Reference alignment of reads

Richard Orton

MRC-University of Glasgow Centre for Virus Research

March 2025

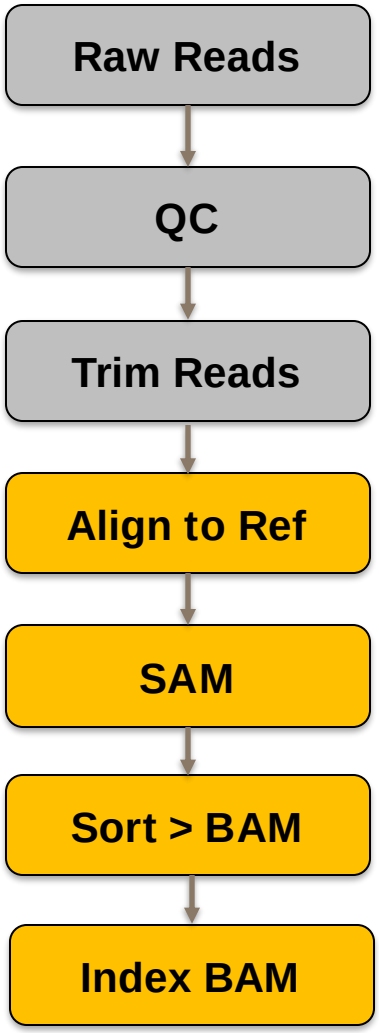
Richard.Orton@glasgow.ac.uk

VIA VERITAS VITA

The way and the truth and the life

Here is the bird that never flew
Here is the tree that never grew
Here is the bell that never rang
Here is the fish that never swam

Previously …



- Previous session we learnt about FASTQ reads and read cleaning/trimming
- Task now is to align these reads to a selected reference sequence

Overview

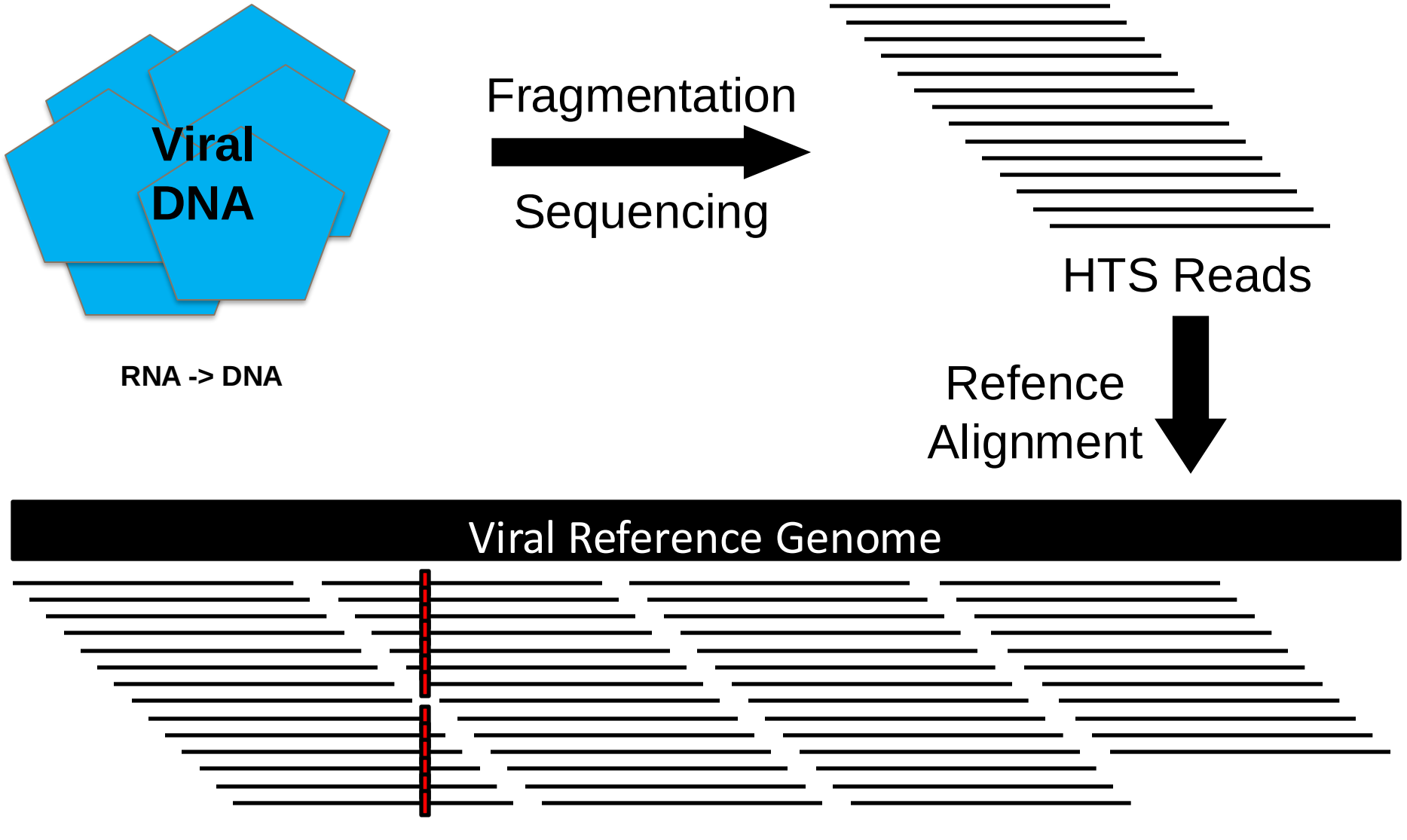
- **What is reference alignment?**
  - How does it work?
- **What tools can you use?**
- **What do the results look like?**
  - Basic statistics
  - Coverage plots
- **Reference alignment practical**
  - Learn the basic steps of refence alignment, SAM/BAM conversion, calculating basic mapping statistics and coverage plots.

Reference alignment

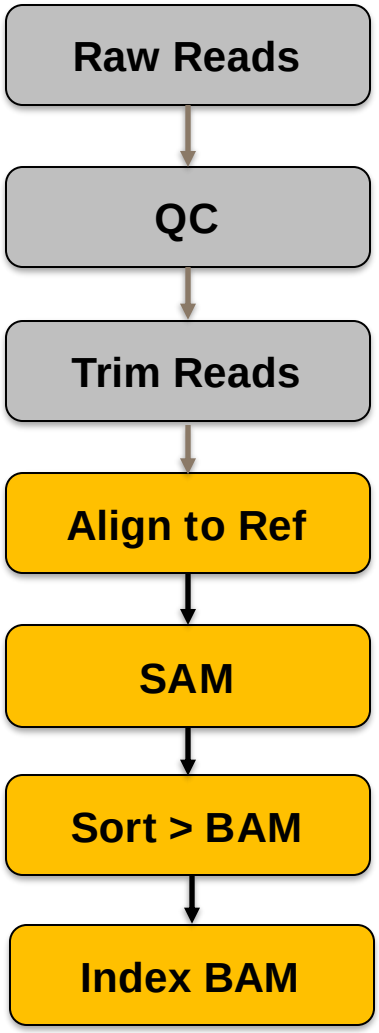
- **Reference alignment:** want to know the **exact position** on the genome a read originates
  - And the **base-to-base** correspondence (to extract mutations, indels)

- **Reference assembly:** assemble reads back together to form a genome
  - Assemble from scratch – *de novo* assembly – using read overlaps, kmers

Aligning reads to a reference genome



Ref alignment basic steps




```
trim_galore -q 25 --length 50 --paired R1.fq R2.fq
```

```
bwa index ref.fasta
bwa mem -r ref.fasta R1_qc.fq R2_qc.fq > my.sam
```

```
samtools sort my.sam -o my.bam
```

```
samtools index my.bam
```

Aligning reads to a reference - needs

- **Need Reads**
  - Single or paired, short or long
  - Typically pre-trimmed & filtered
  - But you can use your raw read files

- **Need a Reference**
  - A suitable reference
  - [More on this later]

- Trimmed reads were aligned to the HCV reference genome (GenBank accession NC_038882) with BWA {Li et, 2009}.

Be careful – aligners tend not to complain

- **Sample**
  - Ebola virus sample from a human patient
- **Reads**
  - Reads were adapter trimmed and quality filtered using trim_galore (quality 25, length 50).
- **Reference**
  - Reads were aligned to the HCV reference genome (GenBank accession NC_038882
- **Result – SAM file of all the reads aligned to the reference**
  - **No errors**
  - **Number of mapped reads (0), coverage statistics (0 cov)**

Unmapped reads

- Reads that could not be aligned to the reference sequence are marked as unmapped
- What are these reads?

Unmapped reads

- Reads that could not be aligned to the reference sequence are marked as unmapped

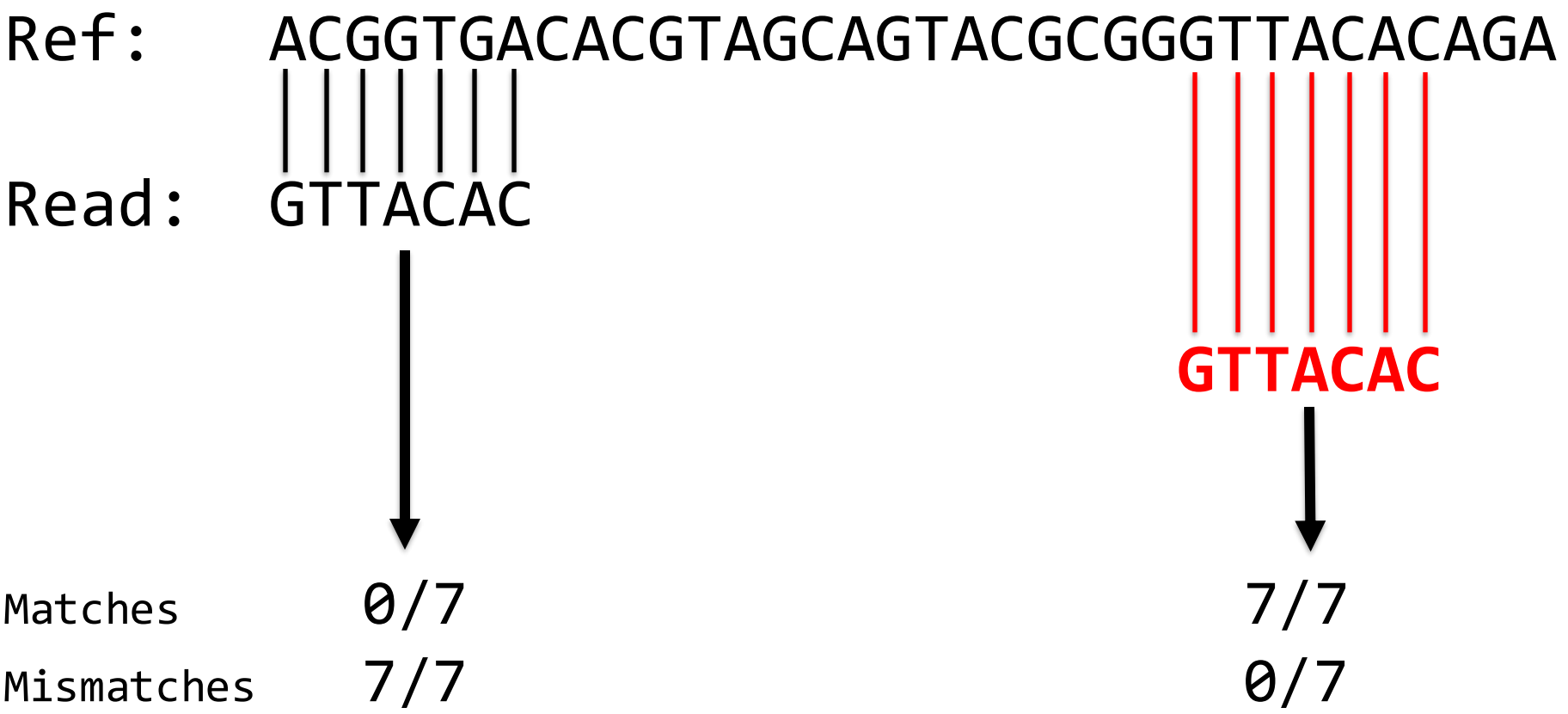
- What are these reads?
  - Host
  - Bacteria, Parasites,
  - Other viruses
  - Random ”low complexity” sequences

- This will be missed as we are “targeting” a specific reference sequence to align against
  - Possible solution: metagenomics

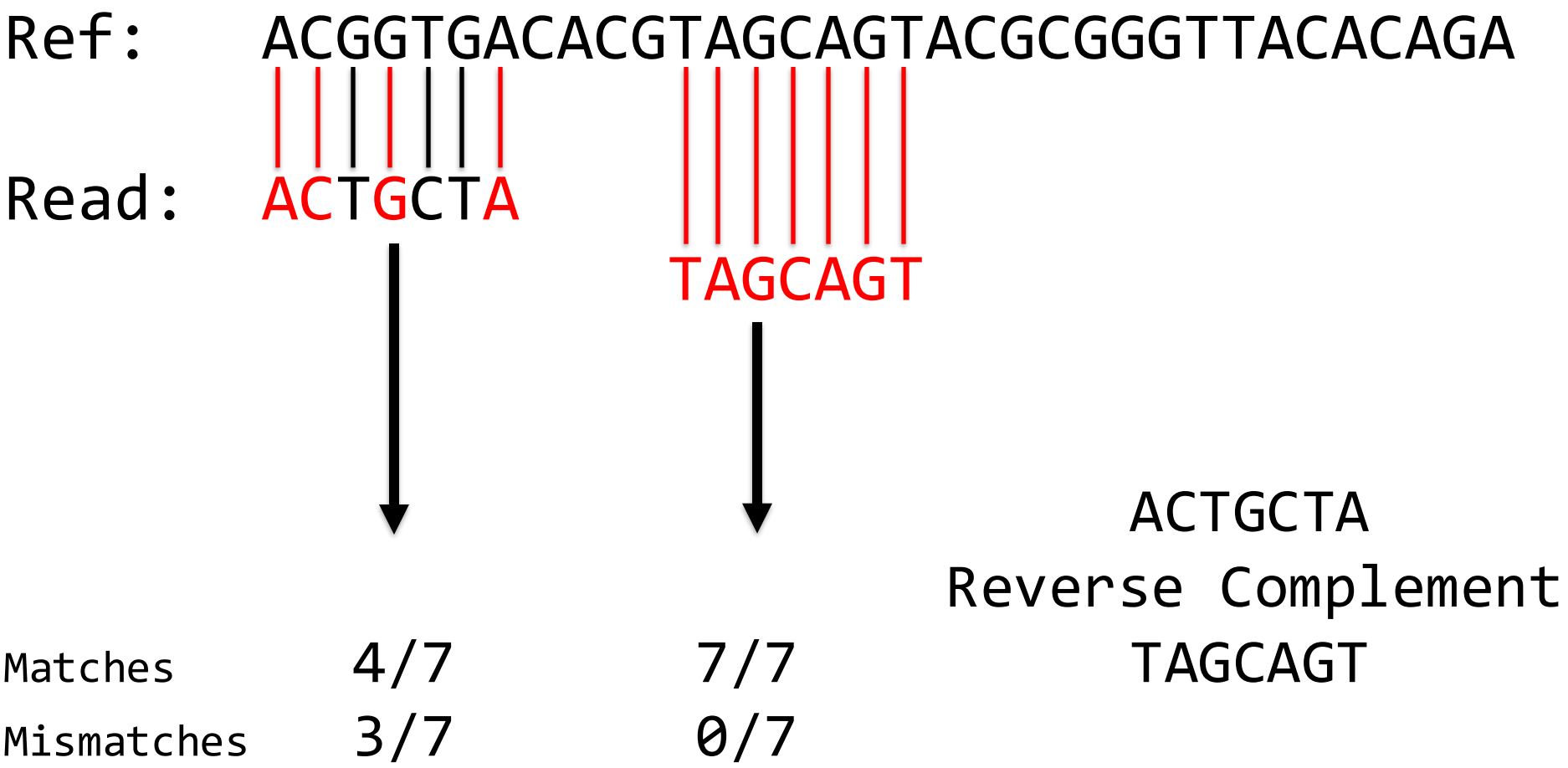
Aligning reads to a references

```
Ref:    ACGGTGACACGTAGCAGTACGCGGGTTACACAGA

                                Read:  GTTACAC
```

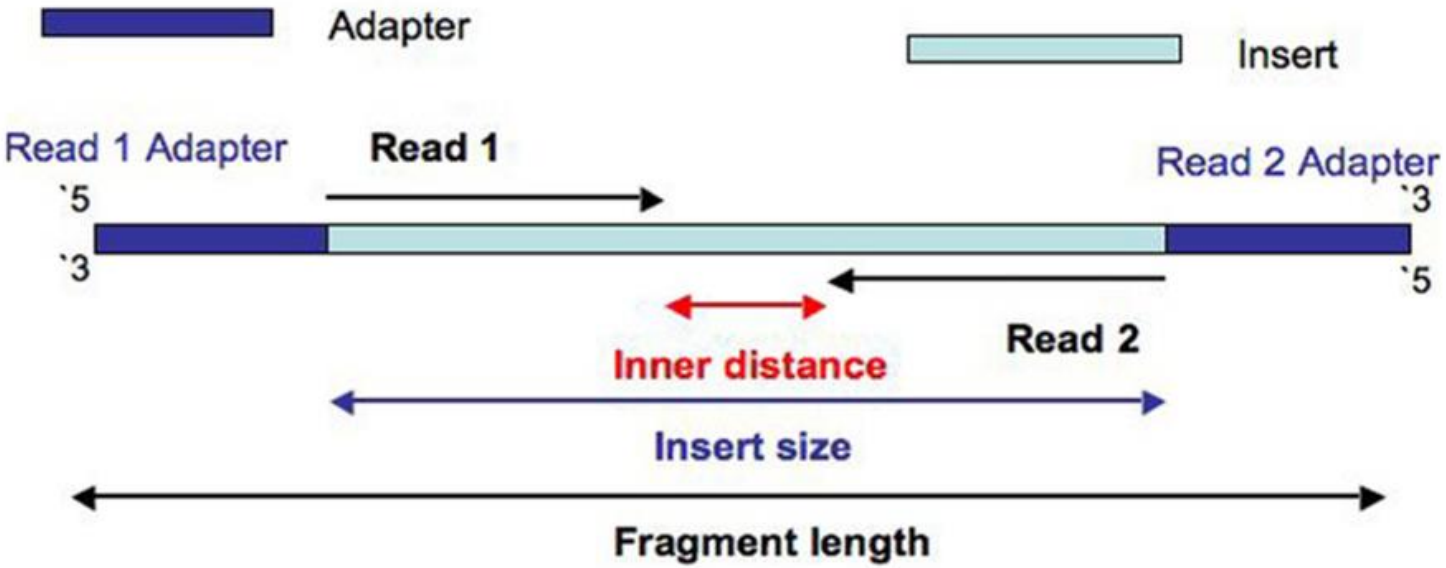
Aligning reads to a references



Aligners check the reverse complement

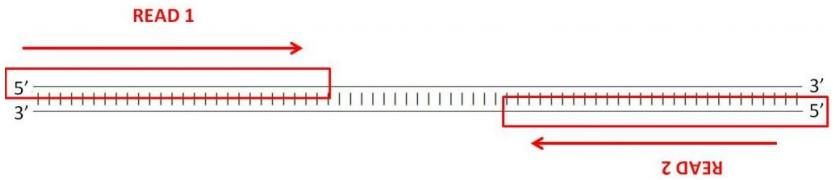


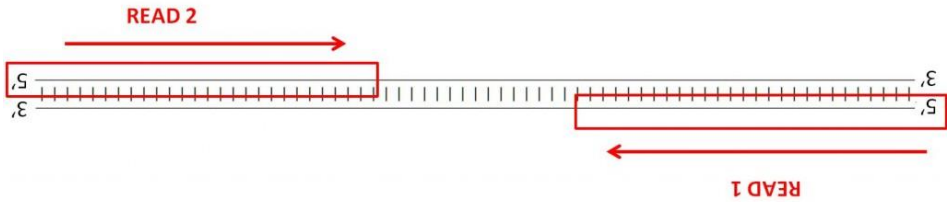
Paired end … Insert Size



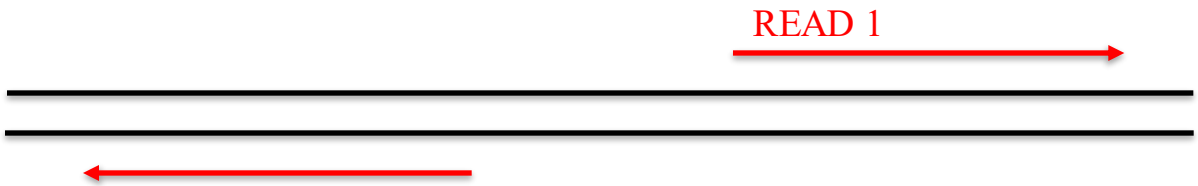
- Typically, the two reads do not overlap, but they can if the reads are long and fragments are short [redundant data, but can be used to correct errors]
- 500bp fragments + 2 x 300bp reads = 100bp overlap
- Turner 2014, Frontiers in Genetics

Concordance & Discordance – paired reads



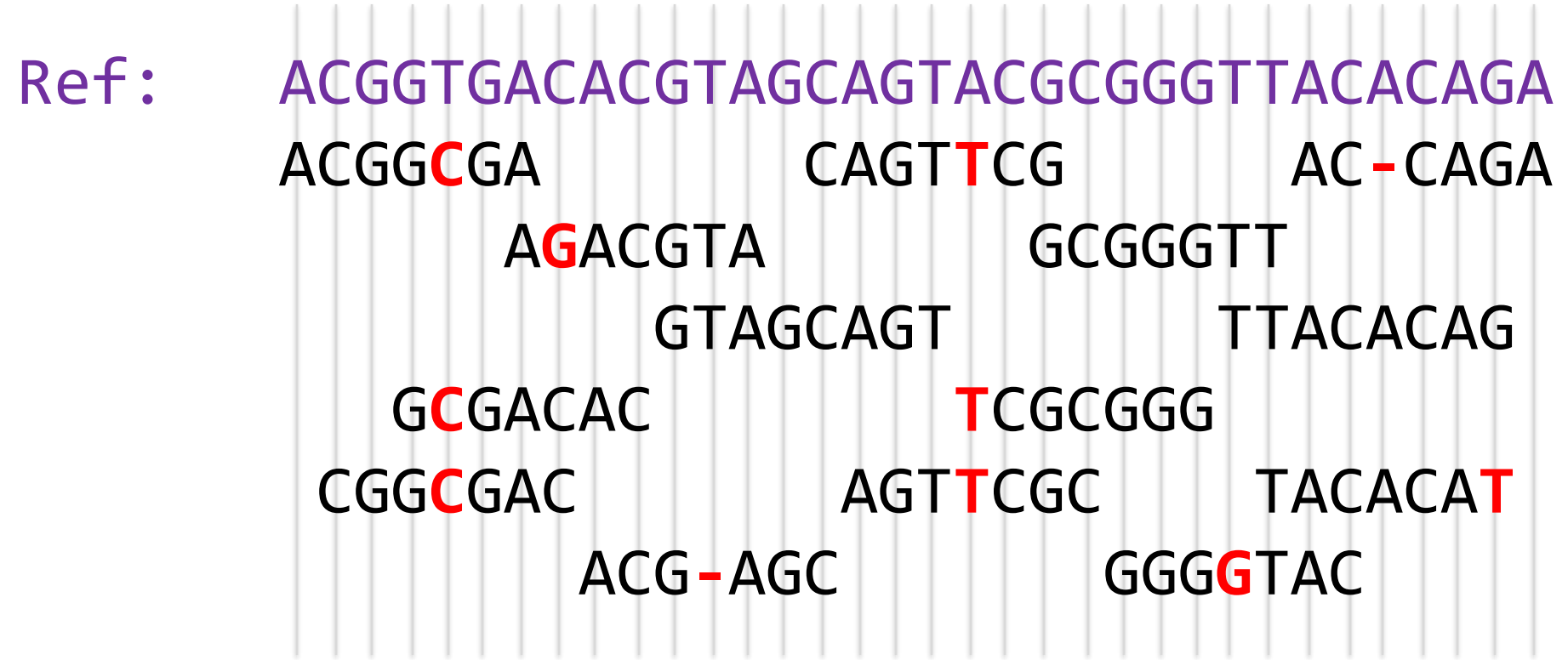


Concordant: consistent orientation of read pairs with respect to reference, have insert size within the expected range (depends on library)

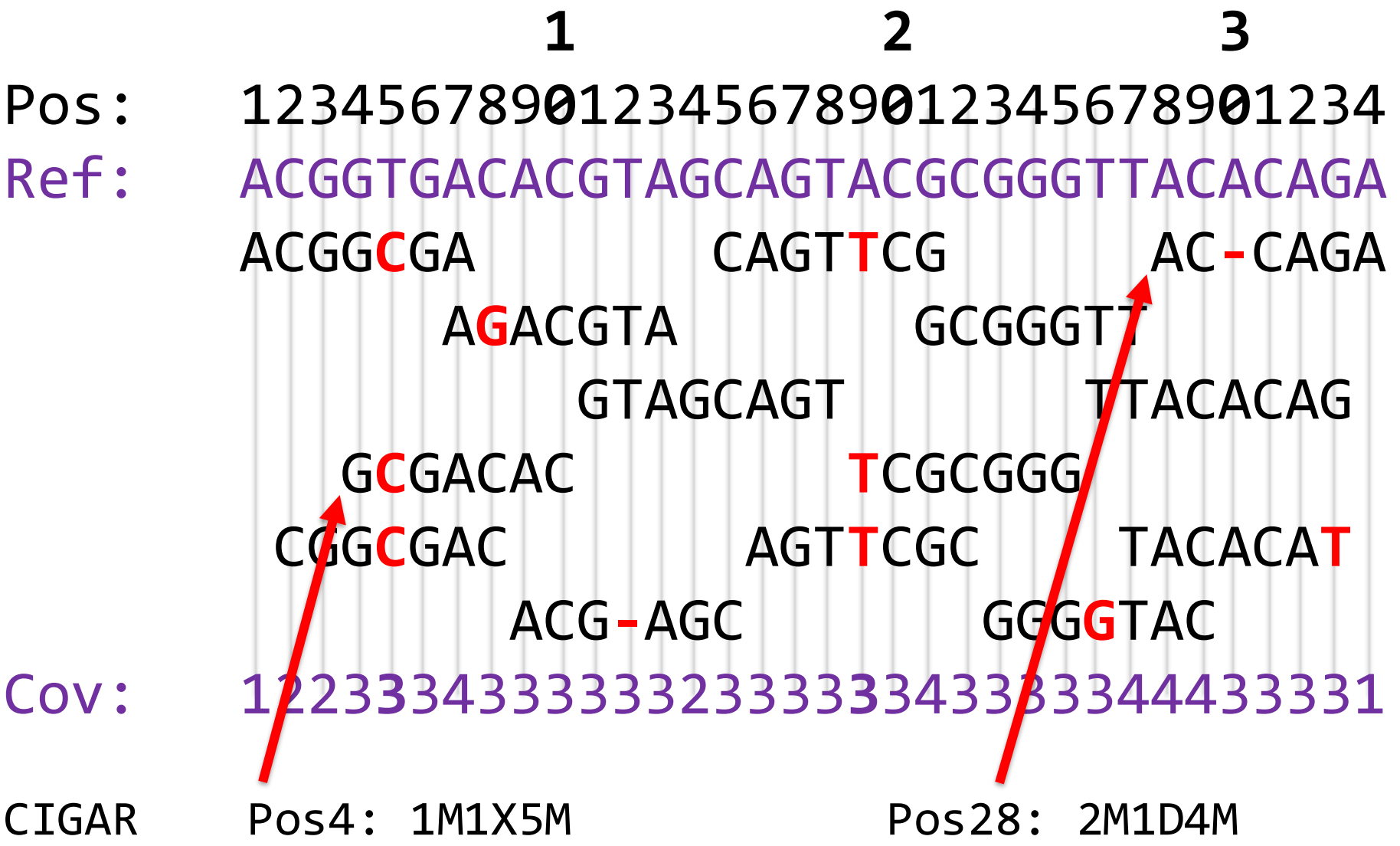


Discordant: inconsistent orientation (mixed up reads? reference issues? Abnormal insert size)

Aligning reads to a reference: Mutations and Indels

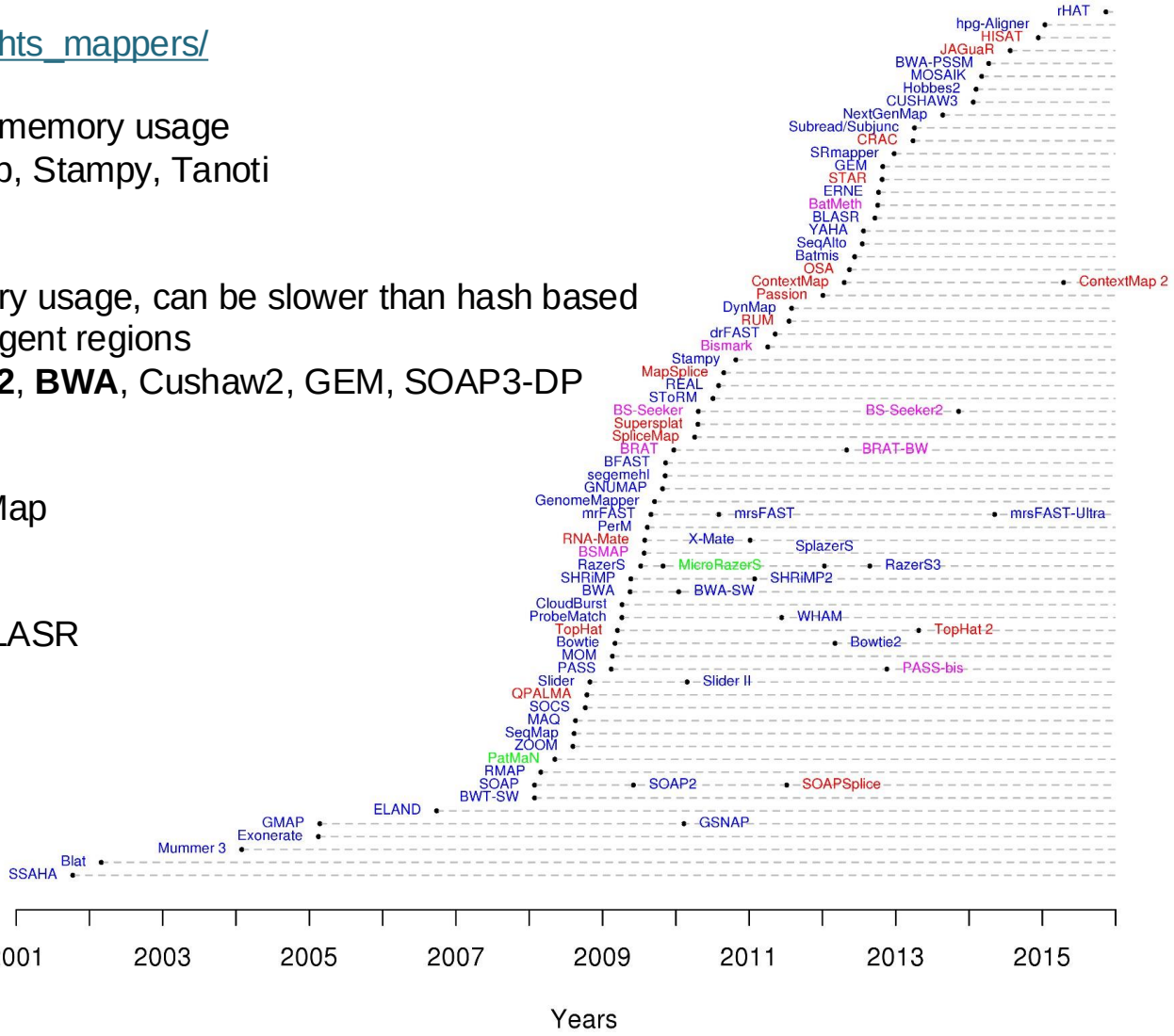


CIGAR
Concise Idiosyncratic Gapped Alignment Report



Aligners – There are Lots

- https://www.ebi.ac.uk/~nf/hts_mappers/
- Hash based - faster, high memory usage
  - Mosaik, NextGenMap, Stampy, Tanoti
- Burrows-Wheeler based
  - Sensitive, low memory usage, can be slower than hash based
  - Can struggle in divergent regions
  - BarraCUDA, **Bowtie2**, **BWA**, Cushaw2, GEM, SOAP3-DP
- RNA-Seq Splice aware
  - HiSAT, TopHat, BBMap
- Long Reads
  - **Minimap2**, LAST, BLASR

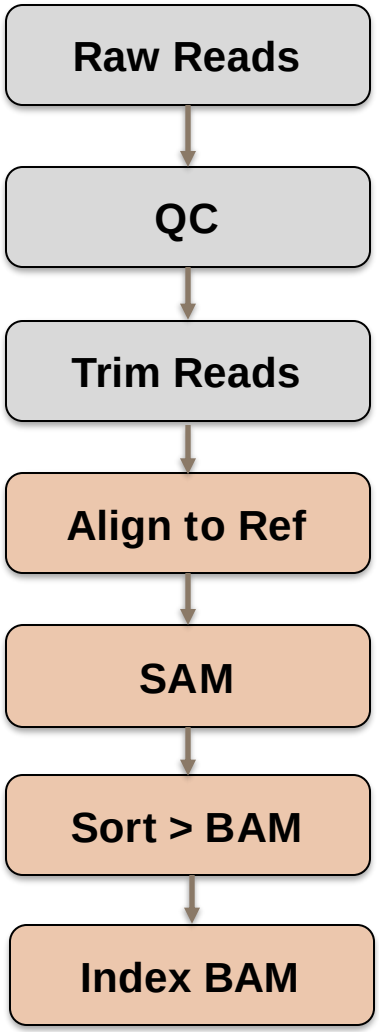


Which aligner to use?

- Bowtie2
- BWA
- Tanoti
- BBMAP
- minimap2
- Mosaik
- …

- Sequencing technology – long vs short reads
- Library/Analysis – e.g. rna-seq
- **Short RNA viral genome - which aligner?**
  - In general aligners are quite consistent in terms of consensus sequence & coverage to a good (close) reference
  - Differences in aligner can be subtle – so may influence **low frequency** variants
- Starting out on a new virus - try a few aligners – not just about most reads aligned – consensus seq and variants

Ref alignment commands - different tools




```
bwa index ref.fasta

bwa mem ref.fasta R1_qc.fq R2_qc.fq > my.sam
```

```
bowtie2-build ref.fasta ref.fasta

bowtie2 --local -x ref.fasta -1 R1_qc.fq -2 R2_qc.fq -S
my.sam
```

```
tanoti -r ref.fasta -i R1_qc.fq R2_qc.fq -o my.sam -p 1
```

```
samtools sort my.sam –o my.bam


samtools index my.bam
```

Which reference sequence?

- **Selecting a suitable reference sequence is an important step.**
  - If your reference is too divergent from your sample if can affect read mapping and possibly the consensus sequence

- **First – obviously want to select the right virus!**
  - If you doing a reference assembly – you probably suspect a particular virus is present in your sample

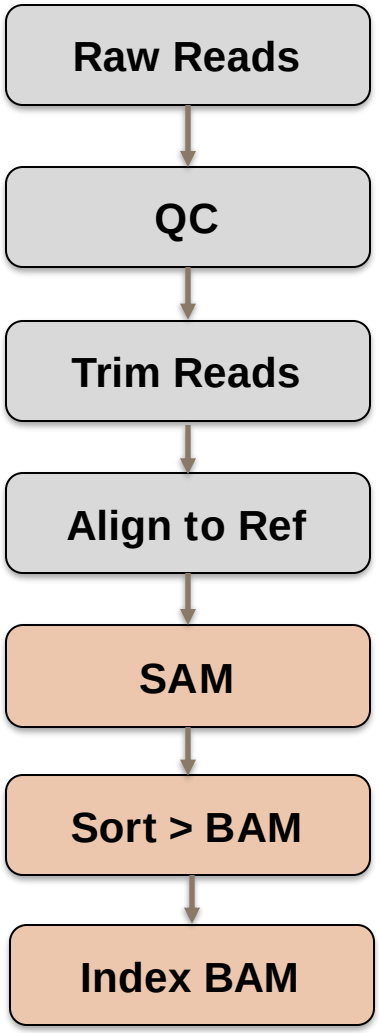
- Second – if a divergent virus e.g. HCV – select the right genotype:
  - Hepatitis C Virus (HCV) - want to select the right genotype – differ by 30–35% at the nucleotide level (subtypes can differ by 15-25% at nucleotide level)

- If unsure what virus is in the sample or suspect it is very divergent
  - **De novo assembly**
  - SHIVER (HIV)
  - Kraken
  - Panel alignment to all genotypes/subtypes – check stats

Multiple Reference sequences

- The reference is in FASTA format: **Need not be one sequence**
  - Segmented virus
    - Influenza: PB1, PB2, PA, NP, HA, NA, M, NS
  - Host
    - Human chromosome 1, 2, 3, 4, 5 etc
  - Panel of viruses
    - HCV 1a, 1b, 1c, 2a, 2b
    - Respiratory viruses
  - Contigs from metagenomics

SAM & BAM files



The result of the alignment step is typically a SAM file

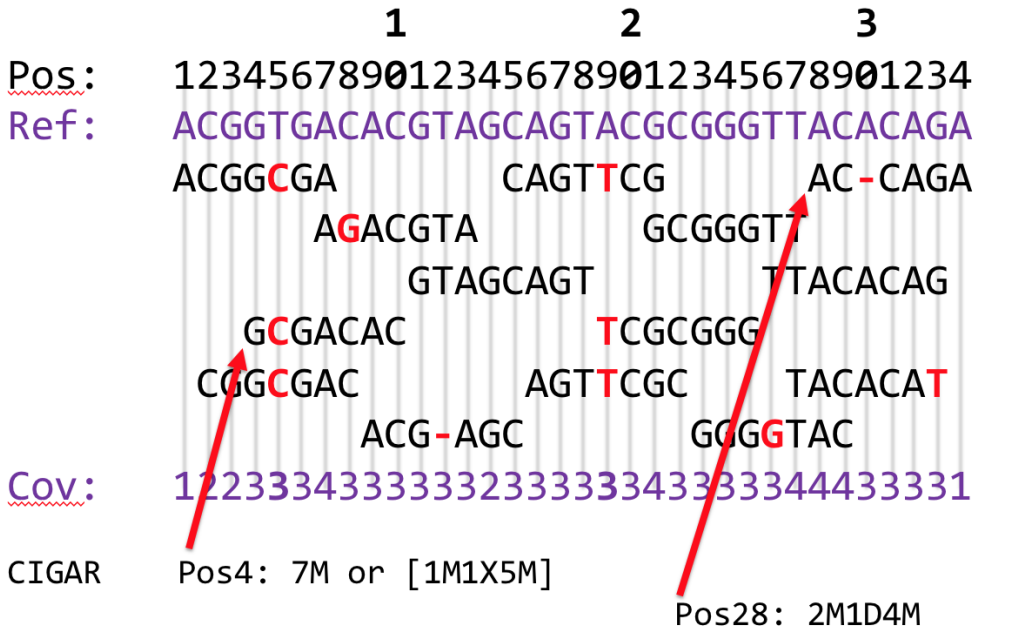
This is then sorted and converted to a BAM file, and indexed

```
samtools sort my.sam -o my.bam
```

```
samtools index my.bam
```

SAM files: Sequence Alignment MAP

- Virtually all aligners output results in **SAM** format
  - **S**equence **A**lignment/**M**ap
- Each line in the SAM file corresponds to a separate alignment
- Sequence and quality strings of the reads stored in the BAM
  - Can extract reads back out of SAM/BAM
  - But always keeps copies of your raw data



The name of the other read in pair

The position the other read in pair is aligned

Template Length/Insert Size

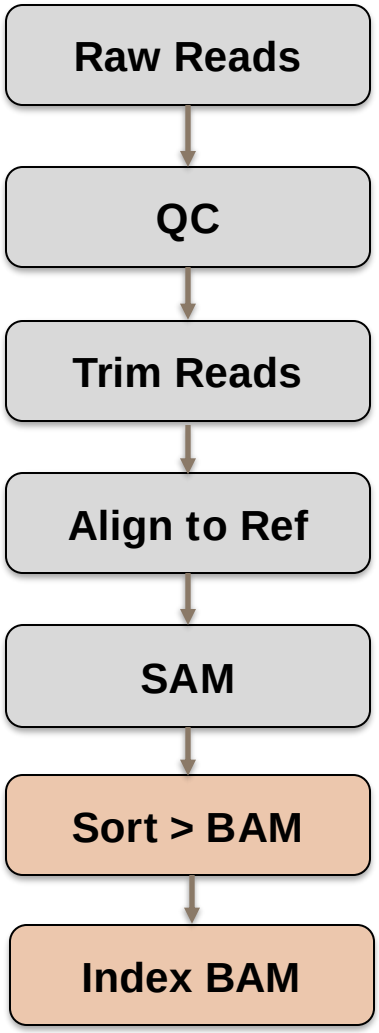
QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUALITY
Read3	10	MyRefSeq	28	52	2M1D4M				ACCAGA	IHGFFF
Read8	10	MyRefSeq	4	57	1M1X5M				GCGACAC	IIIHHGG

Samtools

- **One of the key HTS programs** - provides various utilities for manipulating alignments in the SAM/BAM [and CRAM] formats
  - sorting, merging, indexing and generating alignments in a per-position format.

- Links seamlessly to downstream tools such as VCFTools, BCFTools etc

Converting SAMs to BAMs



- Convert your **SAM** file into **BAM** files
- **BAM B**inary **A**lignment/**M**ap
- BAM files are **compressed** binary versions of the same data (and **faster**)

- Initially the SAM/BAM is sorted by the order the reads were in their files

- **Sort** BAM file
  - All the reads where alignment starts at position 1 first
  - All the reads where alignment starts at position 2, then 3, then 4 etc

  - `samtools sort my.sam -o my.bam`

- **Index** the BAM file
  - Enables downstream tools (consensus/variant calling) to rapidly look up what is aligned to e.g. position 10,456

  - `samtools index my.bam`

SAM Flags – Mapped/Unapped

- **4 =** Read unmapped
- Can be used to give you the most basic of statistics – how many reads are mapped to the reference and how many are unmapped
- Technically, it is counting how many mapped read alignments are in the SAM file

#	Flag	Description
1	**1**	Read paired
2	**2**	Read mapped in proper pair
3	**4**	**Read unmapped**
4	**8**	Mate unmapped
5	**16**	Read reverse strand
6	**32**	Mate reverse strand
7	**64**	First in pair
8	**128**	Second in pair
9	**256**	Not primary alignment
10	**512**	Read fails platform/vendor quality checks
11	**1024**	Read is PCR or optical duplicate
12	**2048**	Supplementary alignment

SAM Flag = 2nd filed of SAM file

QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUALITY
Read3	10	MyRefSeq	28	52	2M1D4M				ACCAGA	IHGFFF
Read8	4	*	0	0					GCGACAC	IIIHHGG

SAM Flags – Mapped/Unapped

- A read can sometimes have multiple alignments
- **256** = not primary = secondary = alternative alignments (equally good or not quite as good)
- **2048** = supplementary alignment = when read is split (spliced) and sections aligned separately

#	Flag	Description
1	**1**	Read paired
2	**2**	Read mapped in proper pair
3	**4**	**Read unmapped**
4	**8**	Mate unmapped
5	**16**	Read reverse strand
6	**32**	Mate reverse strand
7	**64**	First in pair
8	**128**	Second in pair
9	**256**	Not primary alignment
10	**512**	Read fails platform/vendor quality checks
11	**1024**	Read is PCR or optical duplicate
12	**2048**	Supplementary alignment

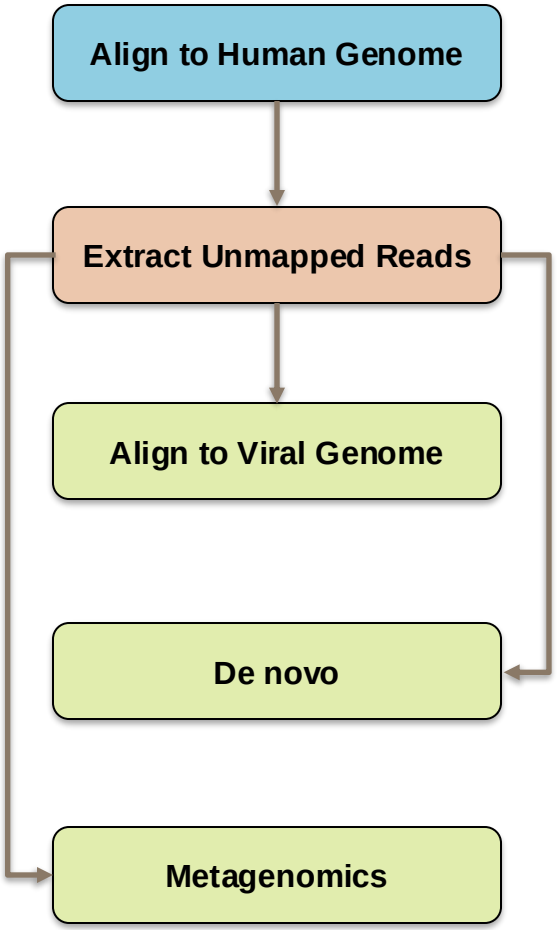
Typically, secondary/supplementary should be few for short RNA virus genome, but if louts it can indicate:

Repeat regions, Large deletions, Poor reference sequence

SAM Flag = 2nd filed of SAM file

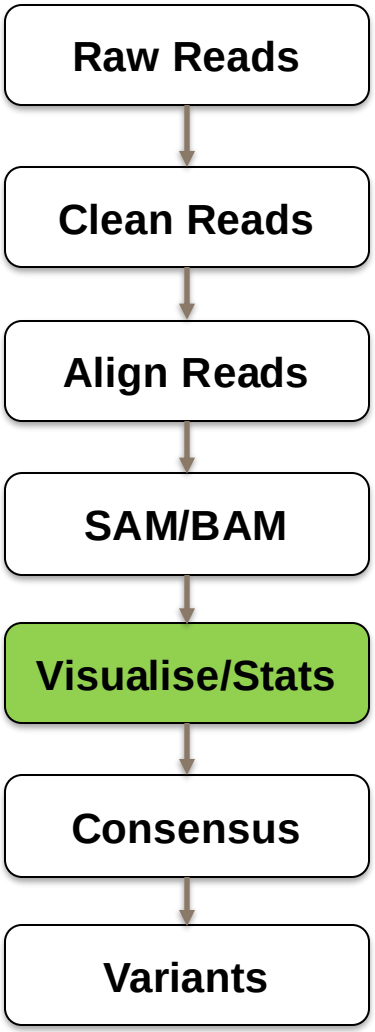
QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUALITY
Read3	10	MyRefSeq	28	52	2M1D4M				ACCAGA	IHGFFF
Read8	4	*	0	0					GCGACAC	IIIHHGG

Host filtering – exploiting flag4 (unmapped)



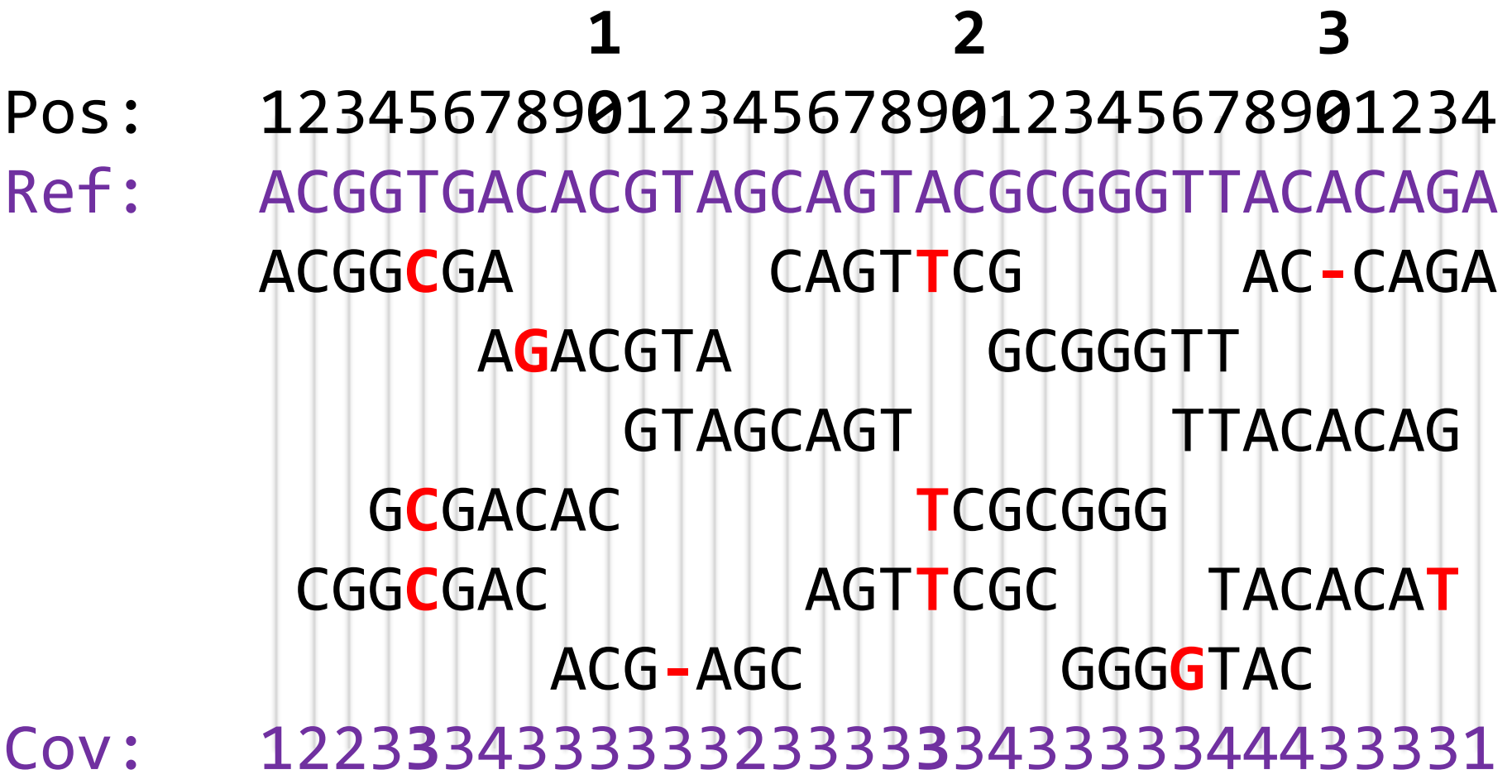
- Create read files without the human host
- samtools has a host of other function available:
  - samtools fastq
  - **samtools depth**
  - samtools stats
  - samtools ampliconclip
  - **samtools idxstats**
  - **samtools flagstat**
  - samtools consensus

Post Assembly – after the BAM



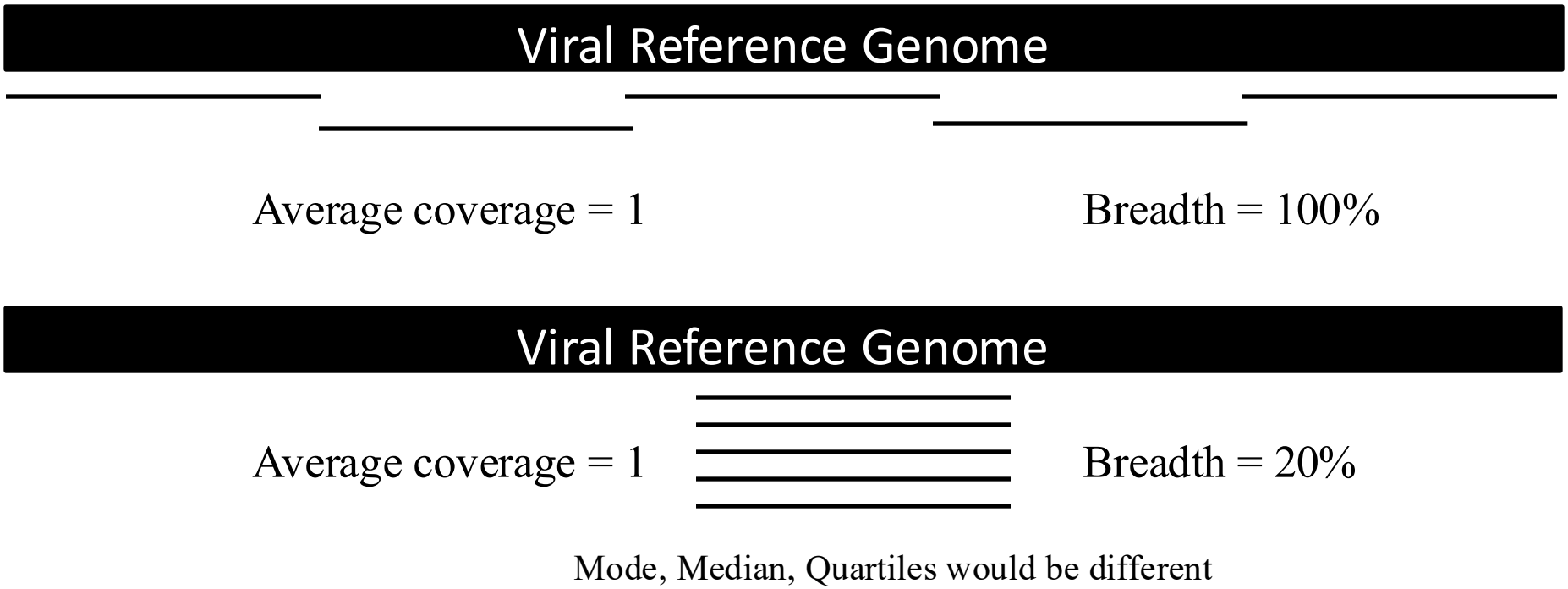
- Post assembly checks
  - Summary statistics:
    - Number of reads aligned
    - Number of reads unaligned
    - **Average depth of coverage**
    - **Breadth of coverage**
  - **Coverage plot**
  - **Visualisation of entire alignment**

Coverage


1 2 3
Pos: 123456789012345678901234567890 1234
Ref: ACGGTGACACGTAGCAGTACGCGGGTTACACAGA
ACGGCGA
CAGTTCG
AC-CAGA
AGACGTA
GCGGGTT
GTAGCAGT
TTACACAG
GCGACAC
TCGCGGG
CGGCGAC
AGTTCGC
TACACAT
ACG-AGC
GGGGTAC
Cov: 122333433333323333334333333444333331

Coverage Depth & Breadth

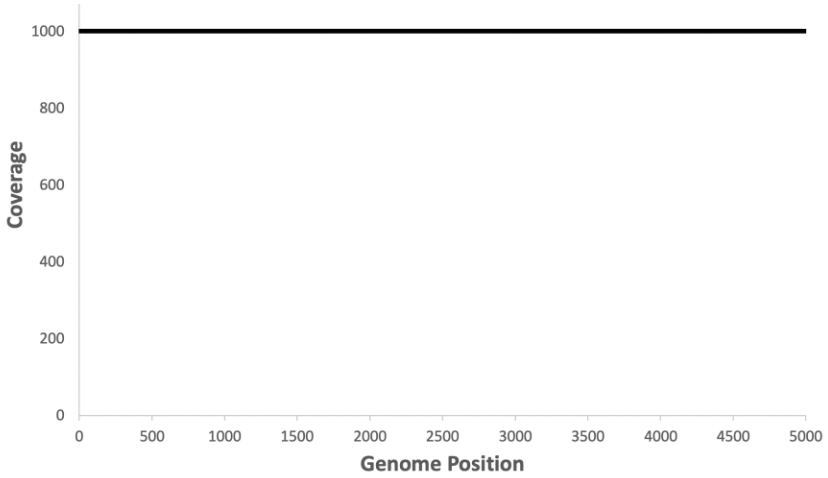
- **Coverage** is the number of reads that "cover" a particular genome coverage
  - **Depth**
- Average (mean) coverage: the average coverage across all genome positions
- Breadth of coverage: how much of the genome is actually covered



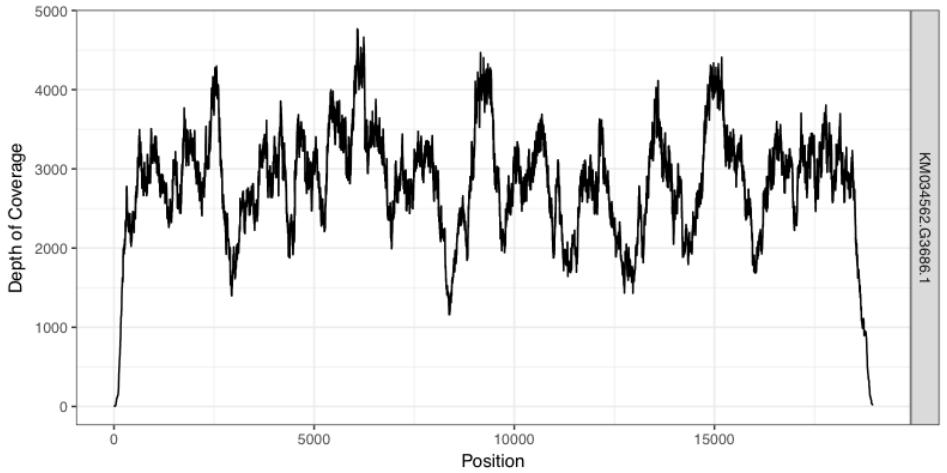
Perfect Coverage Plots

- High uniform coverage across the entire genome
- Biases in library prep fragmentation and PCR (GC content)
  - the terminal ends are typically poorly covered
- Biases in bait capture, amplicon/primer efficiency, extraction methods

Don't think I've ever seen this

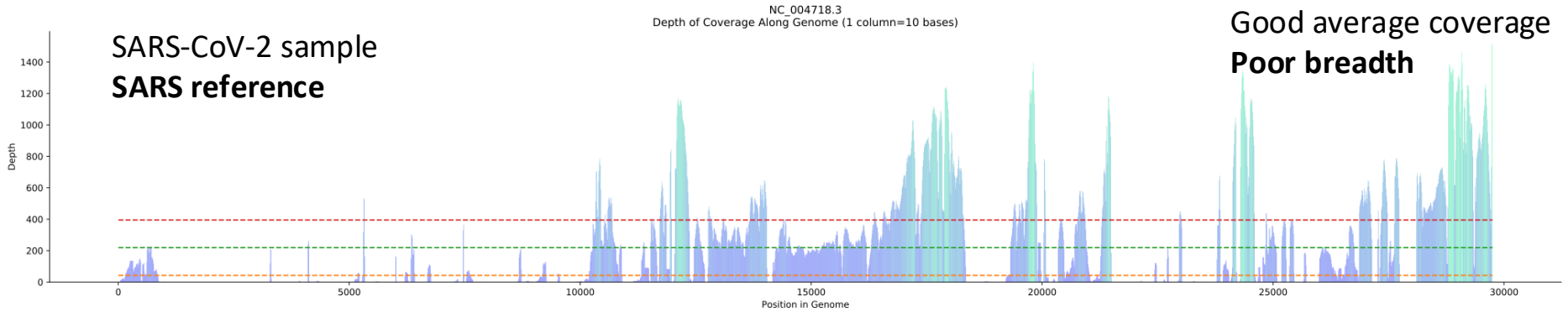


This is perfectly normal

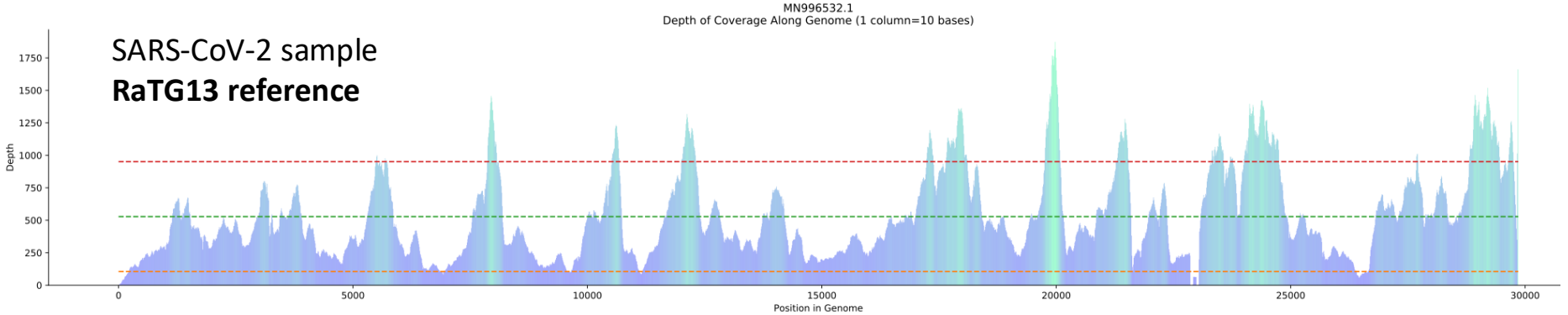


Coverage plots – bad reference

- Sporadic coverage with frequent regions dropping down to zero can indicate a poor reference seq
- The reference is too divergent in many regions and reads can not be aligned at the nucleotide level



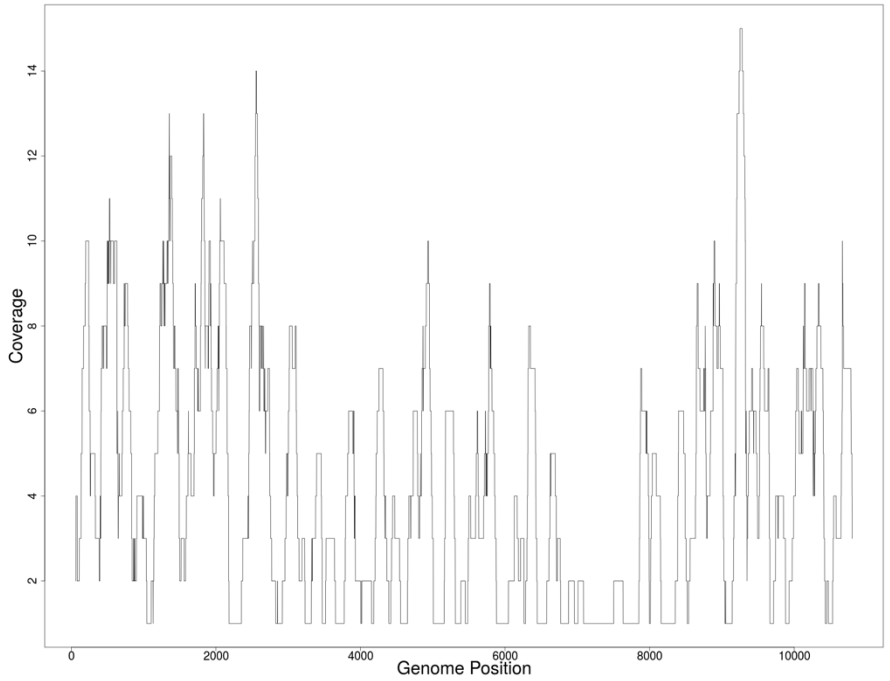
- Viruses can be very diverse – aligning to a different genotype/strain can give obscure results
- Align to different refs, genotype detection tools, **de novo assembly**



Coverage plots – low coverage

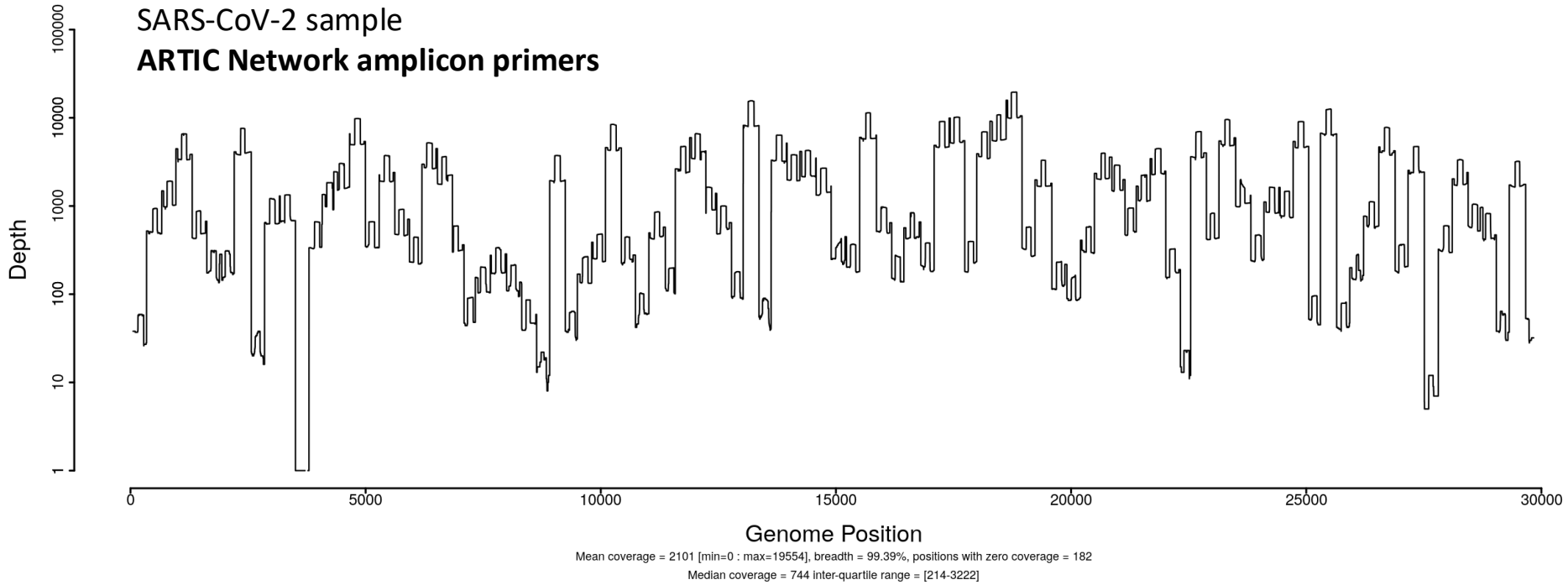
- Sometimes there is just not enough data present
  - Lower read trimming threshold
  - Just use the raw reads
    - Will be noisey
- Re-run the sample
  - Perhaps it was a bad run
  - Combine run data
- PCR amplification
- Bait capture

Louping ill virus sample



Coverage plots – amplicons

- Amplicon data can give step like plots



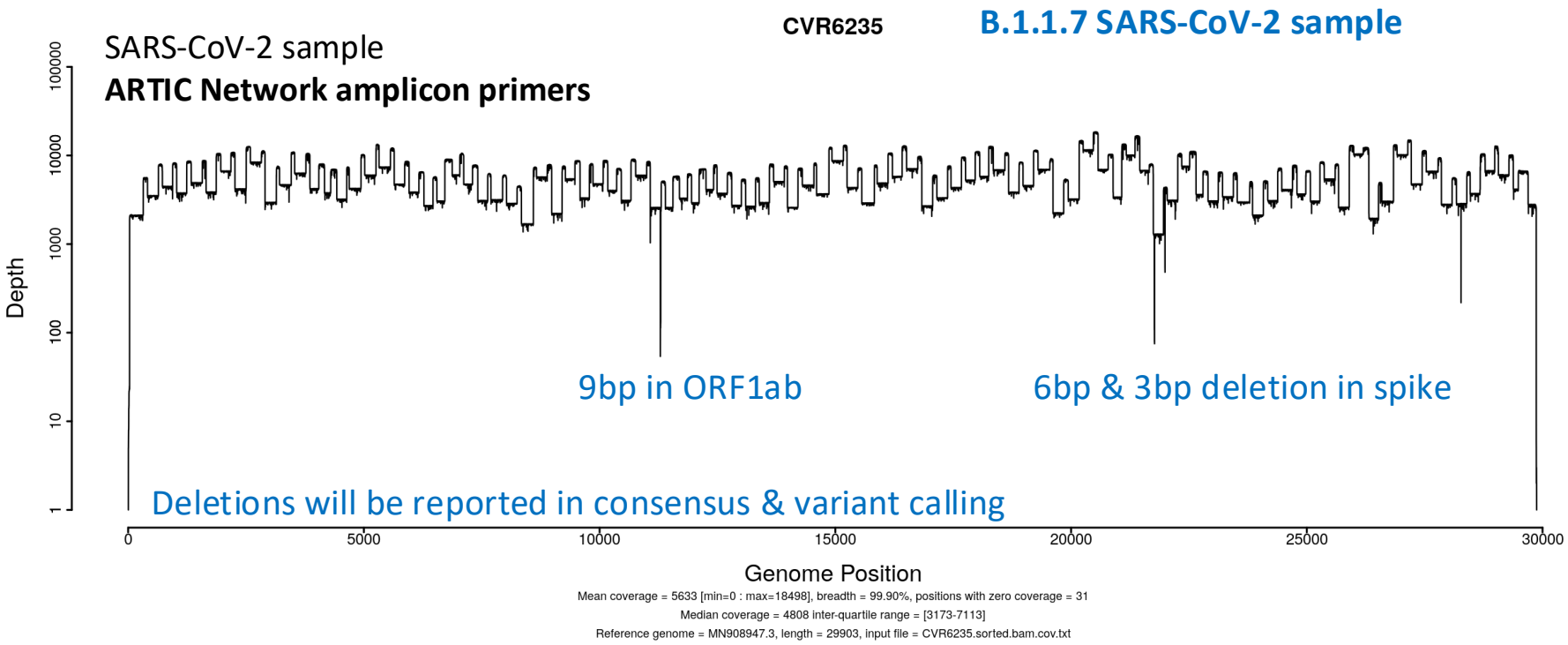
<- Illumina paired read2 250 bases

Illumina paired read1 250 bases->

400bp amplicon

Coverage plots – deletions

- Sudden drops in coverage at a small number of sites can indicate deletions with respect to the reference

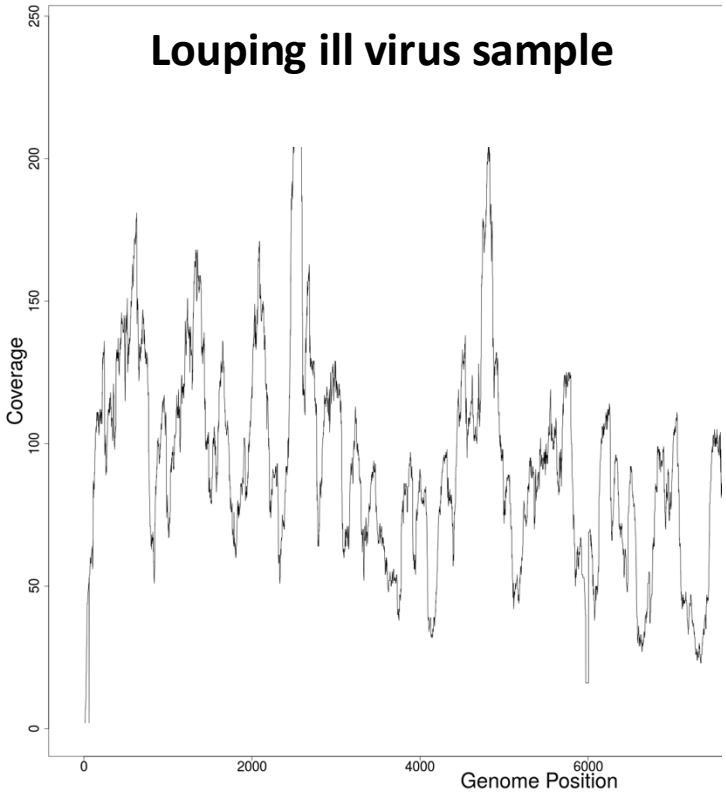


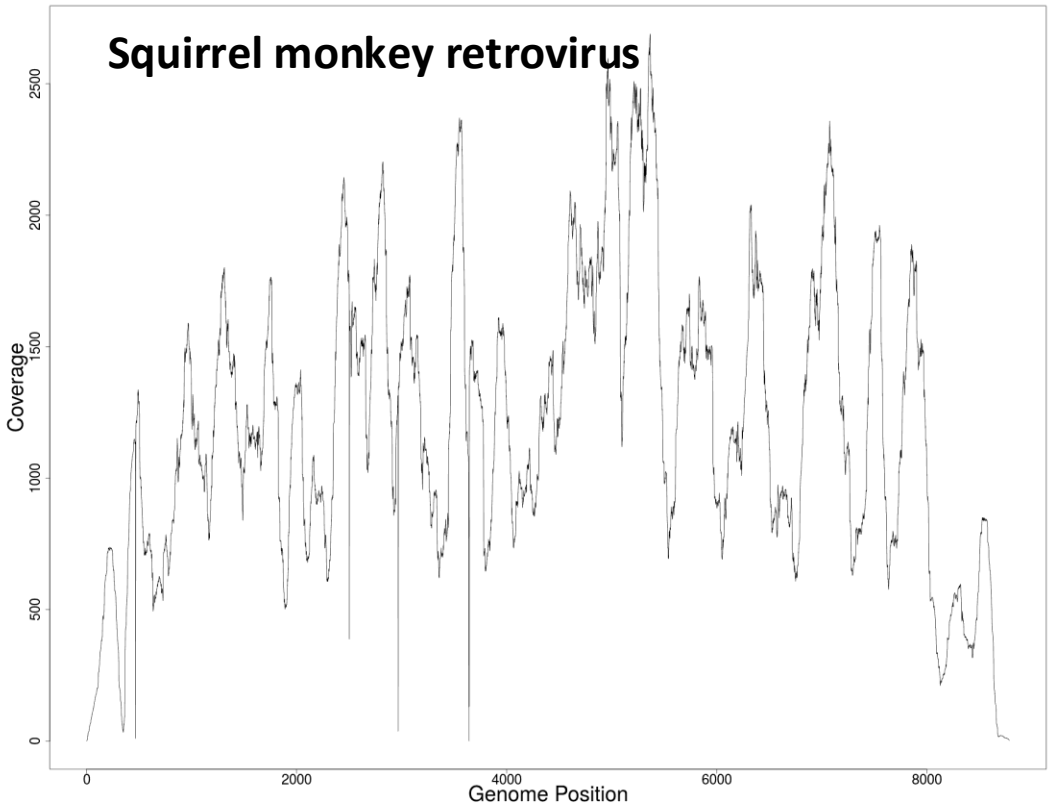
This is a log plot and noisy minion data – the deleted sites still have some coverage but this is nothing compared to the other sites

Reference assembly – tunnel vision

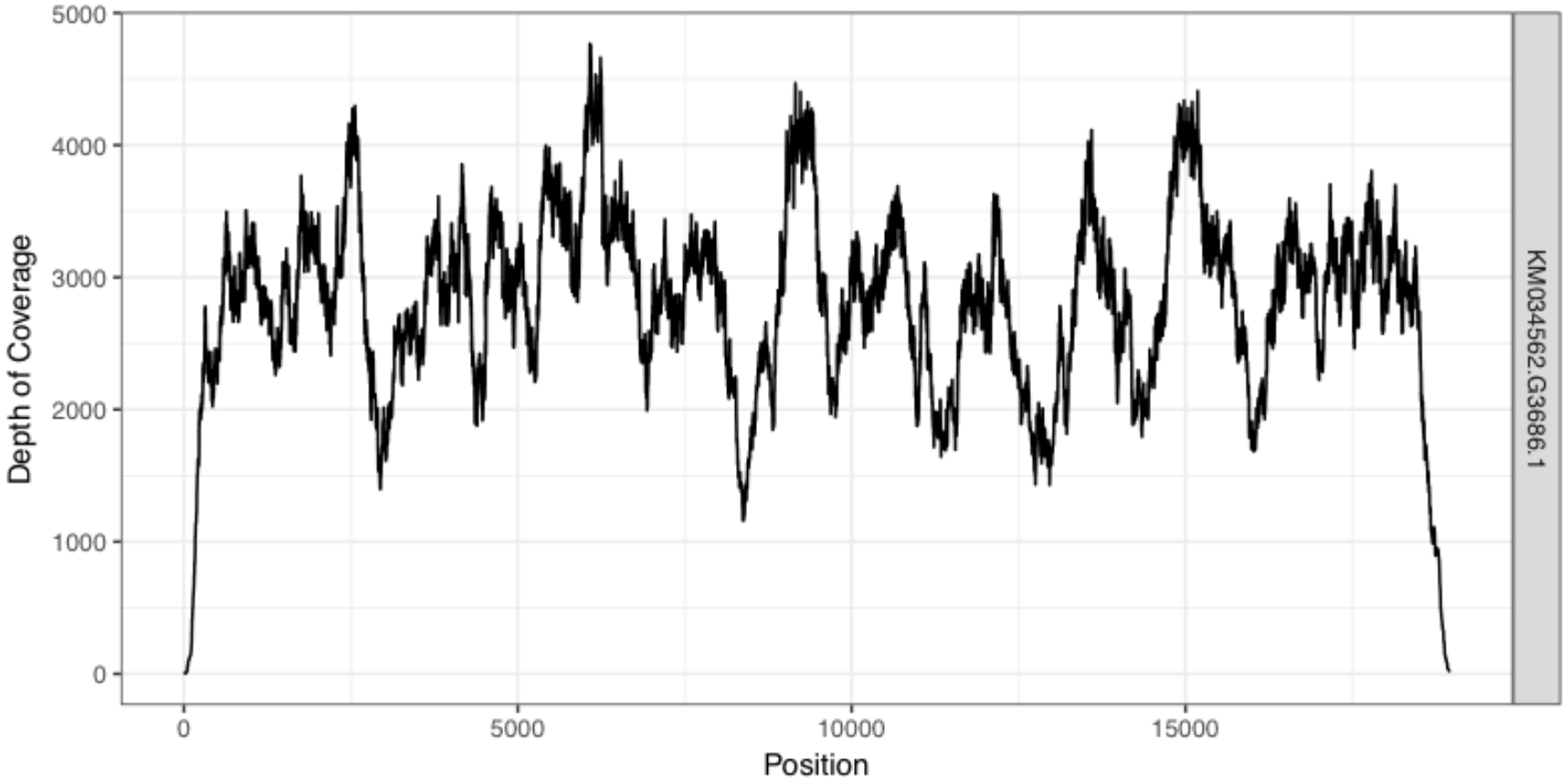
- With reference assembly you automatically focus on a single virus
  - You can align to multiple viruses in one go
  - But you will still need to decide what viruses to investigate

- **Good to run kraken/centriguge on your samples to (viral & mycoplasma contaminants)**





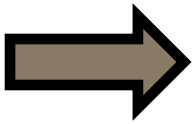
How do you create coverage plots?

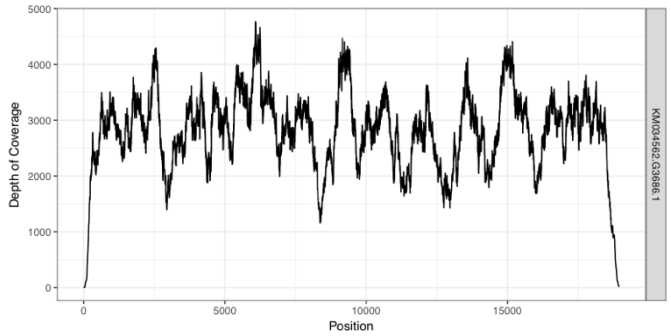


samtools depth

- samtools has a built in function called ‘depth’
- `samtools depth -aa -d 0 my.bam > my_depth.txt`
- -aa: output data for absolutely all positions (even positions with zero coverage)
- -d 0: disable the maximum depth to report [default is 8000]
- 3 column text file:

Chromosome	Position	Depth
MN908947.3	1	0
MN908947.3	2	13
MN908947.3	3	34
…	…	…



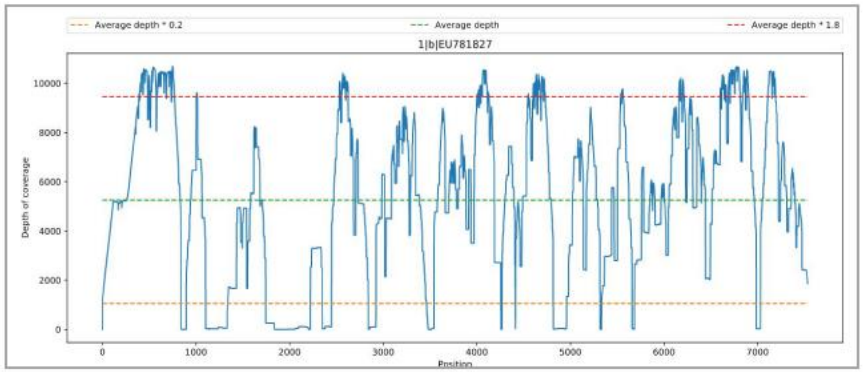


All chromosome will be reported in turn

weeSam - https://github.com/centre-for-virus-research/weeSAM

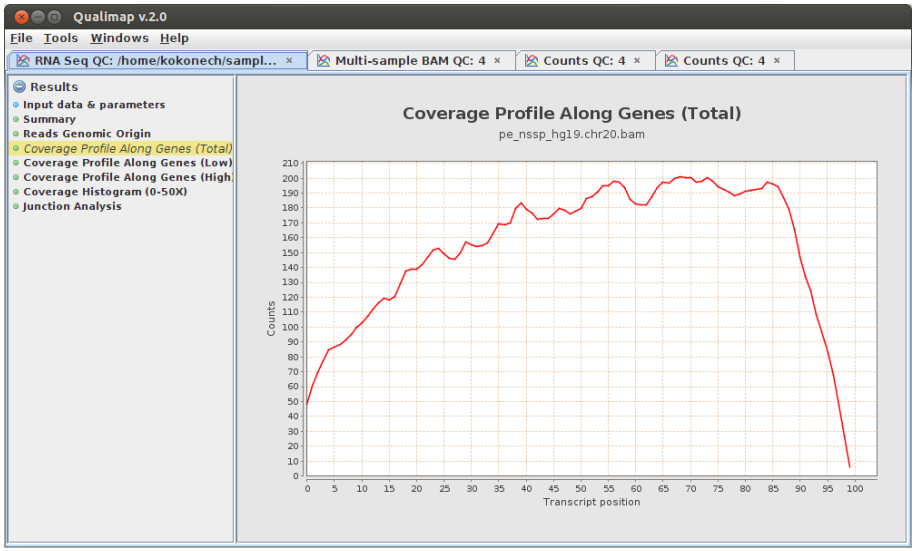
- weeSam is program that can give you information on breadth and depth of coverage as well as generate a coverage plot automatically
- `weeSAM --bam 1a.bam --html 1a`
- `1a_html_results/1a.html`

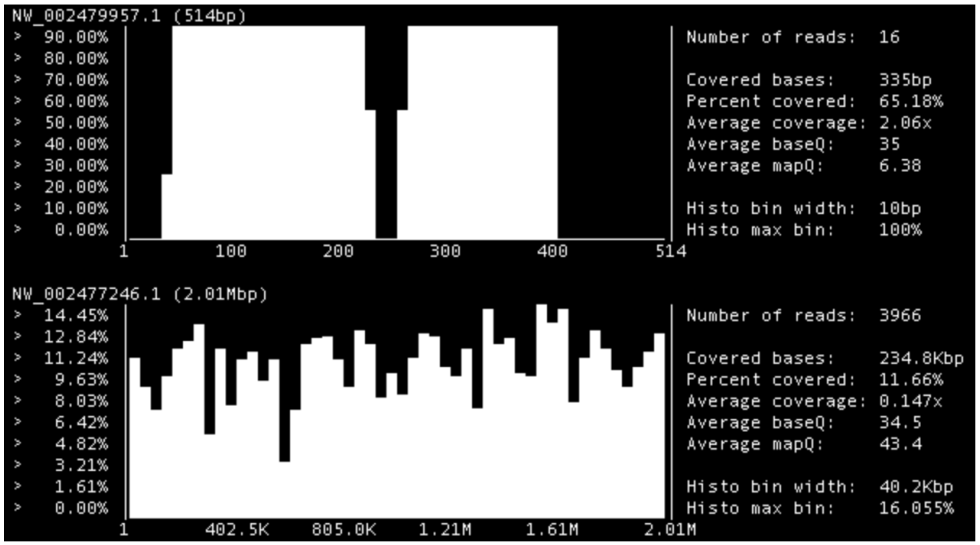
Ref_Name	Ref_Len	Mapped_Reads	Breadth	%_Covered	Min_Depth	Max_Depth	Avg_Depth	Std_Dev	Above_0.2_Depth	Above_1_Depth	Above_1.8_Depth	Variation_Coefficient
NC_004102.1 Hepatitis C virus genotype 1, complete genome	9646	640000	9646	100.00	13	10729	9941.89	1699.34	98.82	90.91	0.00	0.17



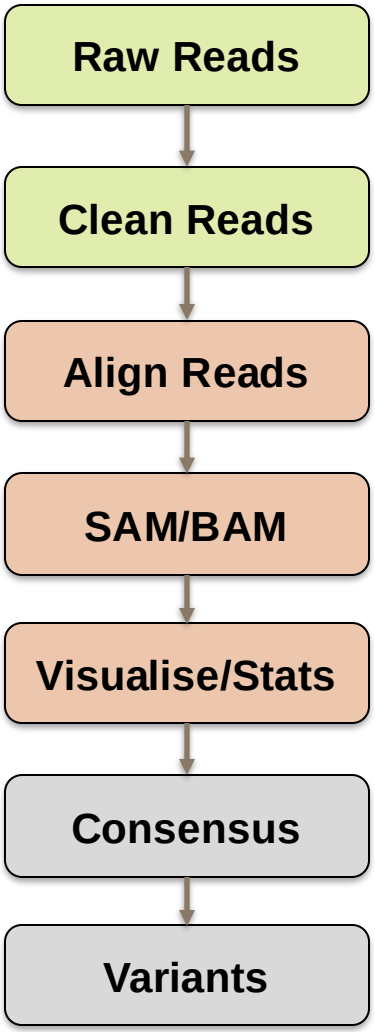
Other tools

- Qualimap: http://qualimap.conesalab.org
- bamCov – https://github.com/fbreitwieser/bamcov



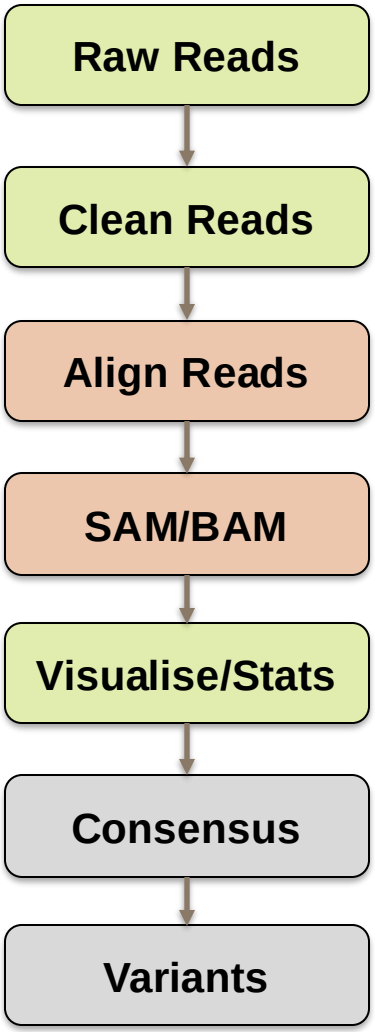


Practical



- Hepatitis C virus (HCV) samples
- **Simulated Illumina paired end reads - R1 R2**
- **HCV Genotype 1a reference genome**
- First sample
  - High quality simulated data set
- 3 other samples
  - Real patient HCV samples (genotype 1a)
- BWA aligner, samtools, SAM/BAM, stats
- weeSAM covereage

Practical – HCV_SIM commands – adapt for another sample



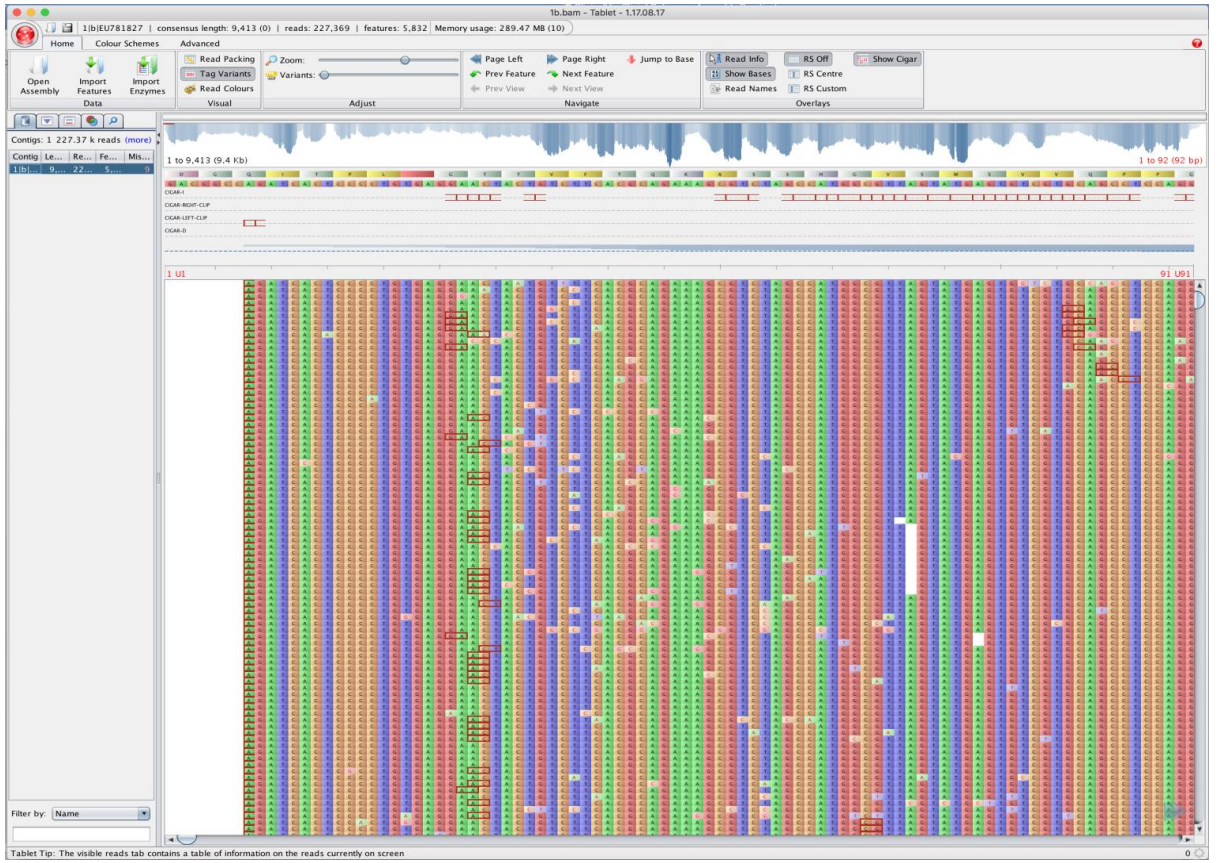

```
bwa index ../1a_hcv_ref.fasta

bwa mem ../1a_hcv_ref.fasta hcv_sim_R1.fq
hcv_sim_R2.fq > 1b.sam

samtools view -b 1b.sam

samtools sort -o 1b.bam

samtools index 1b.bam

rm 1b.sam

samtools view -c -F4 1b.bam

samtools view -c -f4 1b.bam
```

The End … Tablet

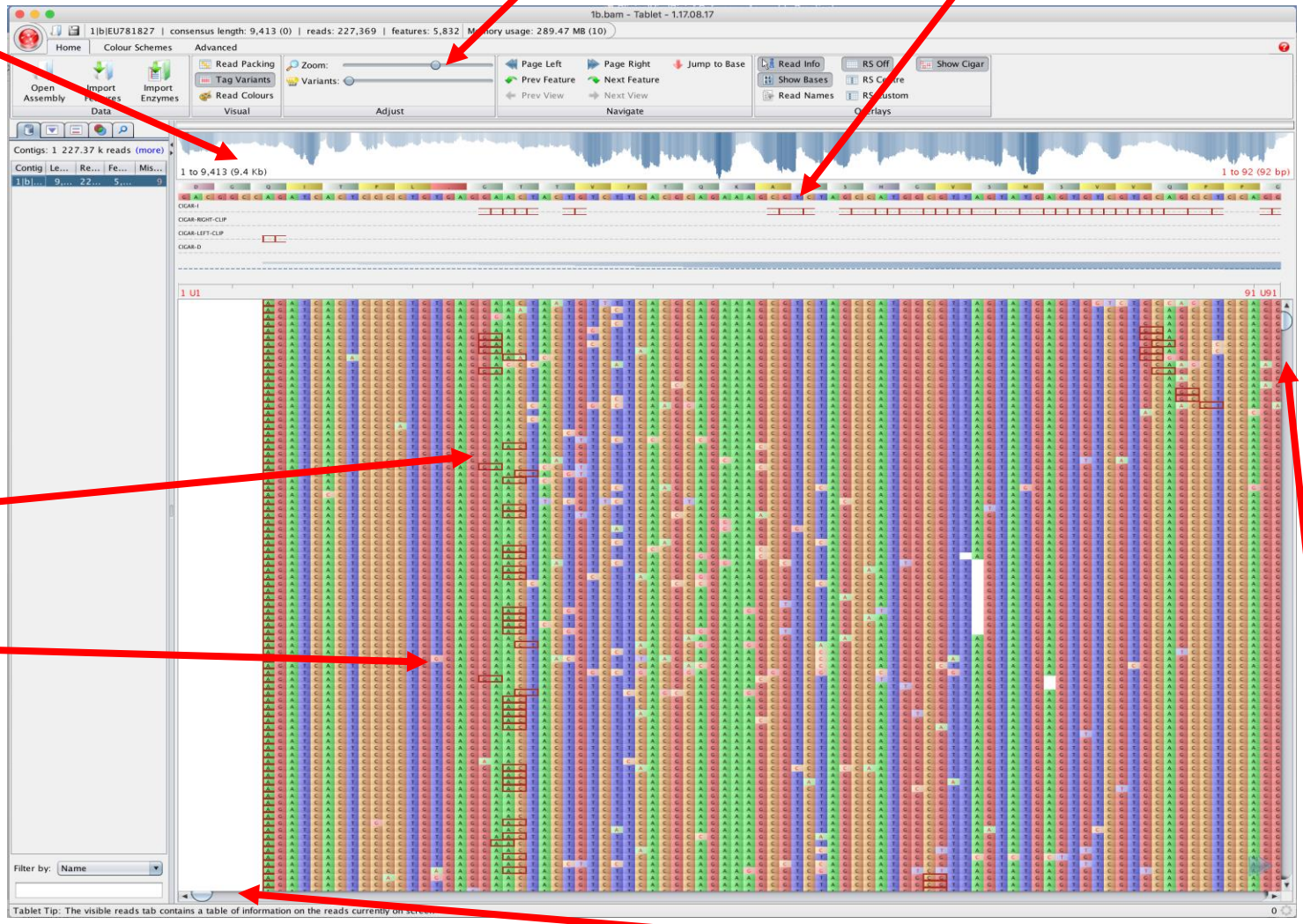
- Tablet demo if time later on

Tablet: https://ics.hutton.ac.uk/tablet/

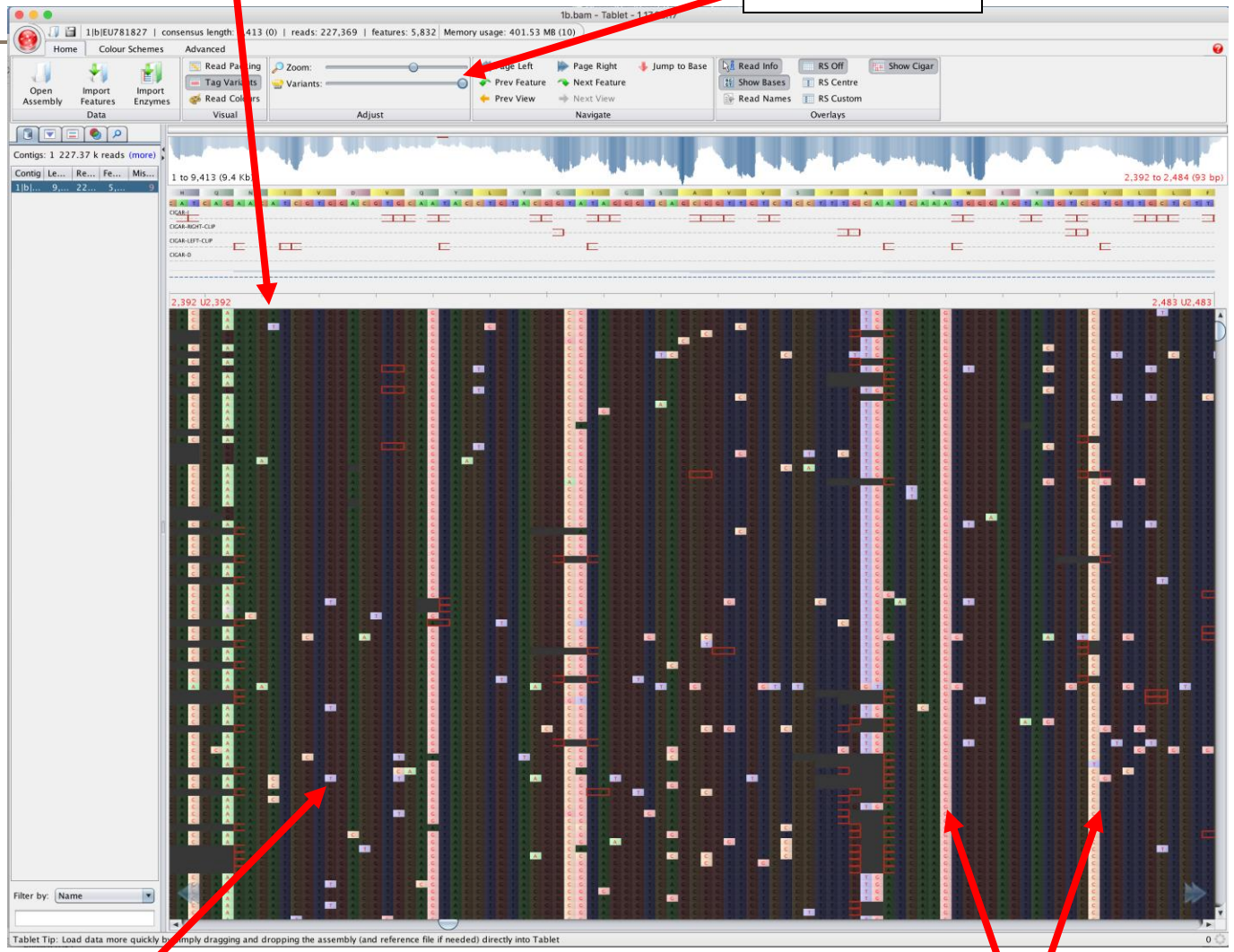
- **`tablet`**
- Zoom, scroll, colour schemes: nucleotides, direction, mutations
- Tablet is a lightweight, high-performance graphical viewer for next generation sequence assemblies and alignments.
  - BAM file
  - Reference file



Tablet


Zoom slider
Reference sequence
Coverage overview
Read display
Mutations
Scroll bars
1b.bam - Tablet - 1.17.08.17
Home
Colour Schemes
Advanced
Open Assembly
Import Features
Import Enzymes
Read Packing
Tag Variants
Read Colours
Zoom:
Variants:
Page Left
Page Right
Jump to Base
Prev Feature
Next Feature
Prev View
Next View
Read Info
Show Bases
Read Names
RS Off
Show Cigar
Contigs: 1 227.37 k reads (more)
1 to 9,413 (9.4 Kb)
1 to 92 (92 bp)
Filter by: Name

Tablet


Mouse Genome Position
Variants slider
Minority variants
Consensus level variants