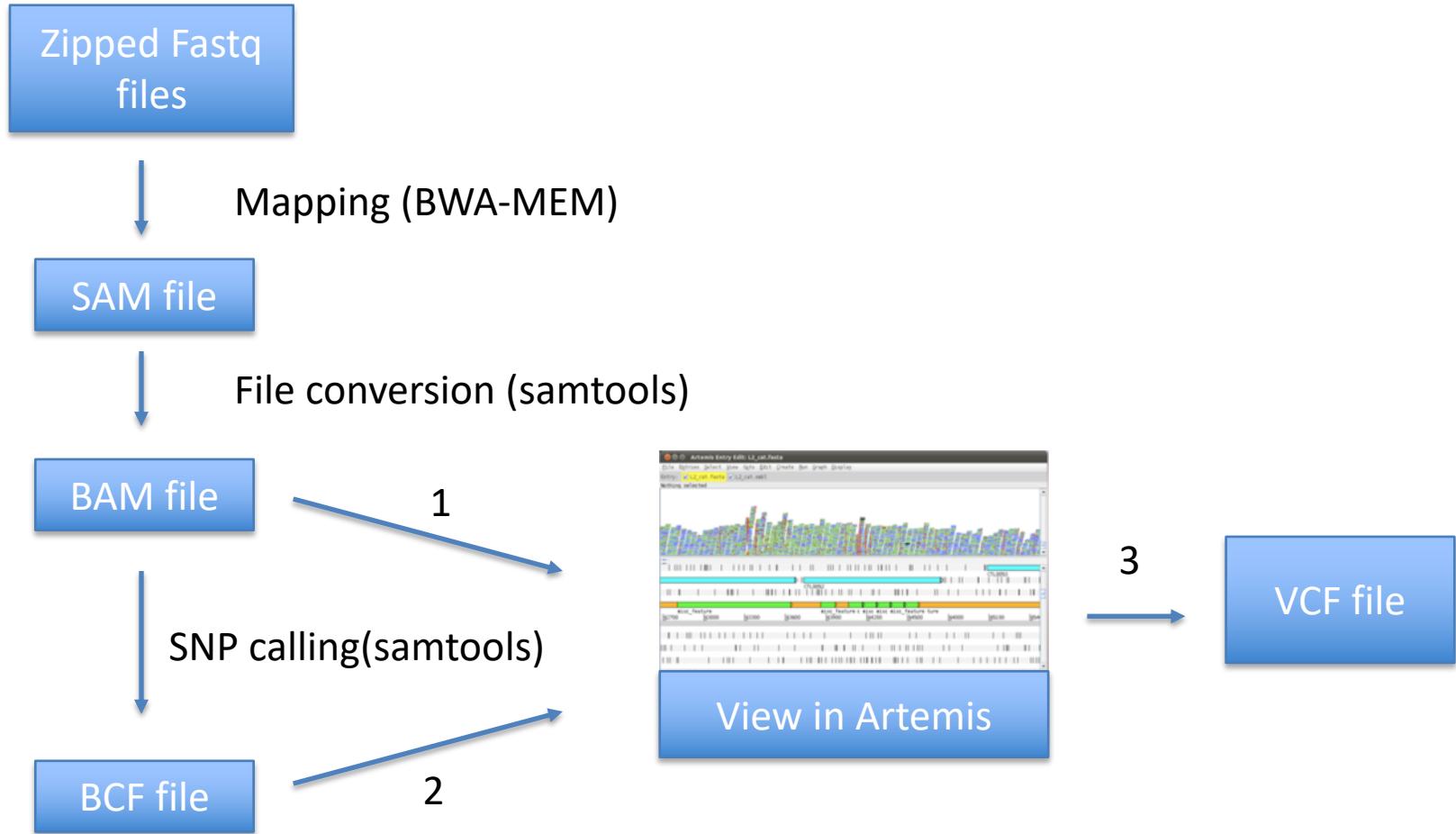
## Splice-aware aligners for RNA-seq

STAR

★ HISAT2

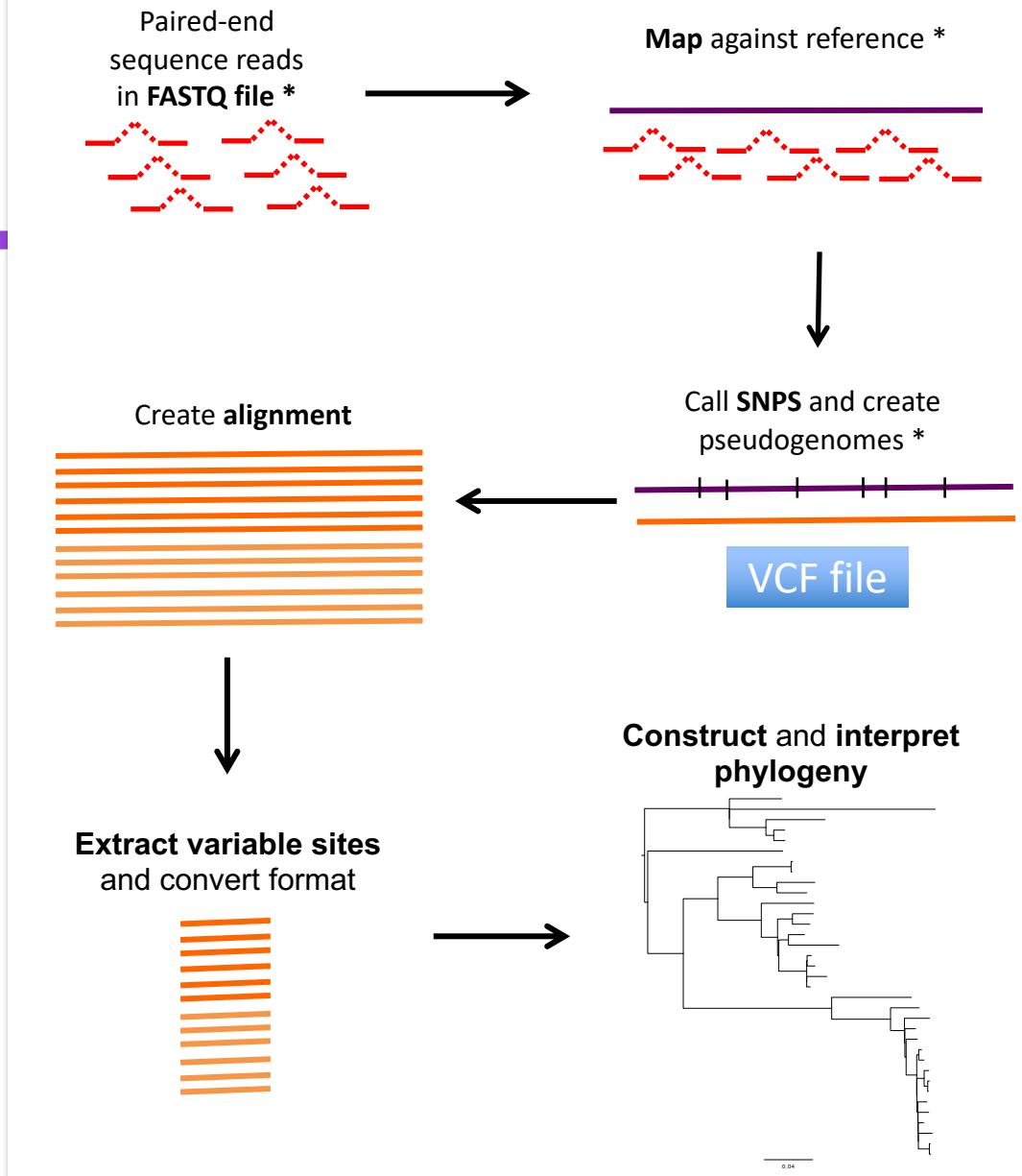


# Module 3: Mapping sequence reads workflow



# Workflow

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



# Choosing your reference is important!

---

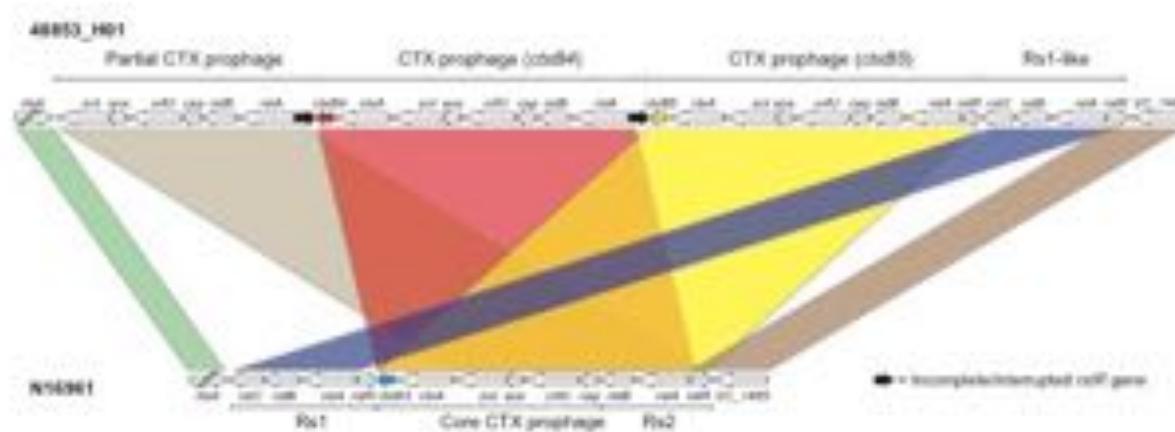
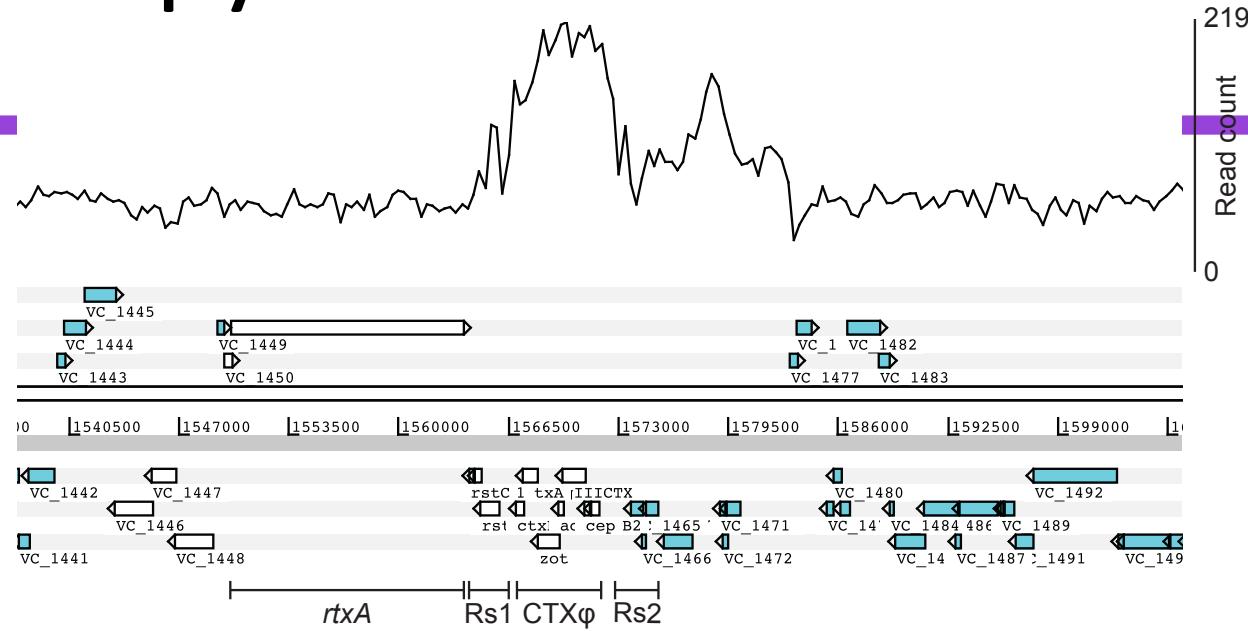
You must think carefully about what reference genome you want to use!

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



# Gene presence/absence: AMR

---

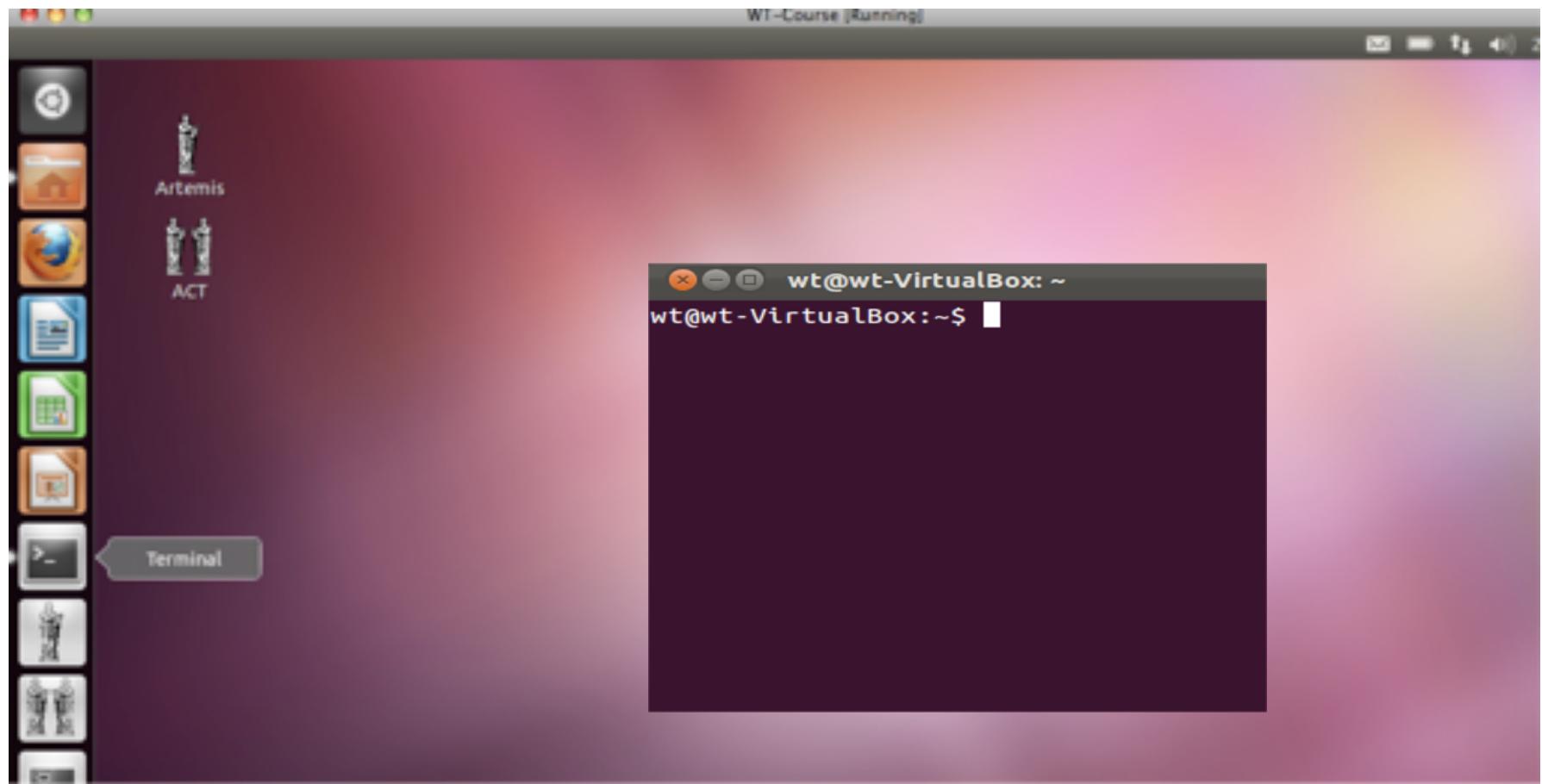
- Reference choice is important!
  - Absence/Deletions is easier to spot
- 
- To identify insertions is a little tricky.

# Gene insertions/novel genes

---

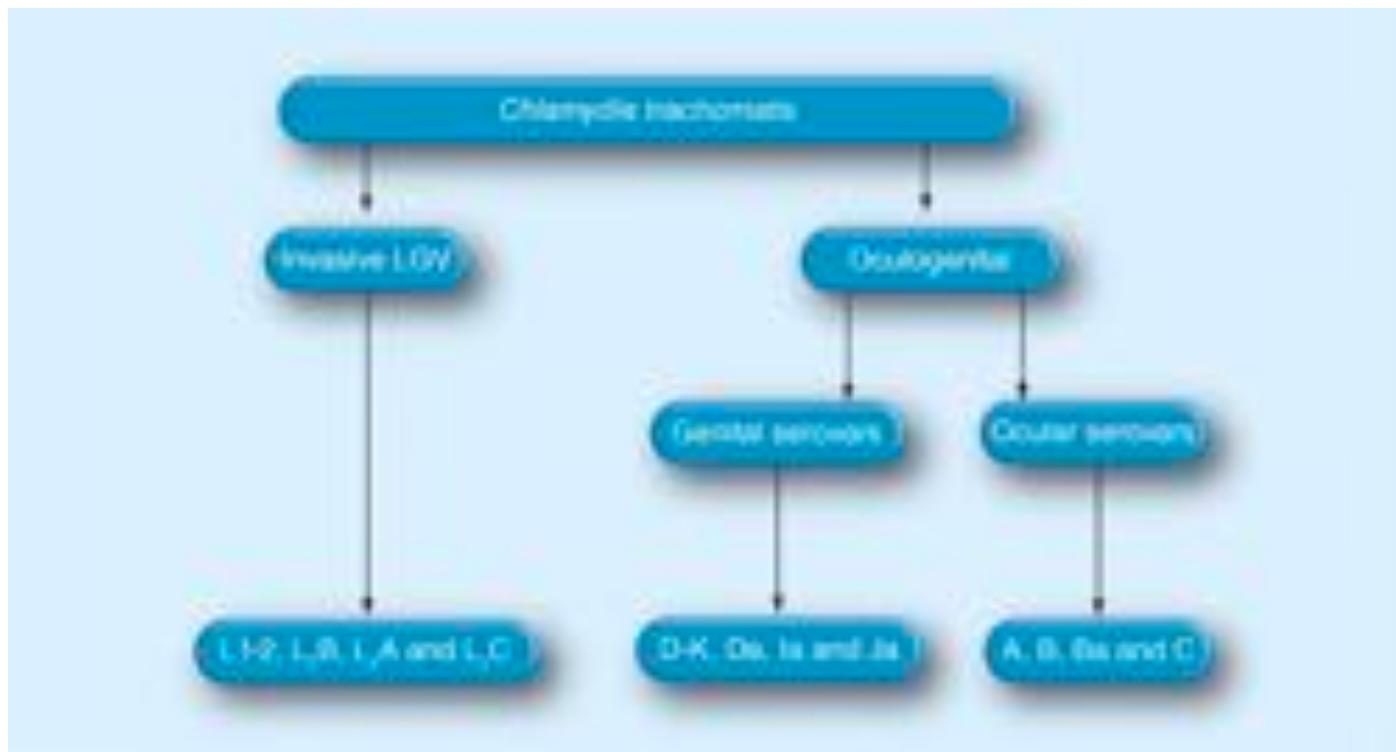
- In this instance you must investigate:
  - Metadata (phenotype)
  - Map to a different reference
  - If AMR/Virulence – map to a database
  - Assembly

# The exercise:



# *Chlamydia trachomatis*

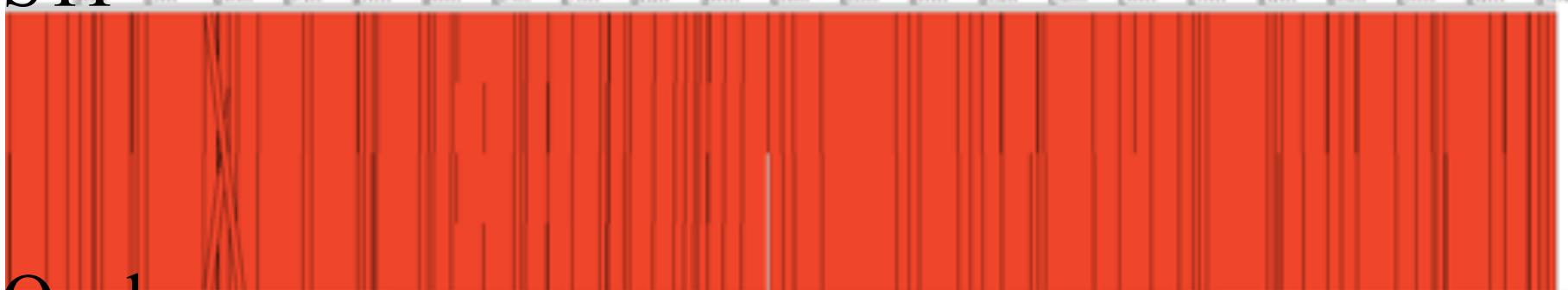
## Classification by tissue tropism



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

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

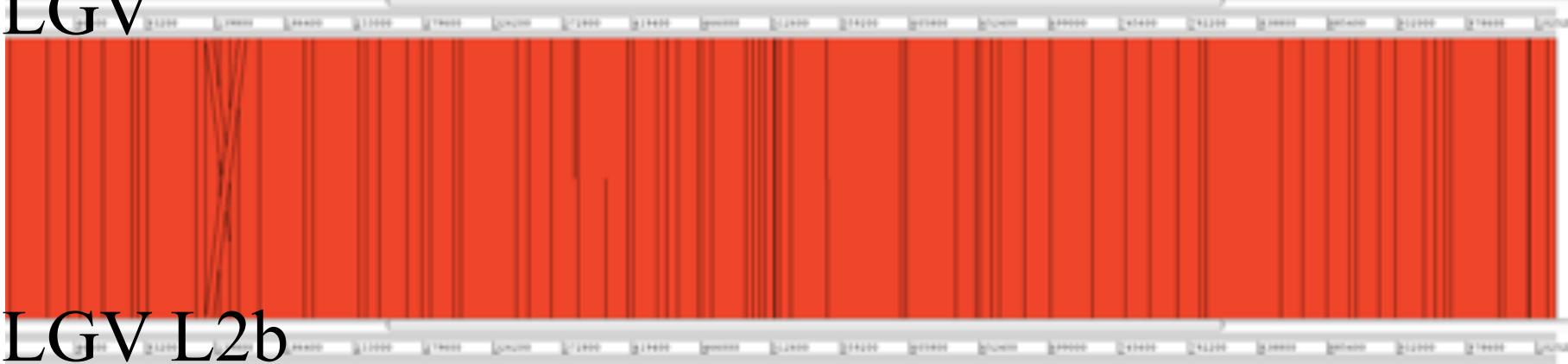
STI



Ocular

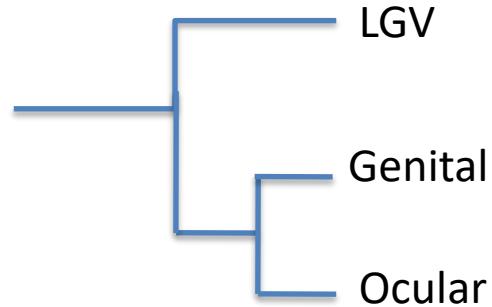
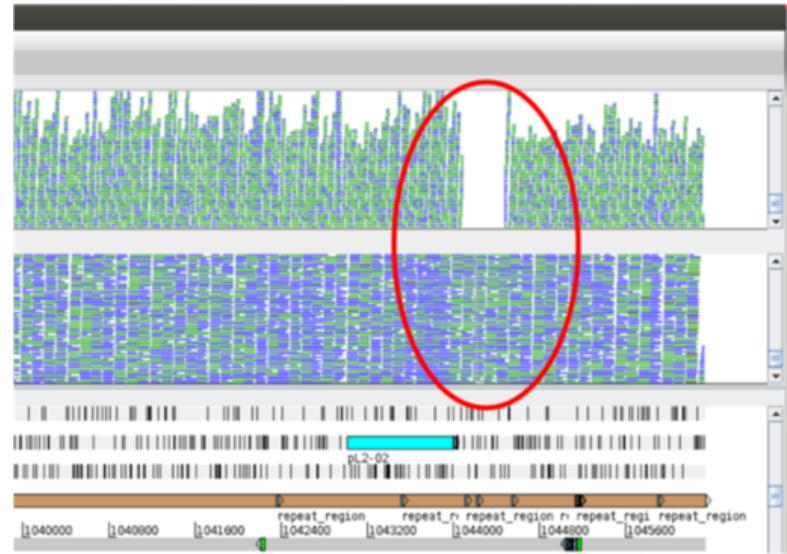
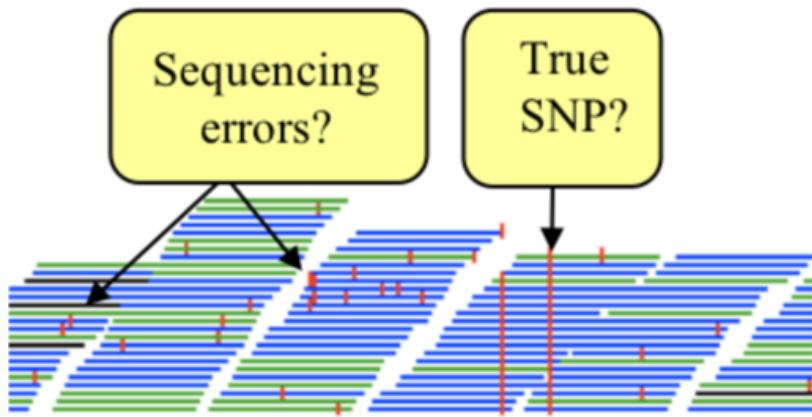


LGV

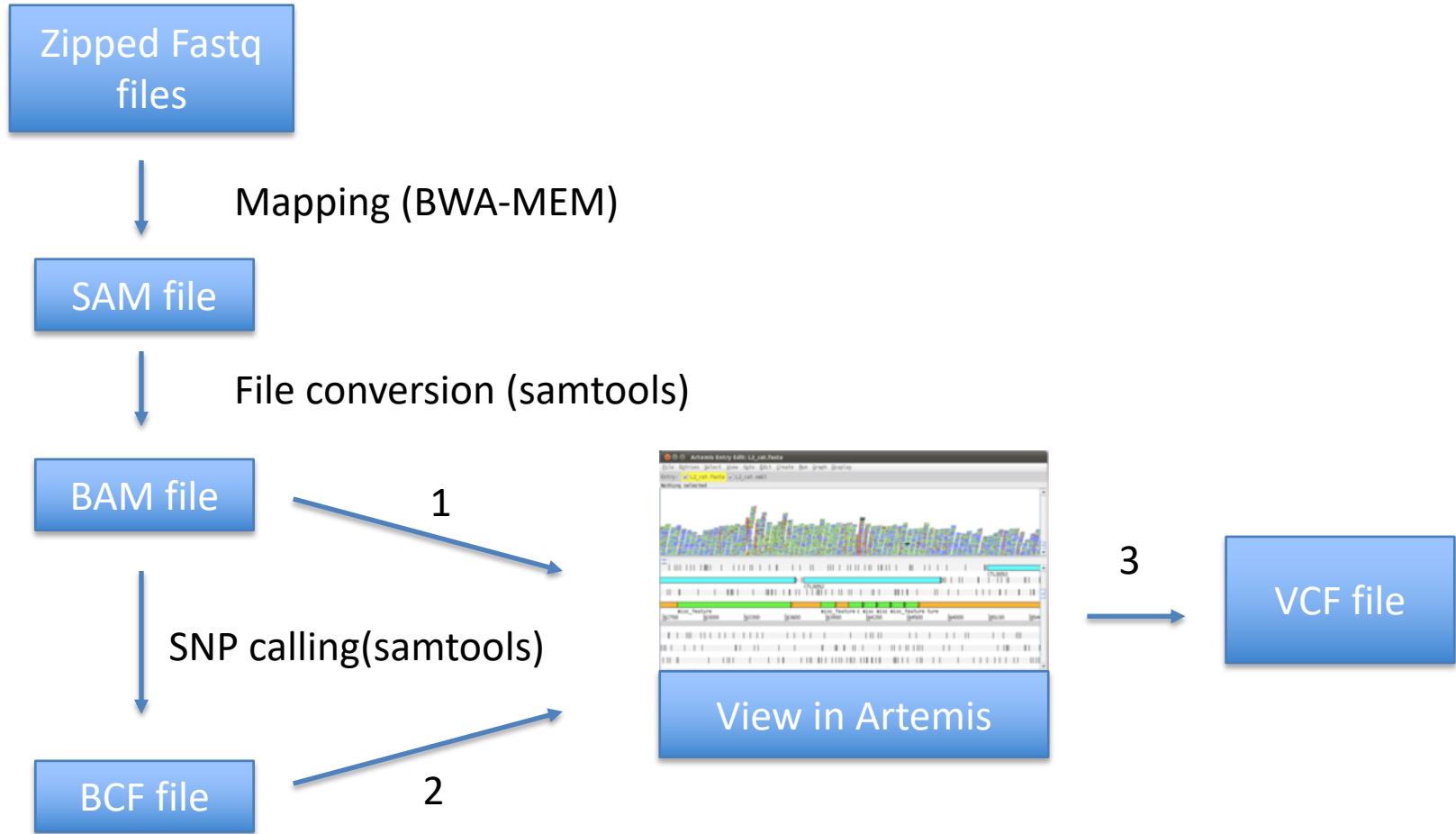


LGV L2b

# SNPs and presence/absence of genes



# Module 3: Mapping sequence reads workflow



# Wrap-up

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

