Module 3
Mapping short reads

Working with pathogen genomes

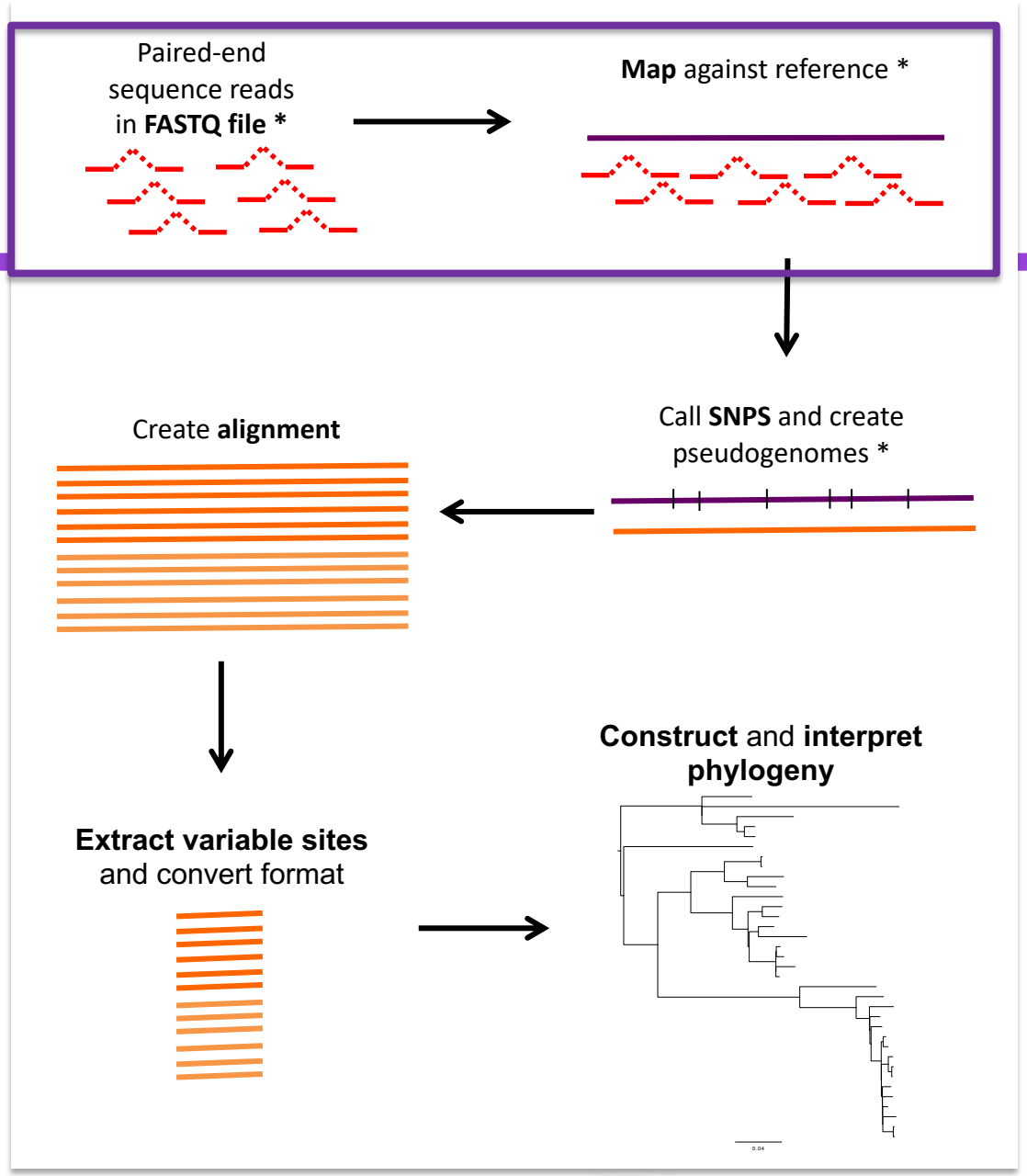
7th - 11th February 2022

Sophie Belman & Sushmita Sridhar

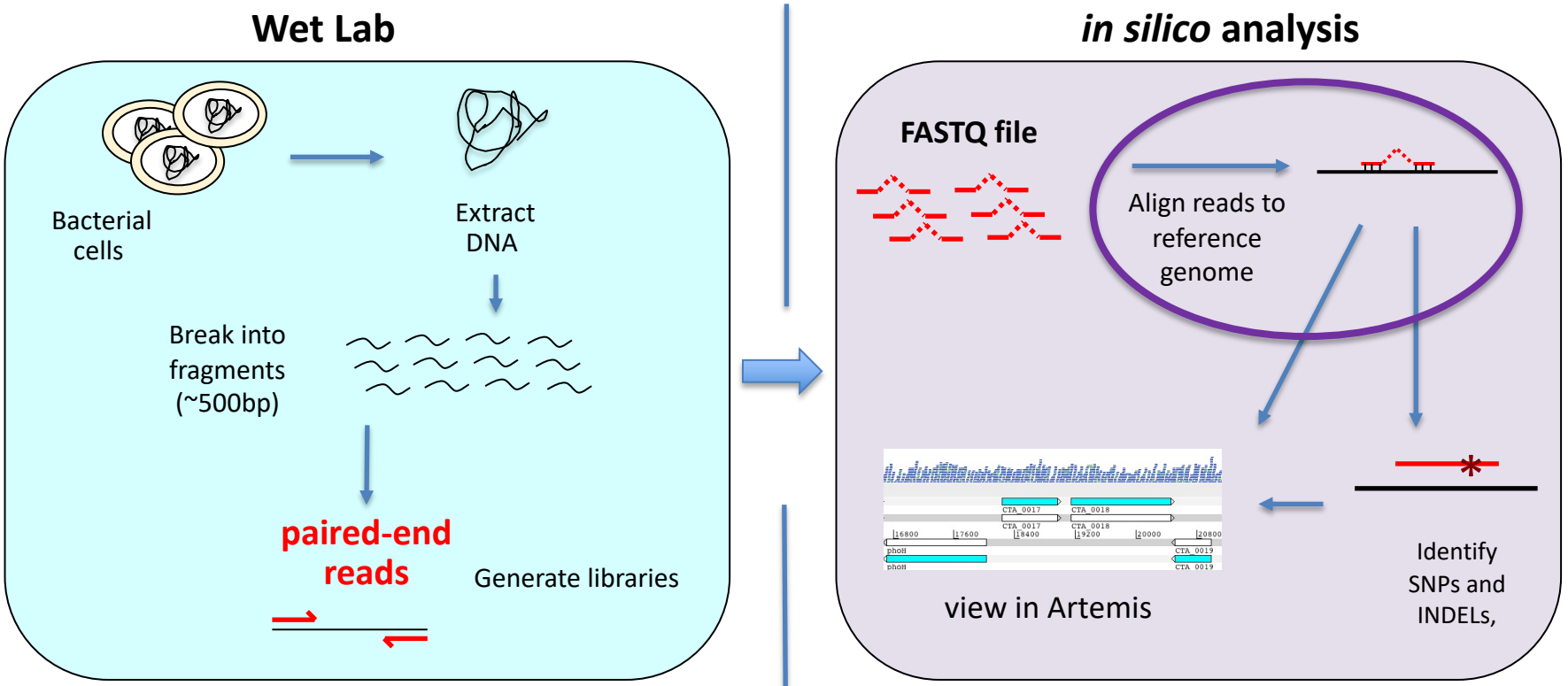
Objectives

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

Workflow:


Paired-end
sequence reads
in **FASTQ file** *
Map against reference *
Call **SNPS** and create
pseudogenomes *
Create **alignment**
Extract variable sites
and convert format
Construct and **interpret**
phylogeny

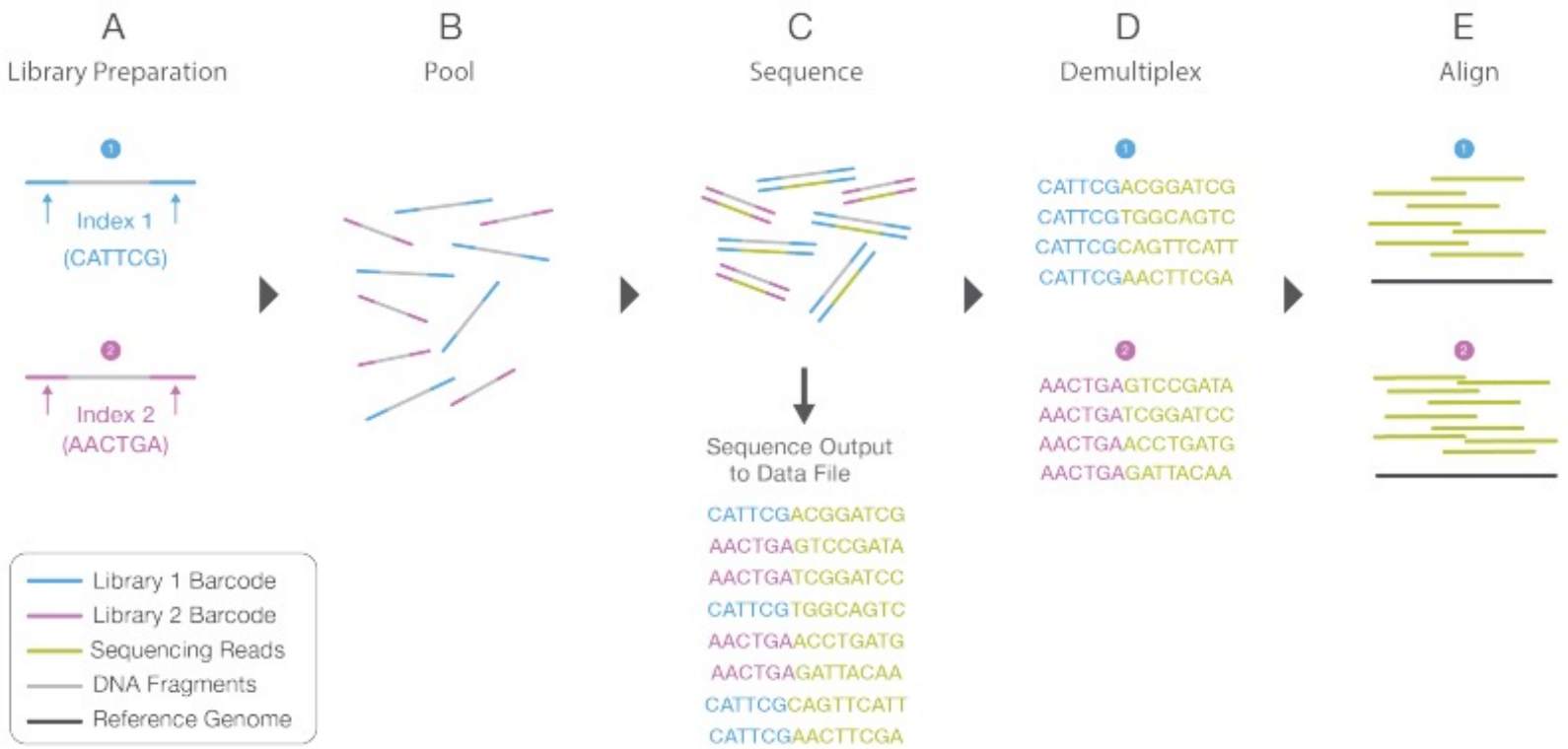
Workflow: generating sequencing reads and *in silico* analysis



DNA extraction and DNA library generation

Mapping is **aligning a sequence to a known reference to determine genetic differences**

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

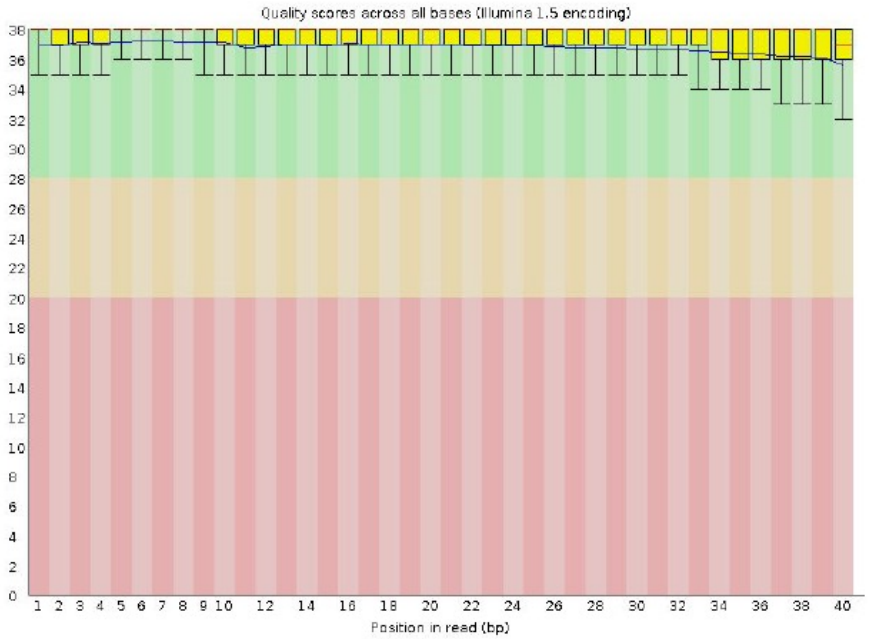
$$Q = -10 \log_{10} P \quad \Longrightarrow \quad P = 10^{\frac{-Q}{10}}$$

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%

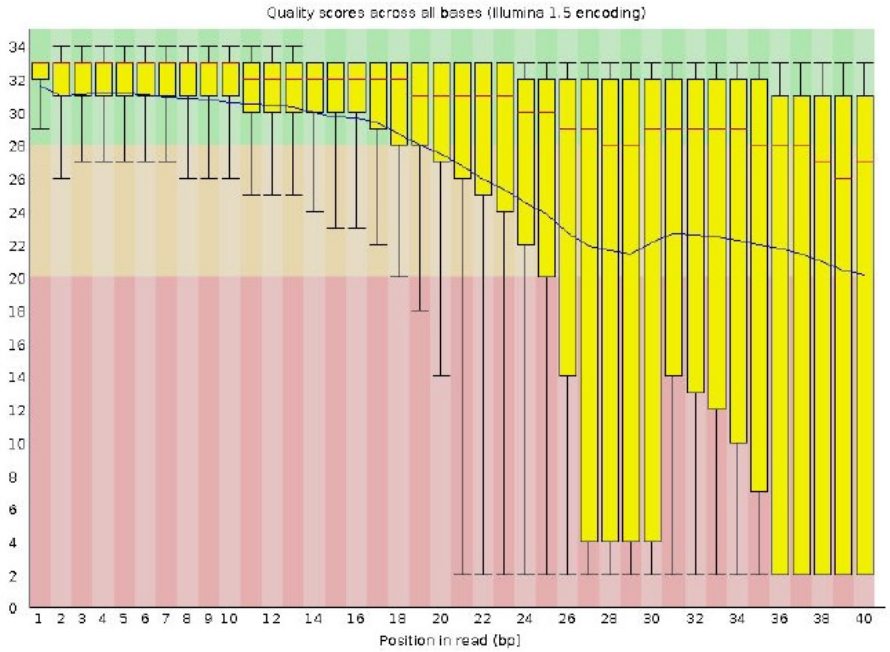
The quality (Q), also called phred score, is the probability (P) that the corresponding basecall is incorrect.

Fastq Quality Check made easy!

Good



Bad



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

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
```

Choose your reference sequence wisely

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

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

wellcome connecting science

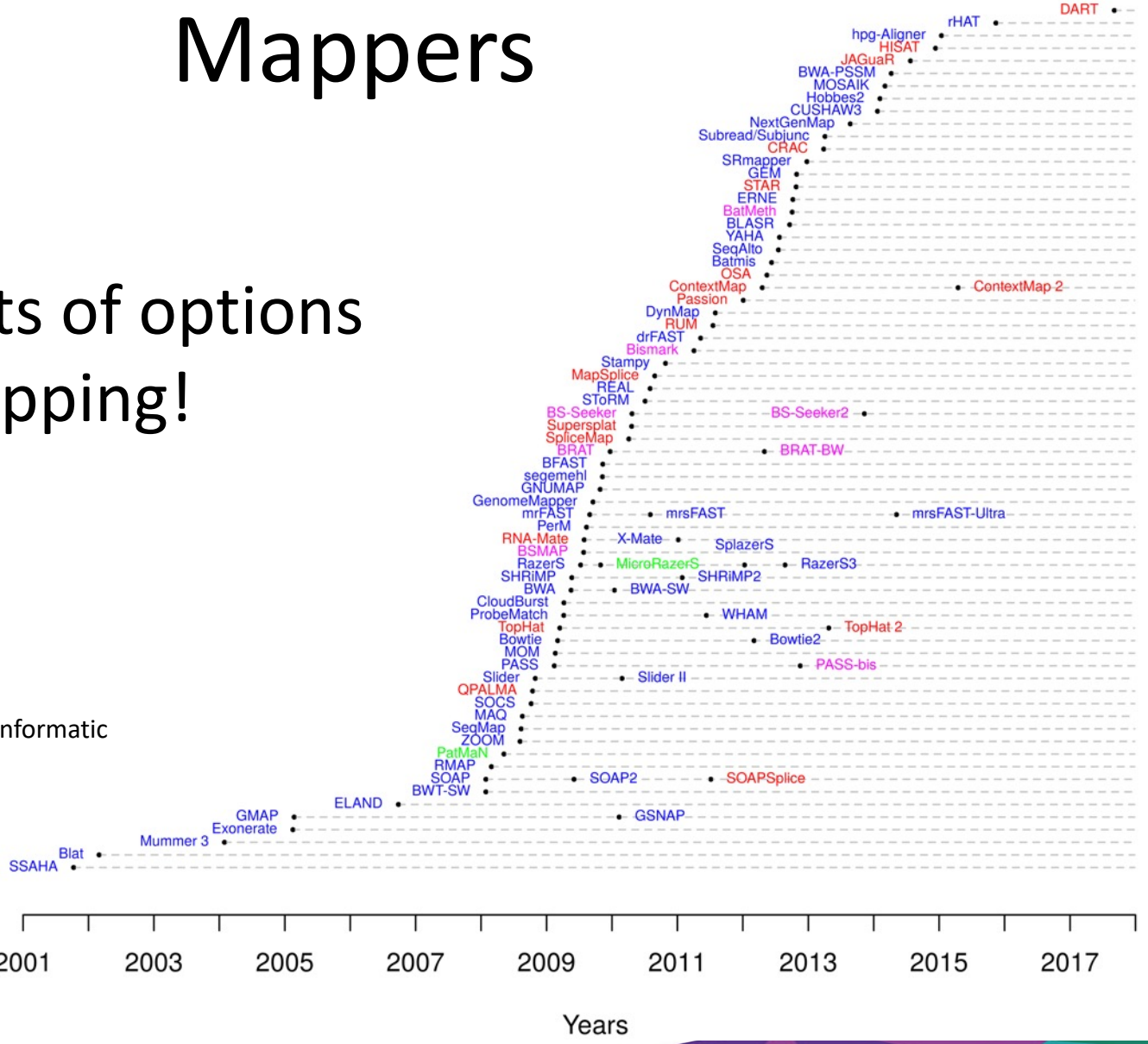
Why do we map reads to a reference?

- Identify variation:
  - Single Nucleotide Polymorphisms (SNPs),
  - insertions and deletions (indels)
  - Copy Number Variants (CNVs) between variants of the same bacteria.
  - Presence / absence of genes (AMR)

Mappers

There are lots of options for mapping!

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



wellcome connecting science

Comparison of different mappers

Mapper	Data	Availability	Version	O.S.	Number Citations	Seq.Plat.	Input	Output	Min. RL	Max. RL	Mismatches	Indels	Gaps	Align.	Reported Alignment	Parallel	QA	PE	Splicing	Index
BatMeth	Bisulfite	OS	1.03	Linux, Unix	34	I	(C)FAST(A/Q)	Native	35	100	5	N	N	B,U	G	N	Y	N	N	Reference
Batmis	DNA	OS	3.0	Linux,Mac	23	I,So	FASTA/Q	SAM			10	0	N	A,U,S		N	Y	Y	N	Reference
BFAST	DNA	OS	0.7.0	Linux,Mac	553	I,So,4, Hel	(C)FAST(A/Q)	SAM TSV		*	Y	Y	Y	B,R,U	G	SM	N	Y	N	Reference
Bismark	Bisulfite	OS	0.7.3	Linux,Mac	887	I	FASTA/Q	SAM	16	10K	Score	Score	N	U		SM	Y	Y	N	
BLASR	DNA	OS	1.4	Linux, Unix		P	FASTA/Q hdf5	SAM TSV	50	100000	0.2	0.2	Y	A,B,R	G L	N	Y	N	De novo	Reference
Blat	DNA	OS	34	Linux,Mac	6252	N	FASTA	TSV BLAST	11	5000K	Score	Score	Y	B	L	N	N	N	De novo	Reference
Bowtie	DNA	OS	0.12.7	Linux,Mac,Windows	11207	I,So,4,Sa,P	(C)FAST(A/Q)	SAM TSV	4	1K	Score	Score	N	A,B,R,S	G L	SM	Y	Y	N	Reference
Bowtie2	DNA	OS	2.0beta5	Linux,Mac,Windows	8586	I,4,Ion	FASTA/Q	SAM TSV	4	5000K	Score	Score	Y	A,B,R,S	G L	SM	Y	Y	N	Reference
BRAT	Bisulfite	OS	1.2.3	Linux	60	I	FASTA/Q	TSV			Y	0	N			N	N	Y	N	Reference
BRAT-BW	Bisulfite	OS	2.0.1	Linux	53	I	FASTA/Q	TSV	32	*	Y	0	N			N	N	Y	N	Reference
BS-Seeker	Bisulfite	OS		Linux,Mac	193	I	FASTA/Q	SAM			3	0	N	U		SM	Y	N	N	
BS-Seeker2	Bisulfite	OS	2.0.0	Linux, Unix, Mac	107	I	FASTA/Q qseq	SAM BAM	10	200	Score	Score	Y	B,U,S	G L	SM	N	Y	N	Reference
BSMAP	Bisulfite	OS	2.73	Linux, Unix, Mac	347	I	FASTA/Q SAM/BAM	SAM BAM Native	20	144	15	1	N	B,R,U	G	SM	N	Y	N	Reference
BWA	DNA	OS	0.6.2	Linux,Mac,Windows	13341	I,So,4,Sa,P	FASTA/Q	SAM	4	200	Y	8	Y	R,S	G	SM	Y	Y	N	Reference
BWA-PSSM	DNA	OS	0.5.11	Linux	26	I,Hel	FASTQ/Q PSSM	SAM BAM	4	200	30	10	N	B,R,U,S	G	SM	Y	Y	N	Reference
BWA-SW	DNA	OS	0.6.2	Linux,Mac,Windows	3494	I,4,Sa,Hel,Ion,P	FASTA/Q	SAM	4	1000K	0.1	0.1	Y	R,S	L	SM	Y	N	N	Both
BWT-SW	DNA	OS	20070916	Linux	133	N	FASTA	TSV		1K	Score	Score	Y	A		N	N	N	N	Reference
CLC Mapper	DNA	Com	4			I,4,So,Sa,Ion,P,Hel	FASTA/Q	SAM BAM								N		Y		
CloudBurst	DNA	OS	1.1	Linux,Mac,Windows	650	N	FASTA	TSV		1K	Y	Y	Y	A,B	G	Cloud	N	N	N	Reads
ContextMap	RNA	OS	2.2	Linux, Unix, Mac	22	I,4,So,Sa,Ion,P,Hel	FASTA/Q	SAM	1	5000	20	10	Y	A,B	G	SM	N	Y	Lib or de novo	Reference
ContextMap 2	RNA	OS	2	Windows, Linux, Unix, Mac	9	I,4,Sa,Ion,P,Hel	FASTA/Q Illumina	SAM BED	20	5000	0.1	10	Y	B	L	No	N	Y	Lib or de novo	Reference
CRAC	RNA	OS	2.0.0	Linux, Unix, Mac	41	I,4,Ion,P	(C)FAST(A/Q) RAW	SAM BAM	50	*	score	score	Y	A,B,U,S	G	SM	N	Y	De novo	Both
CUSHAW3	DNA	OS	v3.0.3	Linux	33	I,So,4,Ion,P	FASTA/Q	SAM	16	4096	score	score	Y	A,B,R,U,S	G L	SM	Y	Y	N	Reference
DART	RNA	OS	1.2.4	Linux	0	I	FASTA/Q	SAM	20		Y	Y	Y	A U	G	SM	N	Y	De novo	Reads
drFAST	DNA	OS	1.0.0.0	Linux, Unix	23	So	CFAST(A/Q)	SAM DIVET	25	200	Score	N	N	A,B	G	N	N	Y	N	Reference
DynMap	DNA	OS	0.0.20	Linux	2	N	FASTA	TSV	18	8K	5	0	N	B	L	N	N	N	N	Reads
ELAND	DNA	Com	1	Linux, Unix, Mac	25	I	FASTA		15	150	2	Score	N	B,S	G	N	Y	Y	N	
ERNE	DNA	OS	1	Windows, Linux, Unix, Mac	14	I	FASTA/Q Illumina	SAM BAM Native	15	600	0.1	5	Y	A,R,U,S	G	SM/DM	N	Y	De novo	Reference
Exonerate	DNA	OS	2.2	Linux,Mac	918	N	FASTA	TSV	20	*	Score	Score	Y	B,S	G L	N	N	N	De novo	Reference
GEM	DNA	Bin	1.x	Linux,Mac	260	I, So	FASTA/Q	SAM Counts		*	Y	Y	Y	A,S	G	SM	Y	Y	Lib and de novo	Reference
GenomeMapper	DNA	OS	0.4.3	Linux,Mac	144	I	FASTA/Q	BED TSV	12	2K	10	10	Y	A,B,R	G	SM	N	N	N	Reference
GMAP	DNA	OS	2012-04-27	Linux,Unix,Mac,Windows	868	I,4,Sa,Hel,Ion,P	FASTA/Q	SAM GFF Native	8	*	Y	Y	Y	B	G L	SM	N	N	De novo	Reference
GNUMAP	DNA	OS	3.0.2	Linux,Mac	80	I	FASTA/Q Illumina	SAM TSV	16	1K	Score	Score	Y	B	G	SM/DM	Y	N	N	Reference
GSNAP	DNA	OS	2012-04-27	Linux,Unix,Mac,Windows	1156	I,4,Sa,Hel,Ion,P	FASTA/Q	SAM Native	17	250	Y	Y	Y	A,B,U,S	G L	SM	N	Y	Lib and de novo	Reference
HISAT	RNA	OS	1	Windows, Linux, Unix, Mac	480	I	FASTA/Q	SAM	50	*	0.1	0.1	N	A,B,R,U,S	G	SM	Y	Y	Lib or de novo	Reference
HISAT2	DNA	OS	2	Windows, Linux, Unix, Mac		I	FASTA/Q	SAM	50		score	score	N	B	G	SM	Y	Y	Lib or de novo	Reference
Hobbes2	DNA	OS	2.1	Linux	13	N	FASTA/Q	SAM	22	200	0.08	0.08	N	A,U,S	G	N	N	Y	No	Reference
hpg-Aligner	DNA	OS	v2.1.0	Linux	1	I,So,4,Sa,Hel,Ion,P	FASTQ	SAM, BAM	10	2000	0.3	0.3	Yes	A,B	G	N	Y	Y	Lib and de novo	Reference
JAGuaR	RNA	OS	2.1	Linux, Unix	15	I	FASTQ	SAM BAM	50	300			Y	B	G	N	Y	Y	Lib	Reference
MapReads	DNA	OS	2.4.1	Linux,Mac,Windows	0	So	FASTA/Q	TSV	10	120	Score	0	N	S		N	Y	N	N	Reference
MapSplice	RNA	OS	1.15.2	Linux	610	I	FASTA/Q	SAM BED			3		Y	B		SM	N	Y	De novo	
MAQ	DNA	OS	0.7.1	Linux,Mac	2592	I,So	(C)FAST(A/Q)	TSV	8	63	Y	Y	N			N	Y	Y	N	Reads
Masai	DNA	OS	0.4	Windows, Linux, Mac	1	I, Ion	FASTA/Q	SAM	20	32678	32	32	N	A, B, U	G	N	N	Y	N	Both
MicroRazerS	miRNA	OS	0.1	Linux	40	N	FASTA	SAM TSV	10	*	Score	0	N	S	G	N	N	N	N	Reference
MIRA	DNA	OS	3	Linux,Unix		I,4,Sa,Ion,P	FASTA/Q PHD EXP	SAM GFF Counts CAF	25	19000	Score	Score	Y	B,R	L	SM	Y	Y	N	Both
MOM	DNA	Bin	0.6	Linux,Mac,Windows	48	I,4	FASTA	TSV			Y	0	N	A	L	SM	N	Y	N	Either
MOSAIK	DNA	OS	2.1	Linux,Unix,Mac,Windows	174	I,So,4,Sa,Hel,Ion,P	(C)FAST(A/Q)	BAM	15	1000	Y	Y	Y	A,B	G	SM	Y	Y	N	Reference
mrFAST	DNA	OS	2.5.0.1	Linux,Unix	602	I	FASTA/Q	SAM DIVET	25	1000	Score	4	N	A,B	G	N	Y	Y	N	Reference
mrsFAST	DNA	OS	2.4.0.4	Linux, Unix	229	I	FASTA/Q	SAM DIVET	25	100	Score	N	N	A	G	N	Y	Y	N	Reference
mrsFAST-Ultra	DNA	OS	3.3.1	Linux, Mac	28	I	FASTA/Q	SAM DIVET	8	500	Score	N	N	A,B,S	G	SM	Y	Y	N	Reference
Mummer 3	DNA	OS	3.23	Linux,Mac	2446	N	FASTA	TSV	10	*	Y	Y	Y	A,B	G	N	N	N	N	Reference
NextGenMap	DNA	OS	0.4.6	Linux	82	I,4,Ion	(C)FAST(A/Q),SAM,BAM	SAM BAM	13	1000	Score	Score	N	R,S	G L	SM	N	Y	N	Reference
Novoalign(CS)	DNA	Bin	V2.08.03	Linux	0	I,So,4,Hel,Ion	(C)FAST(A/Q) Illumina	SAM Native	1	250	Y	Y	Y	A,B,R,U	G	SM/DM	Y	Y	N	Lib Reference
OSA	RNA	Bin	1.0.<	Windows, Linux, Unix, Mac	54	I,4,Ion	FASTA/Q	SAM BAM	15	8000	*	*	Y	A,B,U	G	SM	Y	Y	Lib and de novo	Reference
PASS	DNA	Bin	1.62	Linux,Mac,Windows	142	I,So,4	(C)FAST(A/Q)	SAM GFF3 BLAST	23	1K	Y	Y	Y	A,B	G	SM	Y	Y	De novo	Reference
PASS-bis	Bisulfite	OS	2.01	Linux	14	I,So,4,Sa	FASTA/Q	SAM GFF Counts	14	2000	Score	N	N	A, B, U, S	G	SM	Y	Y	N	Reference
Passion	RNA	OS	1.2.0	Linux,Unix	28	I,4,Sa,P	FASTA/Q	BED			Y	Y	Y	U		SM	Y	Y	De novo	
PatMaN	miRNA	OS	1.2.2	Linux,Mac	140	N	FASTA	TSV	1	*	Y	Y	N	A	G	N	N	N	N	Reads
PerM	DNA	OS	0.4.0	Linux,Unix,Mac,Windows	113	I,So	(C)FAST(A/Q)	SAM TSV	20	128	9	0	Y	A,U	G	DM	Y	Y	N	Reference
ProbeMatch	DNA	OS		Linux,Mac	4	I,4,Sa	FASTA	ELAND	36	50	3	Y	N	A,B		N	N	N	N	Reference
QPALMA	RNA	OS	0.9.2	Linux,Mac	169	I,4	Specific	TSV			Y	Y	Y	B	L	N	Y	N	Lib and de novo	
RazerS	DNA	OS	1.2	Linux,Mac,Windows	165	I,4	FASTQ	TSV ELAND	11	*	Score	Score	Y	A,B,U,S	G	N	N	Y	N	Reference
RazerS3	DNA	OS	3.1	Windows, Linux, Mac	81	I	FASTA/Q	SAM TSV GFF	11	*	0.5	Y	N	A,B,U,S	G	SM	N	Y	N	Reads
REAL	DNA	OS	0.0.28	Linux	32	I	FASTA/Q	TSV	4	*	Score	N	N	B,U	G	SM	Y	N	N	Reference

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

Good general aligners

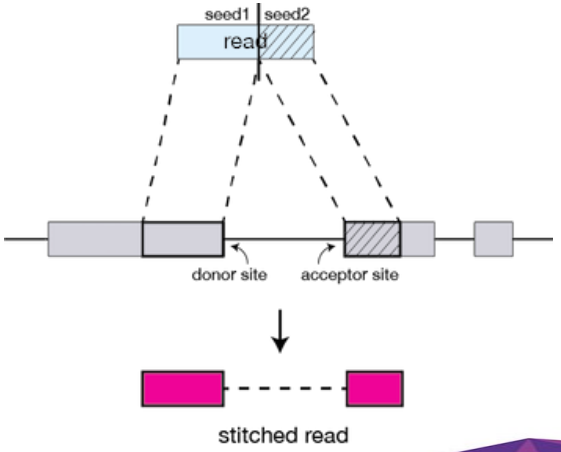
★ bwa
bowtie2
minimap2

→ Fast, sensitive and easy to use!

Splice-aware aligners for RNA-seq

STAR

★ HISAT2



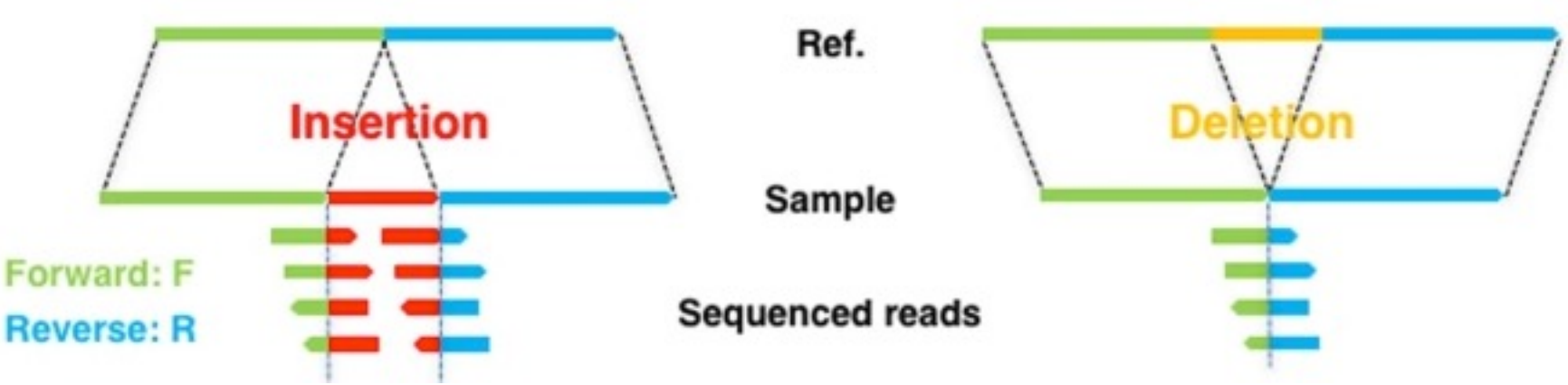
Why do we map to a reference?

- Identify variation:
  - Single Nucleotide Polymorphisms (SNPs),
  - insertions and deletions (indels)
  - Copy Number Variants (CNVs) between variants of the same bacteria.
  - Presence / absence of genes (AMR)

Single Nucleotide Polymorphisms (SNPs)

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

INDELS



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

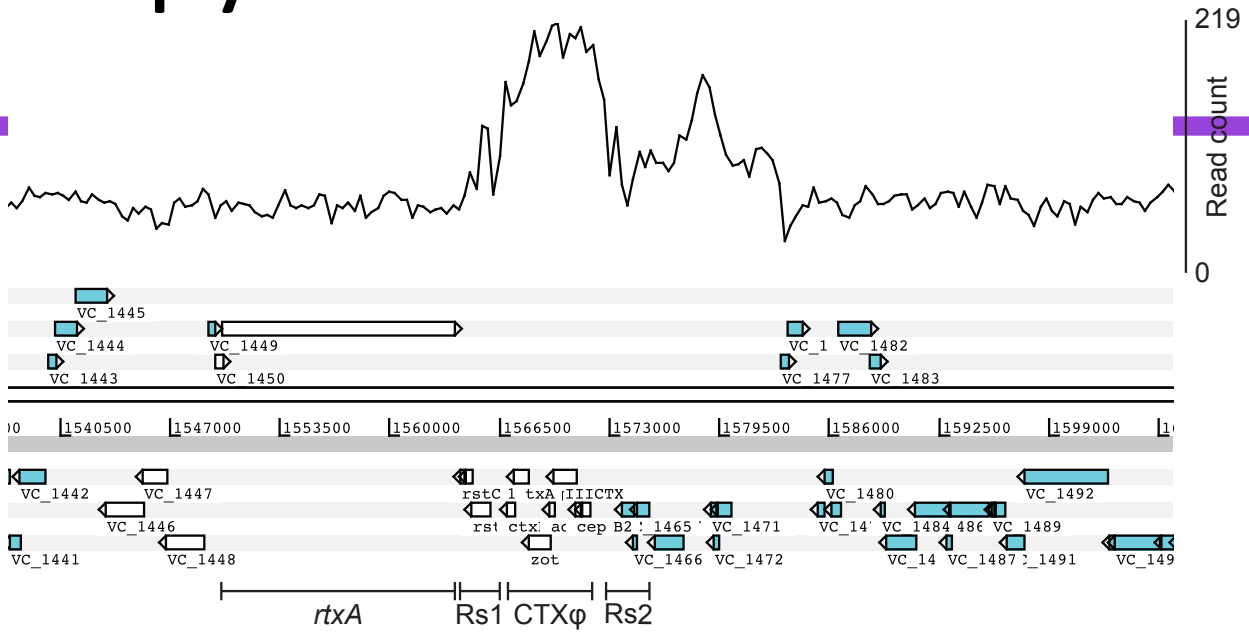
Visualize in Artemis

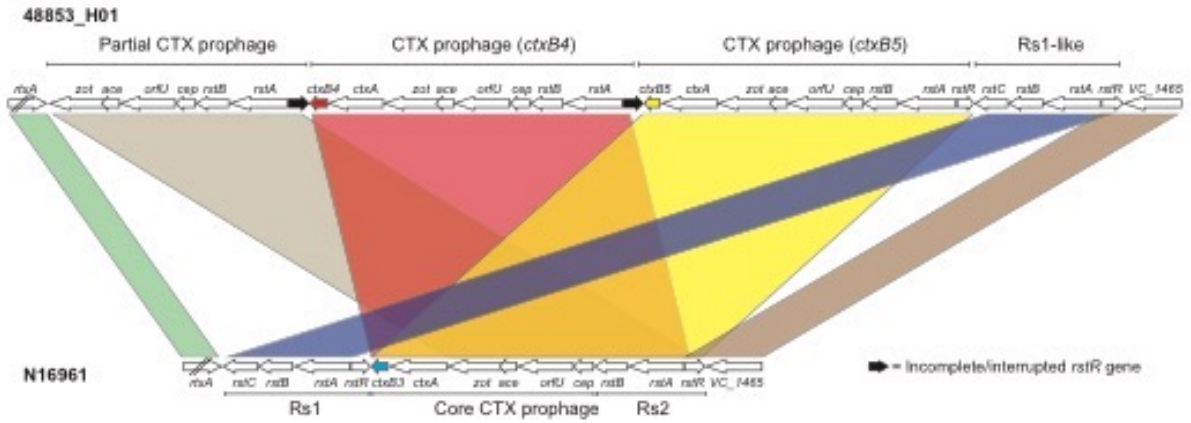
wellcome connecting science

Why do we map to a reference?

- Identify variation:
  - Single Nucleotide Polymorphisms (SNPs),
  - insertions and deletions (indels)
  - Copy Number Variants (CNVs) between variants of the same bacteria.
  - Presence / absence of genes (AMR)

Copy number variation



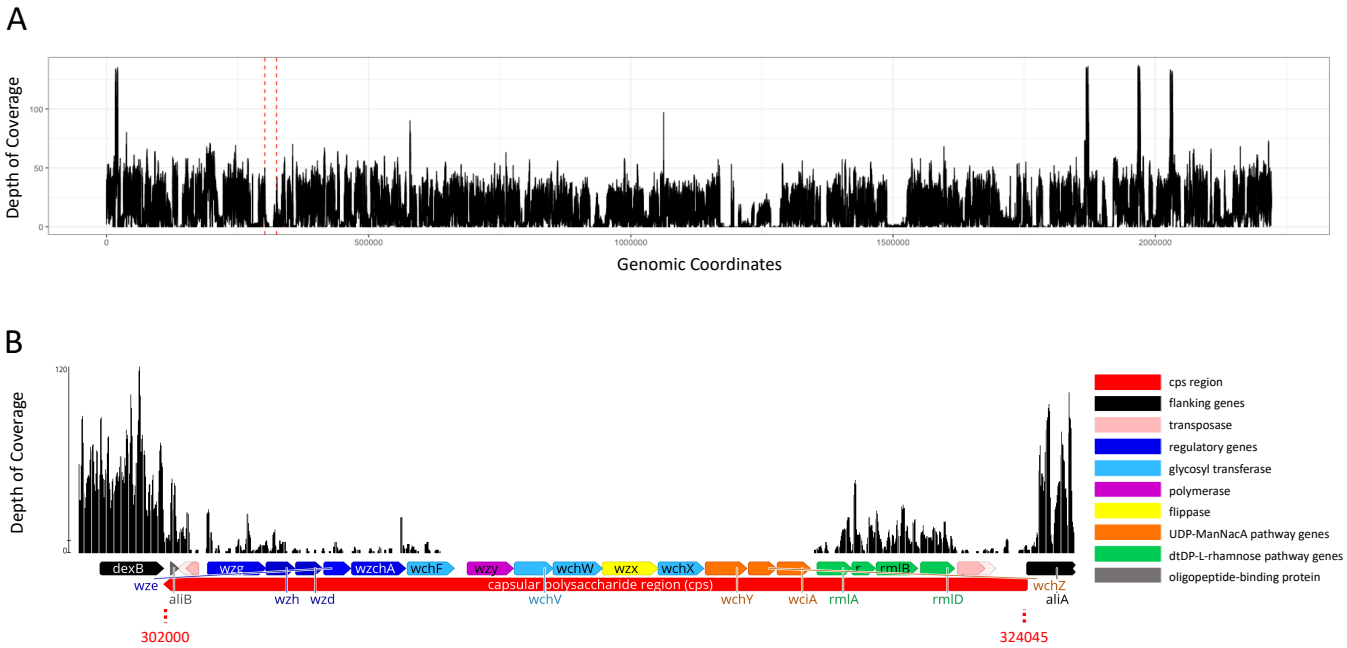


wellcome connecting science

Gene presence/absence: AMR

- Absence/Deletions is easier to spot
- *To identify insertions is a little tricky.

Deletion:

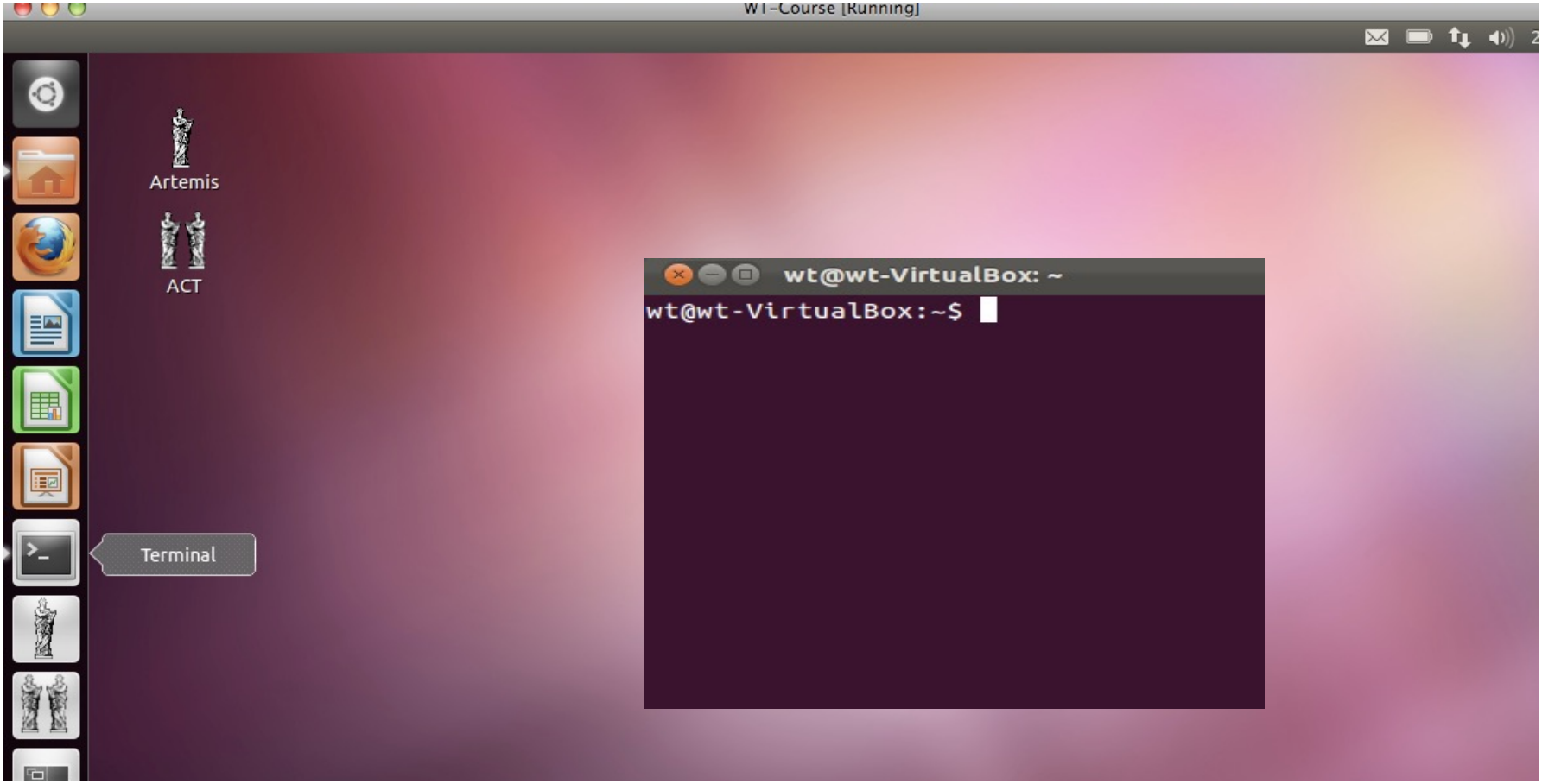


wellcome connecting science

Gene insertions/novel genes

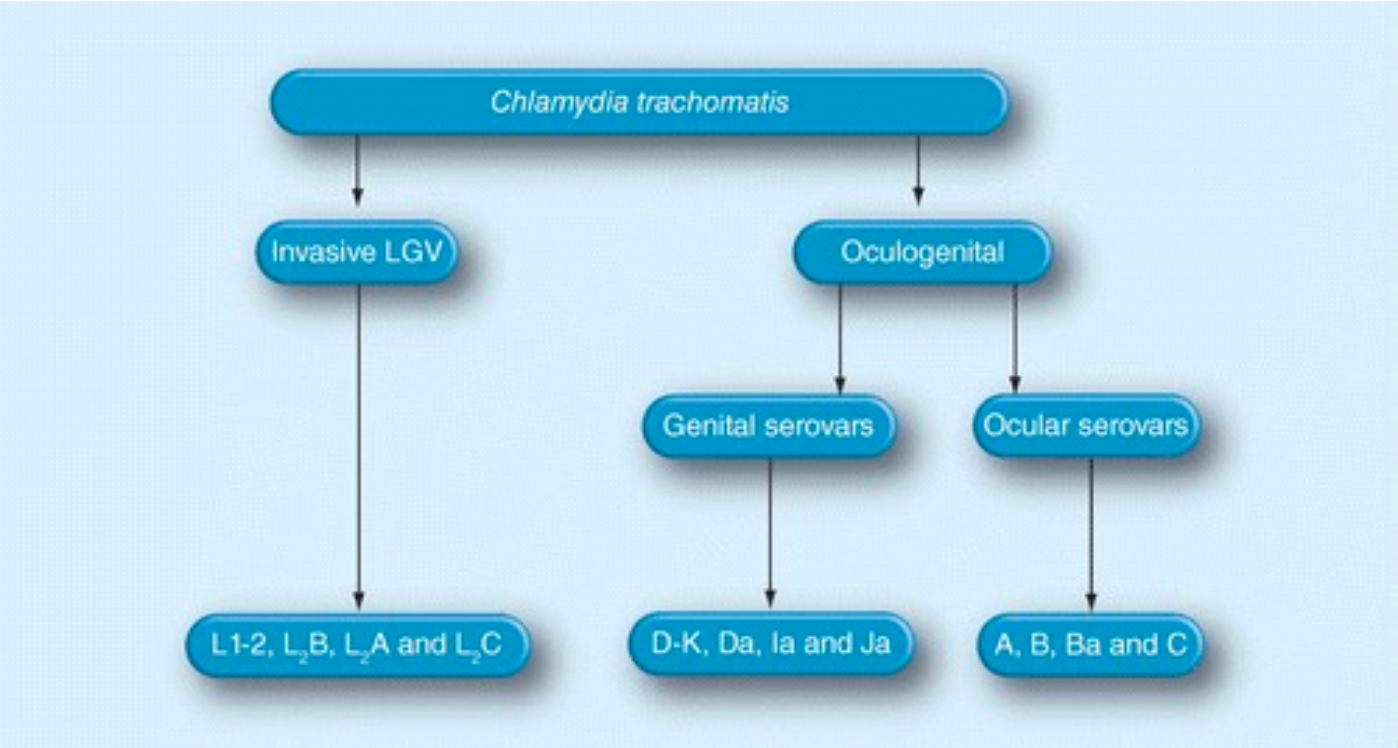
- In this instance you must investigate:
  - Metadata (phenotype)
  - Map to a different reference
  - If AMR/Virulence – map to a database
  - Assembly

The exercise:



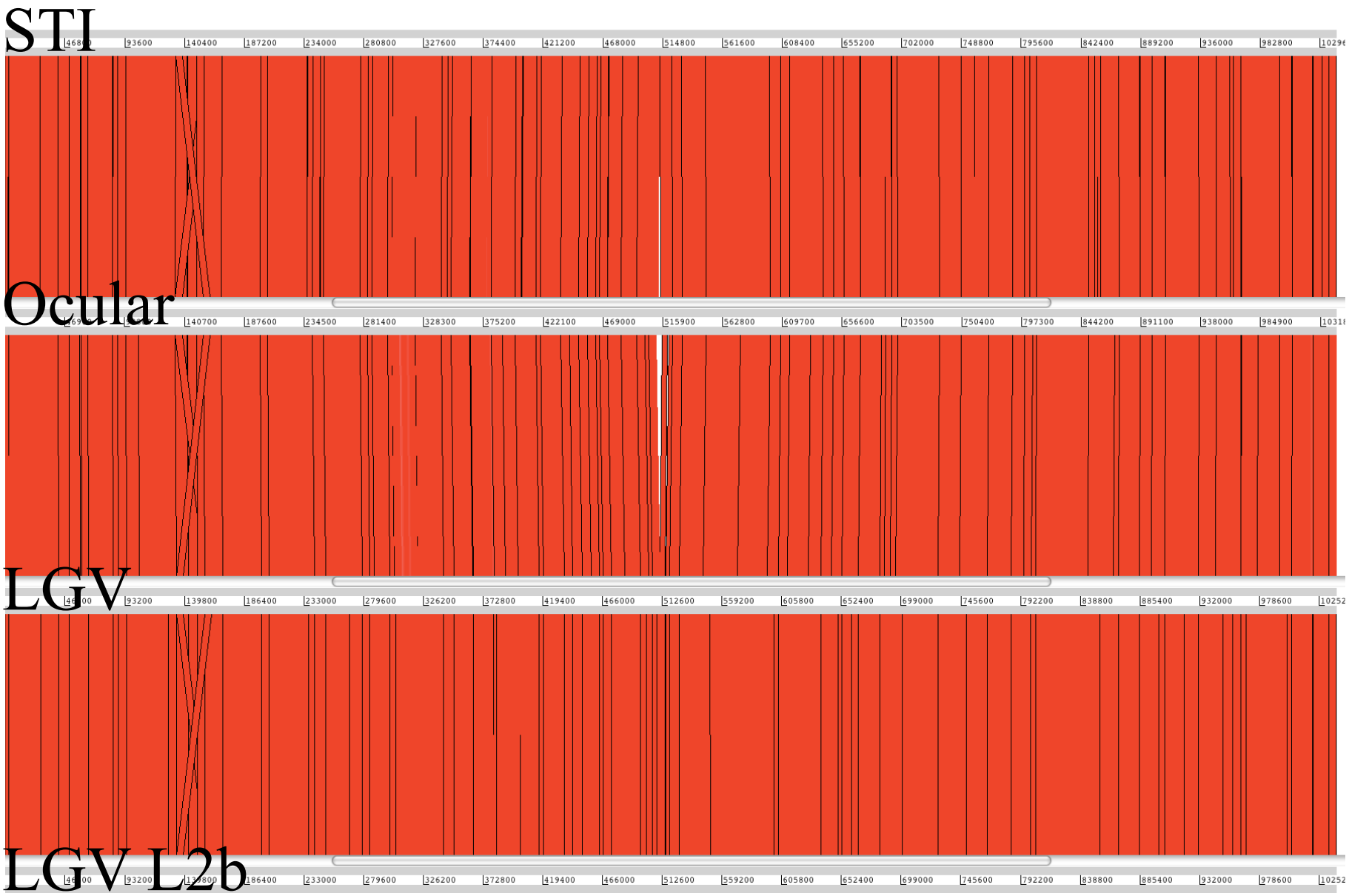
Chlamydia trachomatis

Classification by tissue tropism

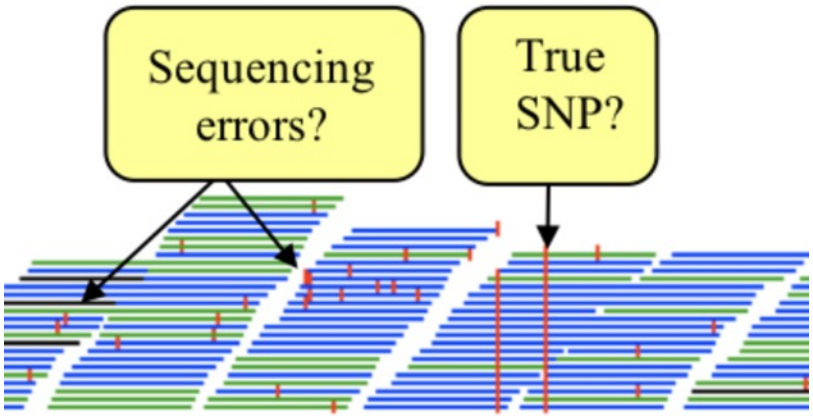


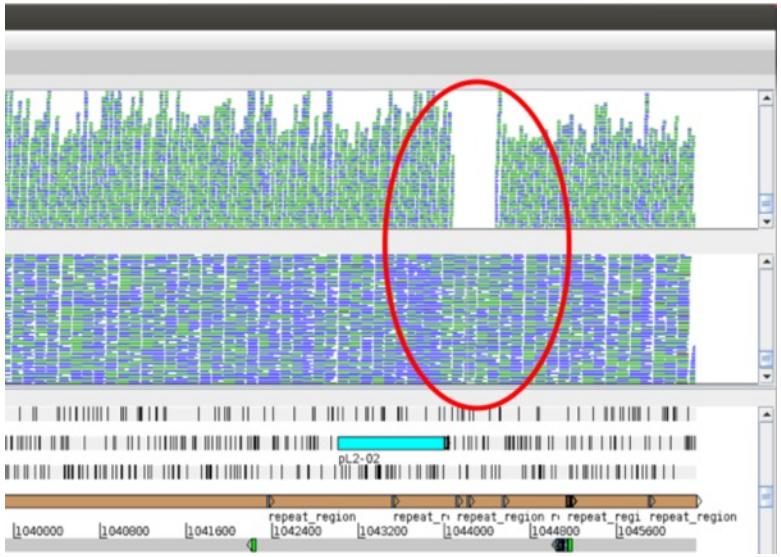
https://www.futuremedicine.com/doi/full/10.2217/fmb.13.80

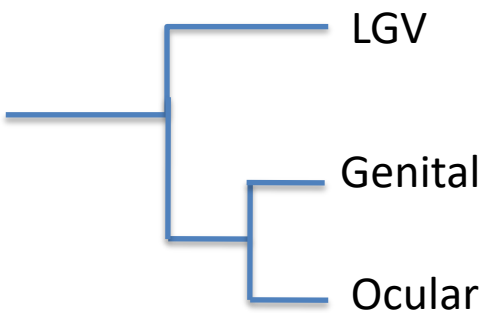
Whole Genome alignments. How do you distinguish between the strains?



SNPs and presence/absence of genes

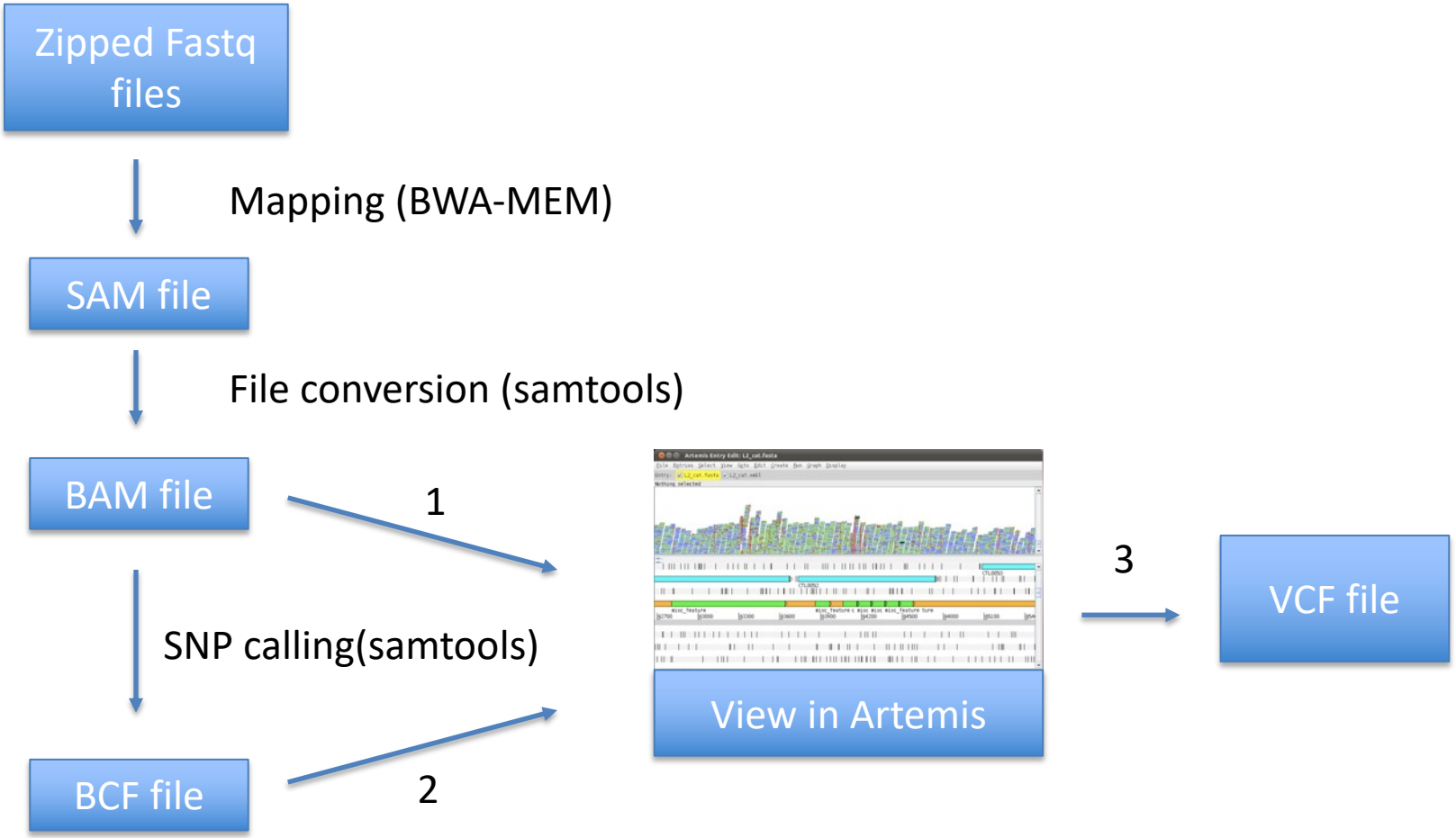






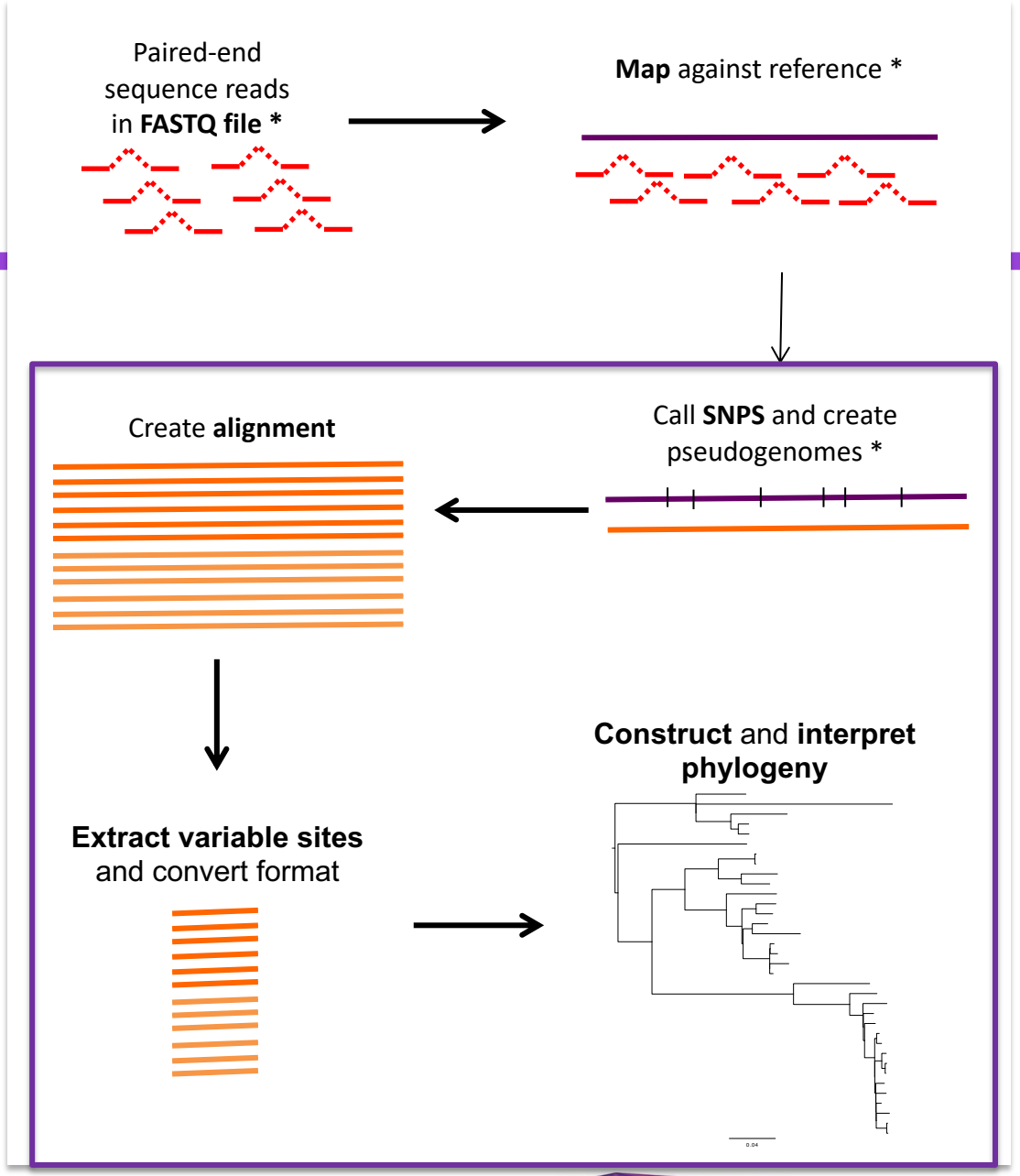
wellcome connecting science

Module 3: Mapping sequence reads workflow



Wrap-up

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



Additional resources

- Illumina sequencing platforms:

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

- Illumina sequencing by synthesis:

https://www.youtube.com/watch?v=fCd6B5HRaZ8

- IGV:

https://software.broadinstitute.org/software/igv/