



Reference alignment of reads

Richard Orton

MRC-University of Glasgow Centre for Virus Research

February 2024

Richard.Orton@glasgow.ac.uk



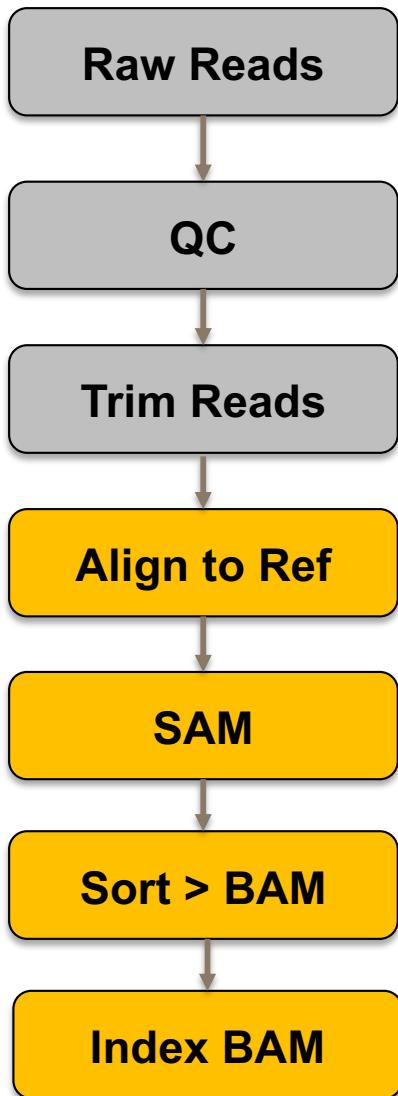
**Here is the bird that never flew
Here is the tree that never grew
Here is the bell that never rang
Here is the fish that never swam**



the way and the truth and the life



Previously ...



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

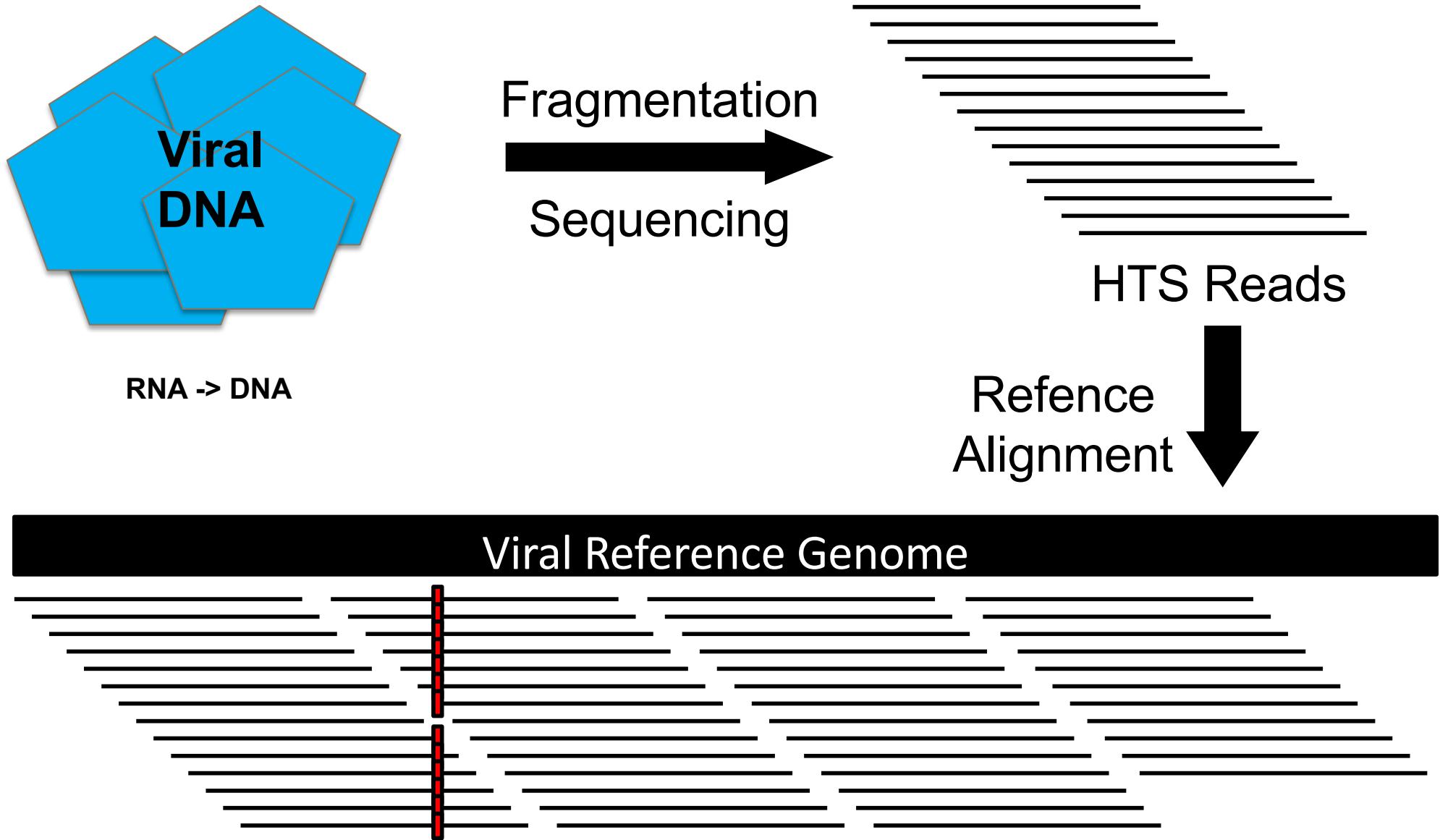
Overview

- **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.

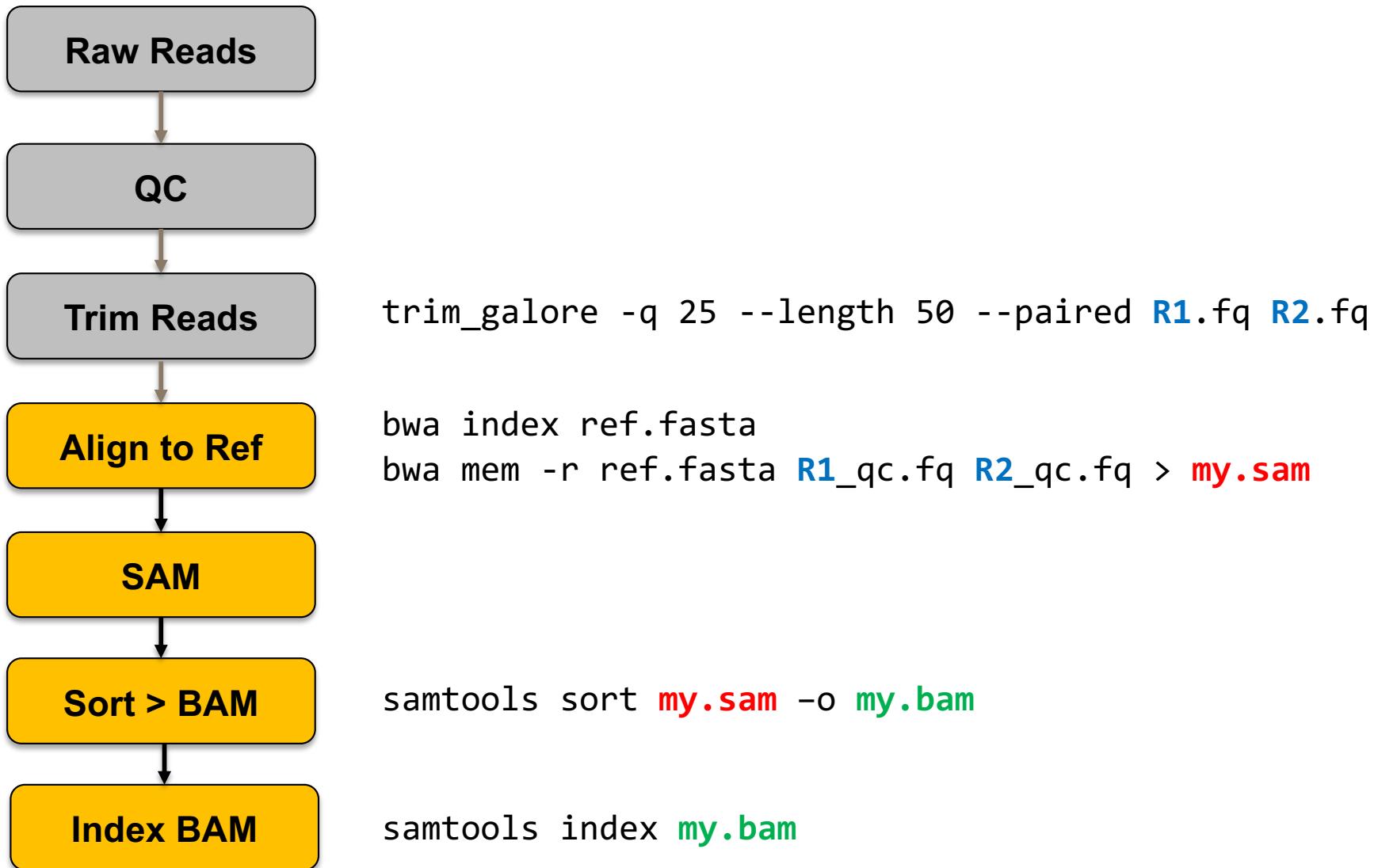
Reference alignment

- **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



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: ACGGTGACACGTAGCAGTACGCAGGTTACACAGA

Read: GTTACAC

Aligning reads to a references

Ref: ACGGTGACACGTAGCAGTACGCGGGTTACACAGA

Read: GTTACAC

↓

Matches 0/7 7/7

Mismatches 7/7 0/7

Aligners check the reverse complement

Ref: ACGGTGACACGTAGCAGTACGCAGGTTACACAGA

Read: ACTGCTA



TAGCAGT



ACTGCTA

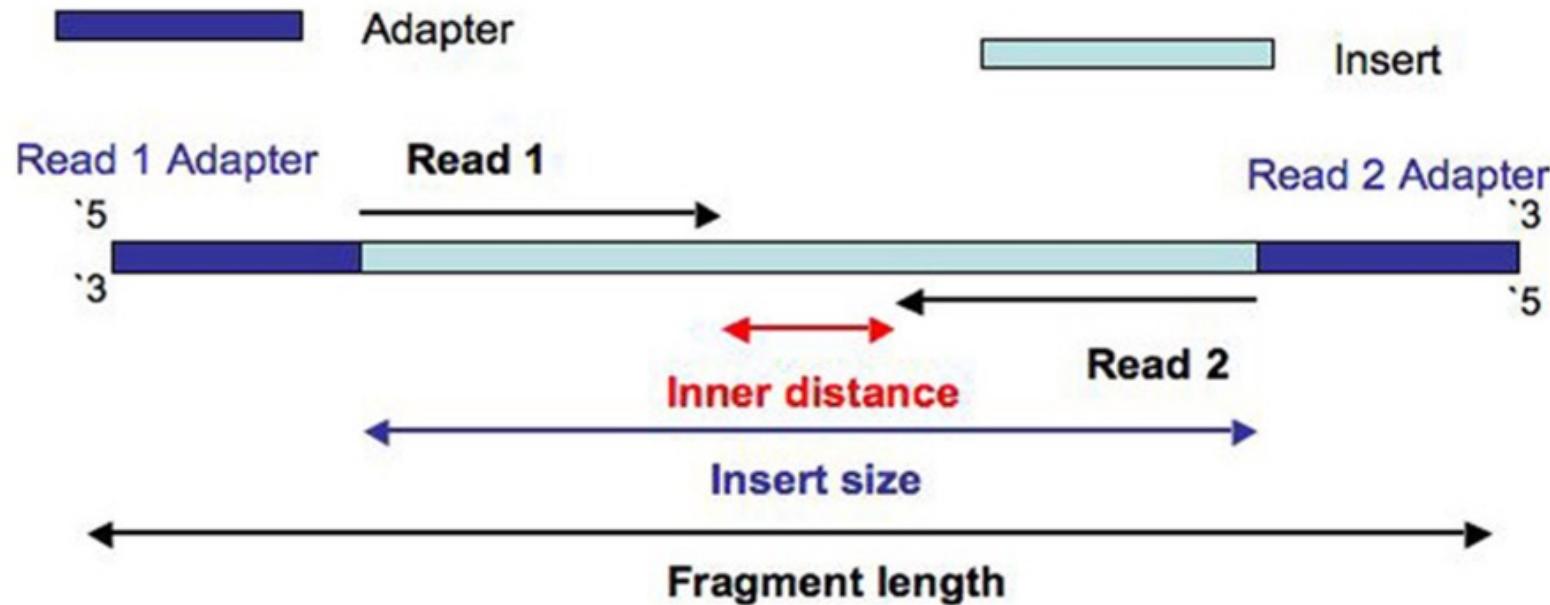
Reverse Complement

TAGCAGT

Matches	4/7
Mismatches	3/7

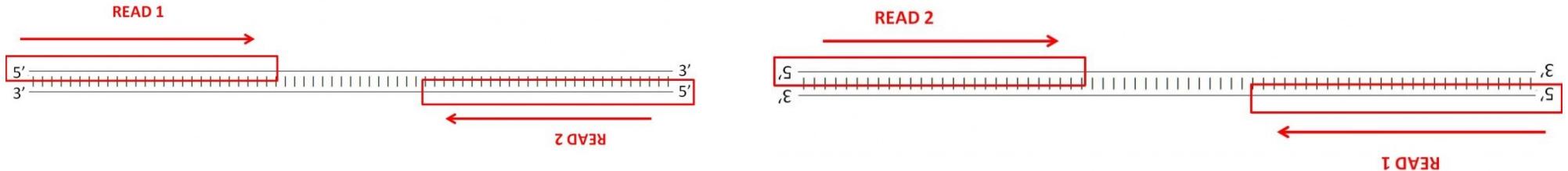
7/7
0/7

Paired end ... Insert Size

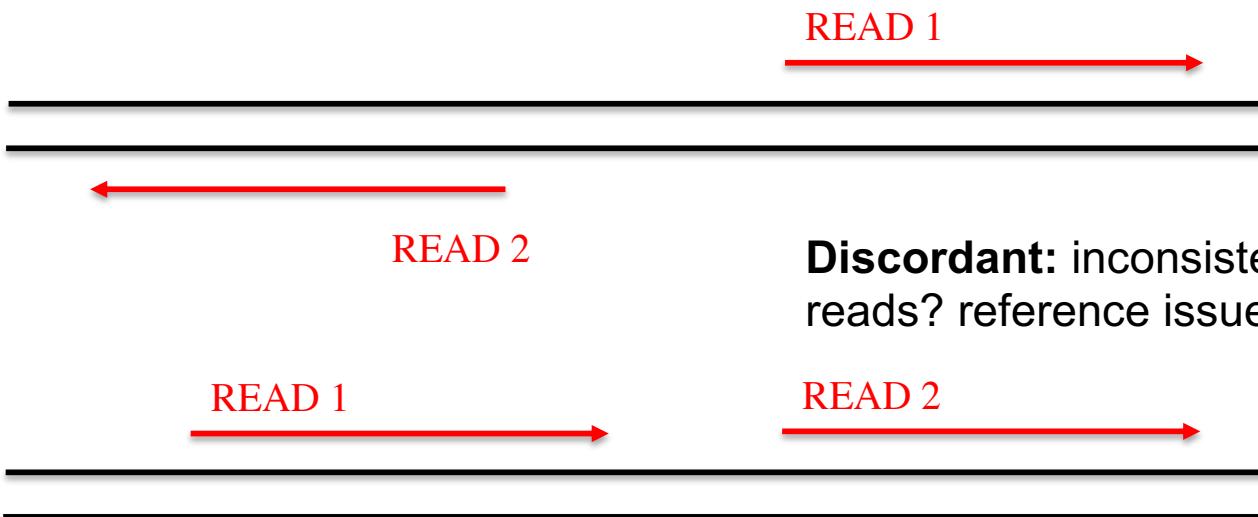


- Typically, the two reads do not overlap, but they can if the reads are long and fragments are short [redundant data, but can be used to correct errors]
- 500bp fragments + 2 x 300bp reads = 100bp overlap
- Turner 2014, Frontiers in Genetics

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

Ref: ACGGTGACACGTAGCAGTACGCGGGTTACACAGA
ACGG**C**GA CAGT**T**CG AC-**C**AGA
 AGACGTA GC**GGG**TT
 TTACACAG
GCGACAC **T**CG**CGGG**
CGG**C**GAC AGT**T**CGC TACACAT**T**
 ACG-AGC GGG**G**TAC

CIGAR

Concise Idiosyncratic Gapped Alignment Report

1

2

3

Pos: 1234567890123456789012345678901234

Ref: ACGGTGACACGTAGCAGTACGCGGGTTACACAGA

ACGG**CGA**

CAGT**TCG**

AC-CAGA

AGACGTA

GC~~GGG~~T~~T~~

GTAGCAGT

TTACACAG

GCGACAC

TCGCGGG

CGGCGAC

AGT**TCG**C

TACACAT**T**

ACG-AGC

GGGGTAC

Cov: 1223334333332333333433334443331

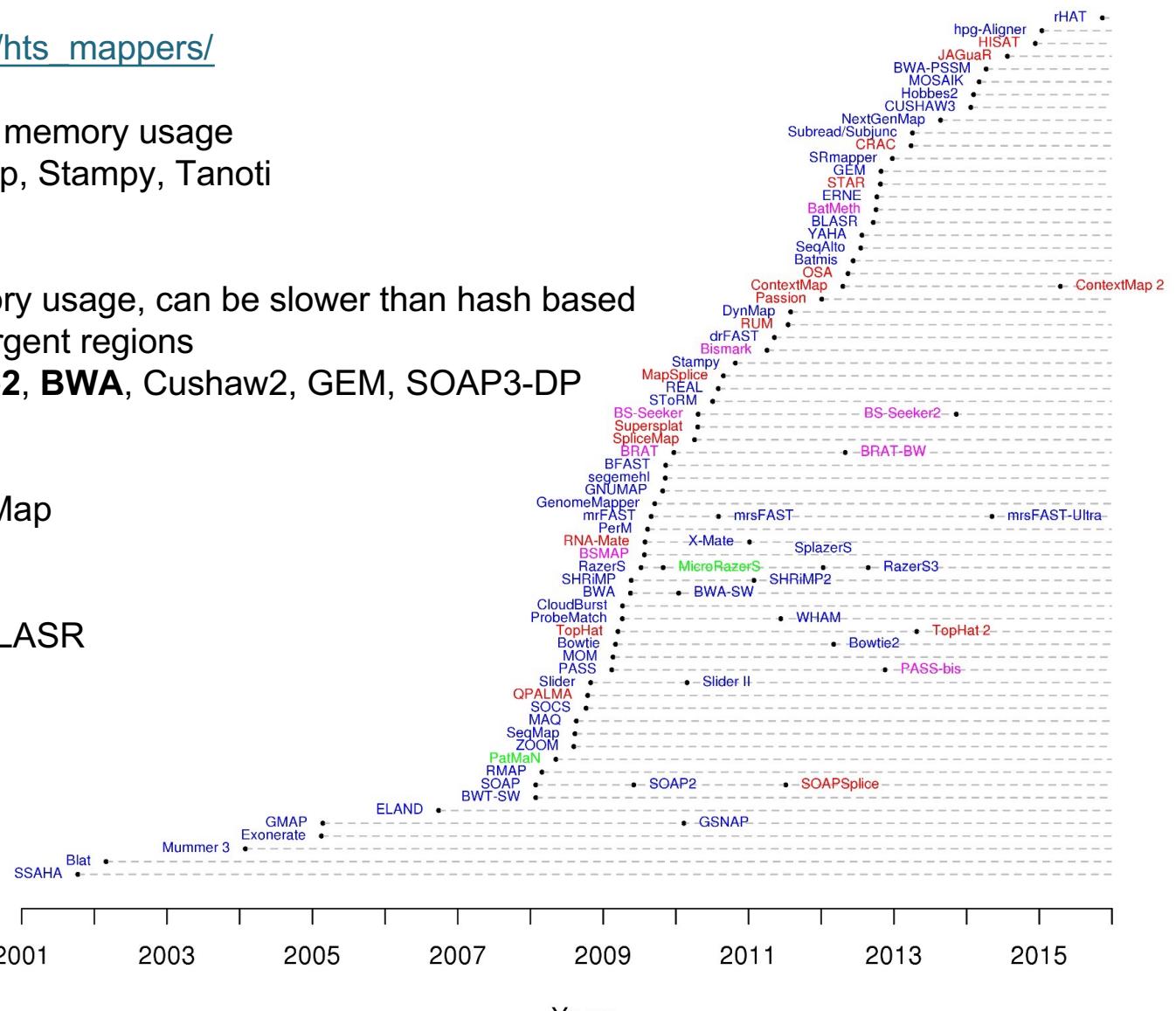
CIGAR

Pos4: 1M1X5M

Pos28: 2M1D4M

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**, Cusheaw2, GEM, SOAP3-DP
- RNA-Seq Splice aware
 - HiSAT, TopHat, BBMap
- Long Reads
 - **Minimap2**, LAST, BLASR



Which aligner to use?

Bowtie2

- Sequencing technology – long vs short reads

BWA

- Library/Analysis – e.g. rna-seq

Tanoti

- **Short RNA viral genome - which aligner?**

- In general aligners are quite consistent in terms of consensus sequence & coverage to a good (close) reference

BBMAP

- Differences in aligner can be subtle – so may influence **low frequency** variants

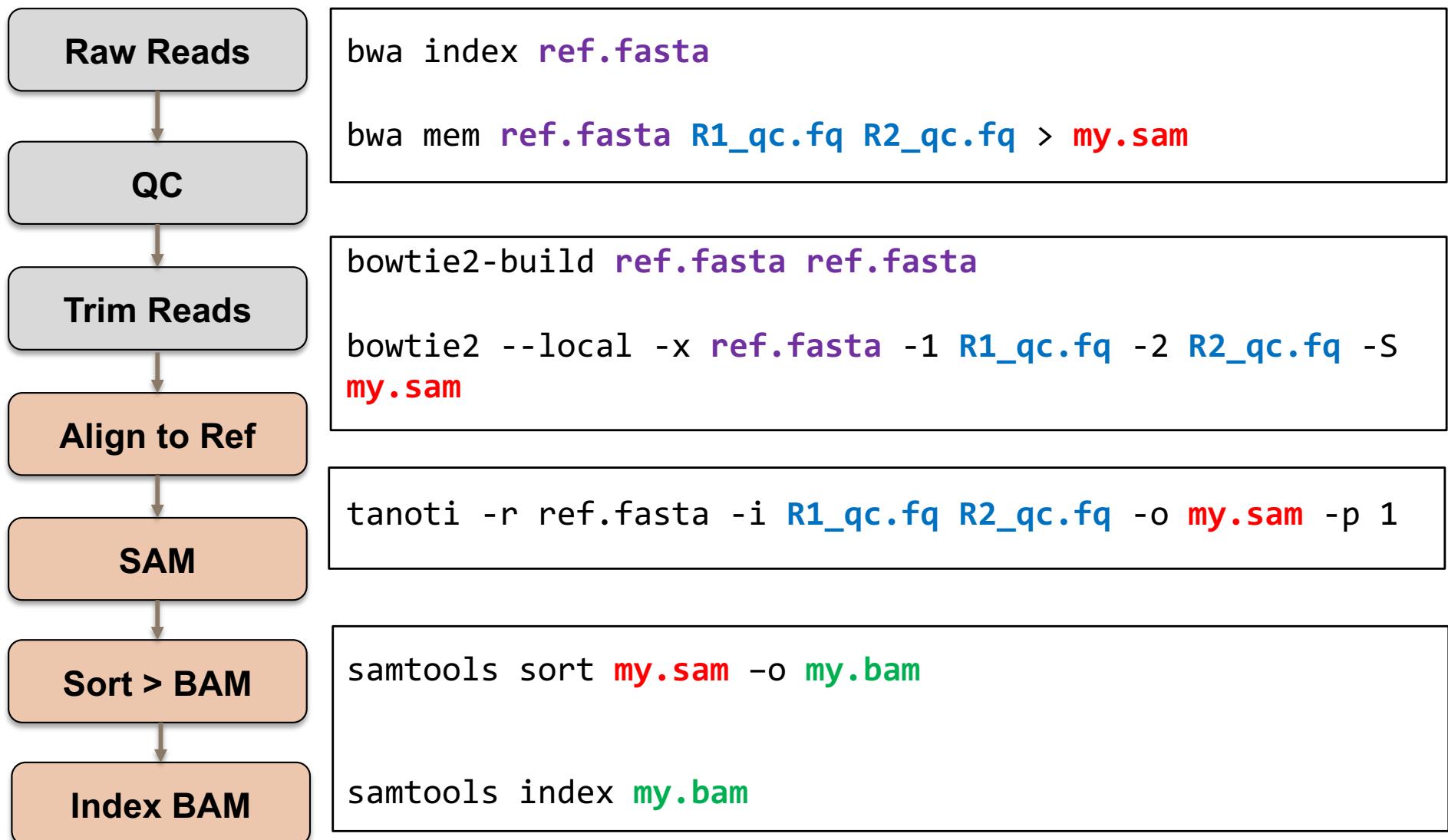
minimap2

- Starting out on a new virus - try a few aligners – not just about most reads aligned – consensus seq and variants

Mosaik

...

Ref alignment commands - different tools



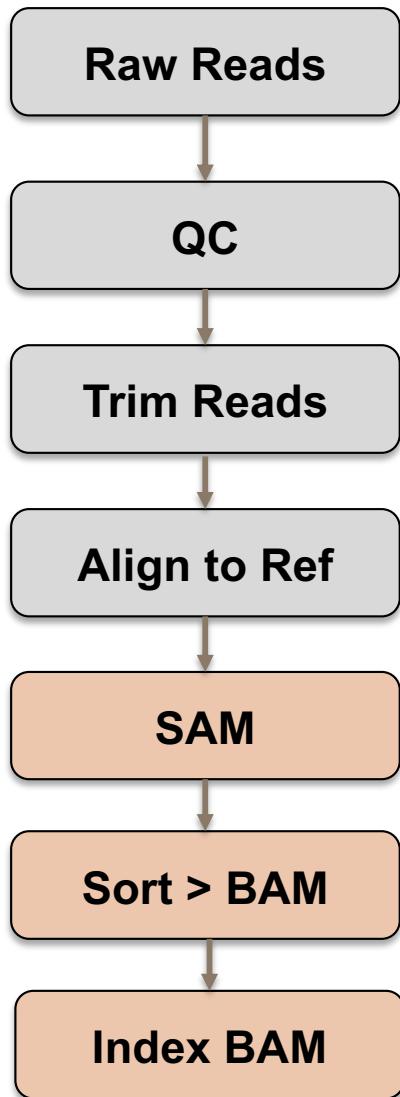
Which reference sequence?

- **Selecting a suitable reference sequence is an important step.**
 - If your reference is too divergent from your sample it 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

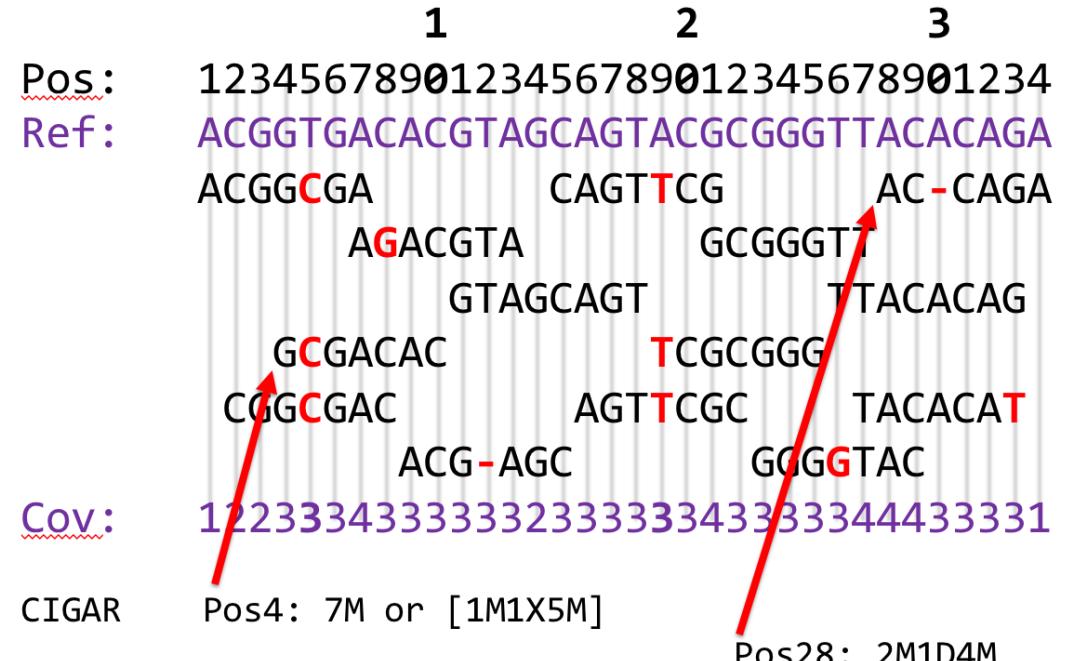
This is then sorted and converted to a BAM file, and indexed

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

```
samtools index my.bam
```

SAM files: Sequence Alignment MAP

- Virtually all aligners output results in **SAM** format
 - **Sequence Alignment/Map**
- 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

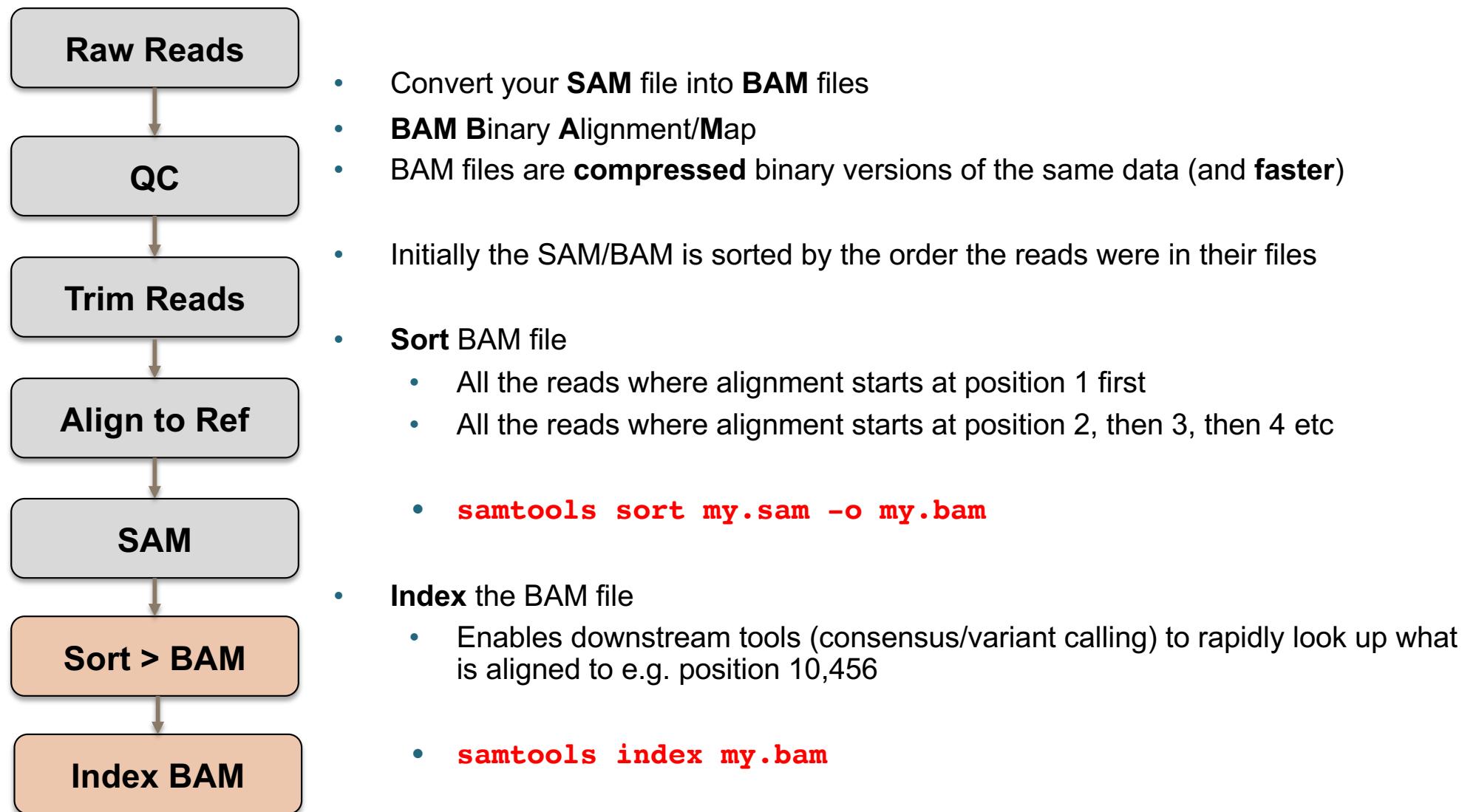


QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUALITY
Read3	10	MyRefSeq	28	52	2M1D4M				ACCAGA	IHGFFF
Read8	10	MyRefSeq	4	57	1M1X5M				GCGACAC	IIHHGG

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



SAM Flags – Mapped/Unmapped

- **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 field 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	IIHHGG

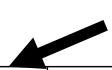
SAM Flags – Mapped/Unmapped

- 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

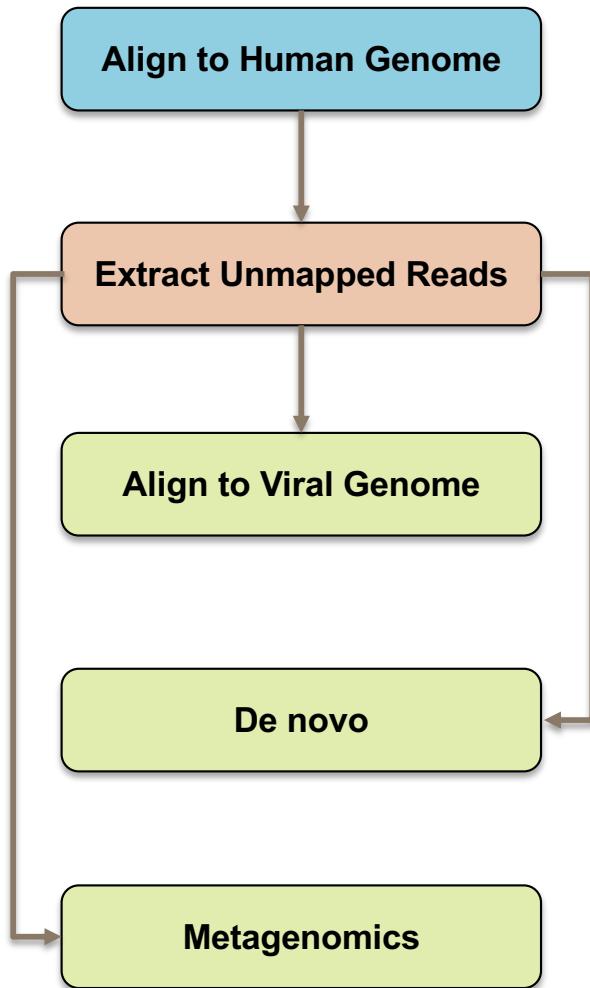
Typically, secondary/supplementary should be few for short RNA virus genome, but if lots it can indicate:

Repeat regions, Large deletions, Poor reference sequence

 SAM Flag = 2nd field of SAM file

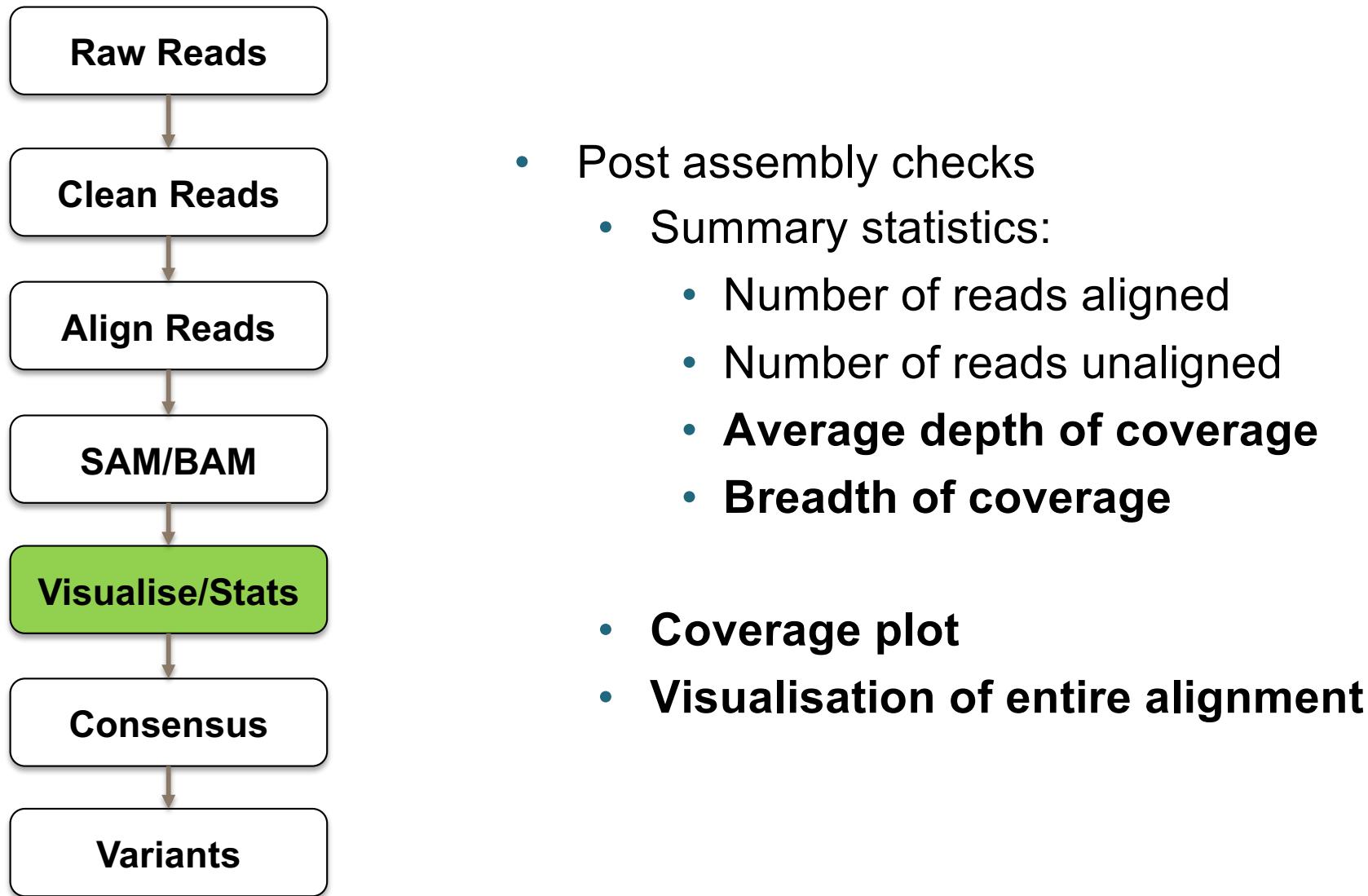
QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	QUALITY
Read3	10	MyRefSeq	28	52	2M1D4M				ACCAGA	IHGFFF
Read8	4	*	0	0					GCGACAC	IIHHGG

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



Coverage

	1	2	3
Pos:	1234567890123456789012345678901234		
Ref:	ACGGTGACACGTAGCAGTACGCGGGTTACACAGA		
	ACGG CGA	CAGT TCG	AC- CAGA
	A GACGTA		GC GGGTT
		GTAGCAGT	TTACACAG
	G CGACAC	T CGCGGG	
	CGG CGAC	AGT TCGC	TACACAT T
	ACG- AGC		GGGG TAC
Cov:	1223334333332333334333334443331		

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

Viral Reference Genome



Average coverage = 1

Breadth = 100%

Viral Reference Genome



Average coverage = 1

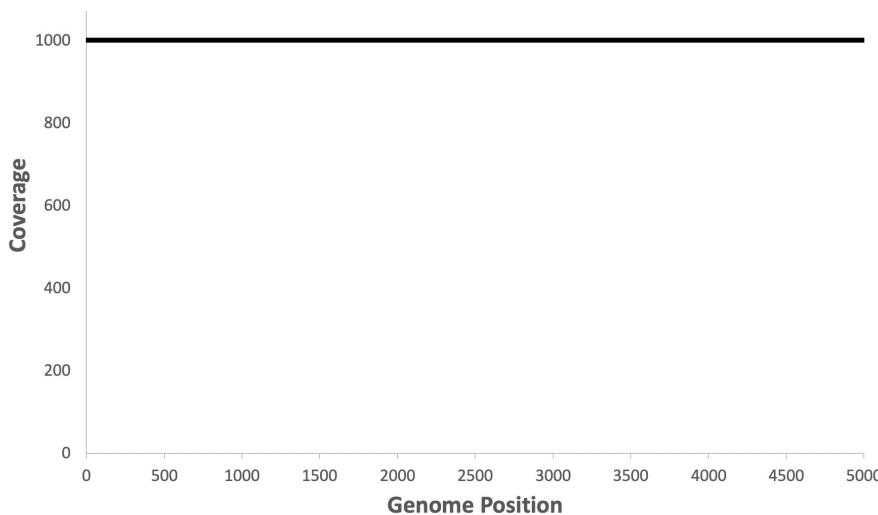
Breadth = 20%

Mode, Median, Quartiles would be different

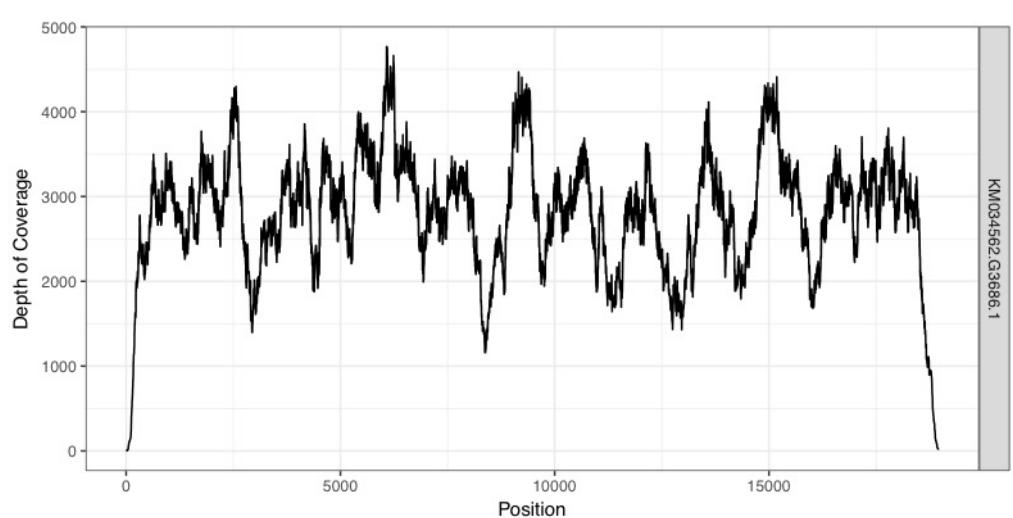
Perfect Coverage Plots

- 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

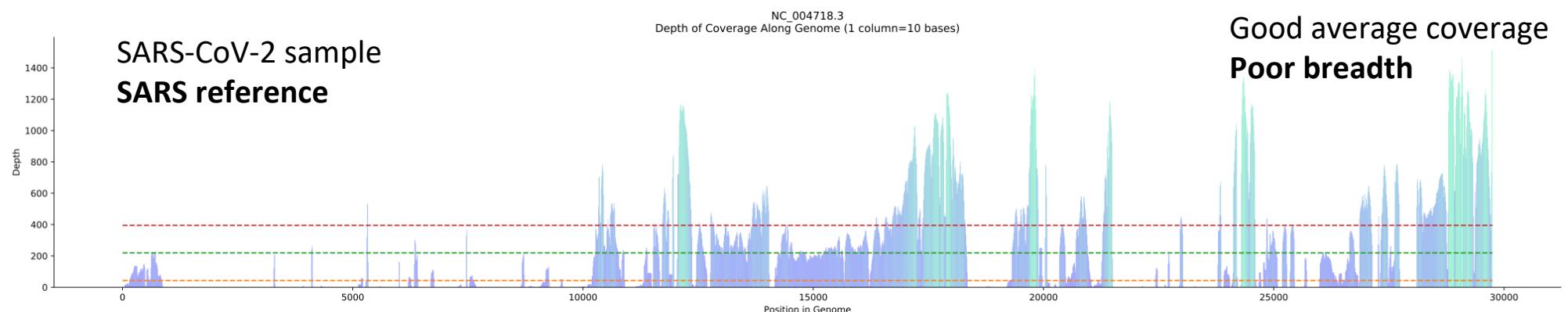


This is perfectly normal

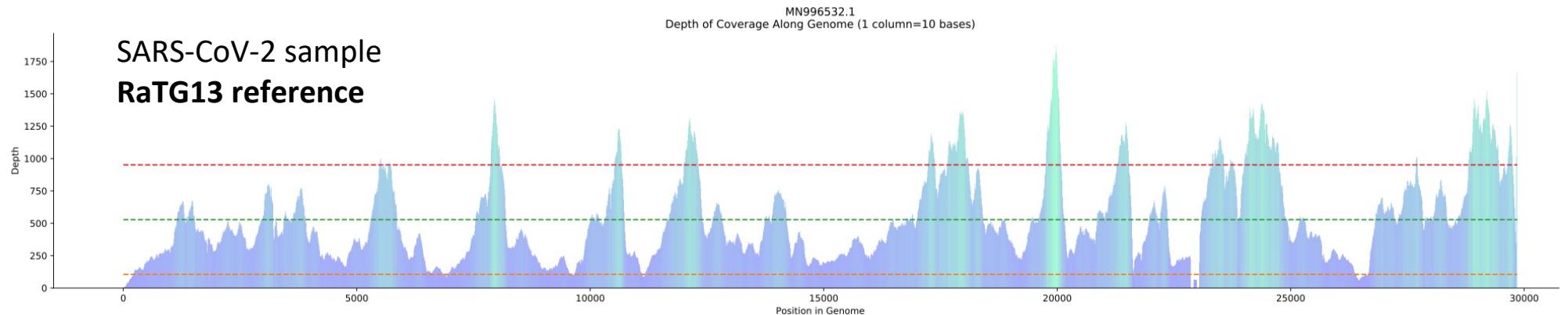


Coverage plots – bad reference

- 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



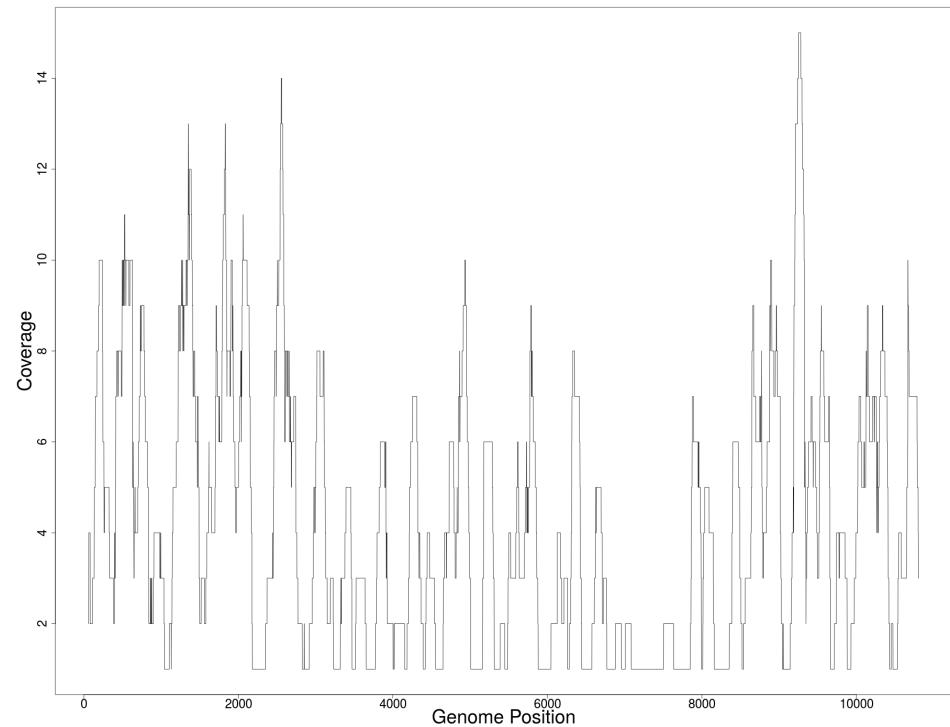
- Viruses can be very diverse – aligning to a different genotype/strain can give obscure results
- Align to different refs, genotype detection tools, **de novo assembly**



Coverage plots – low coverage

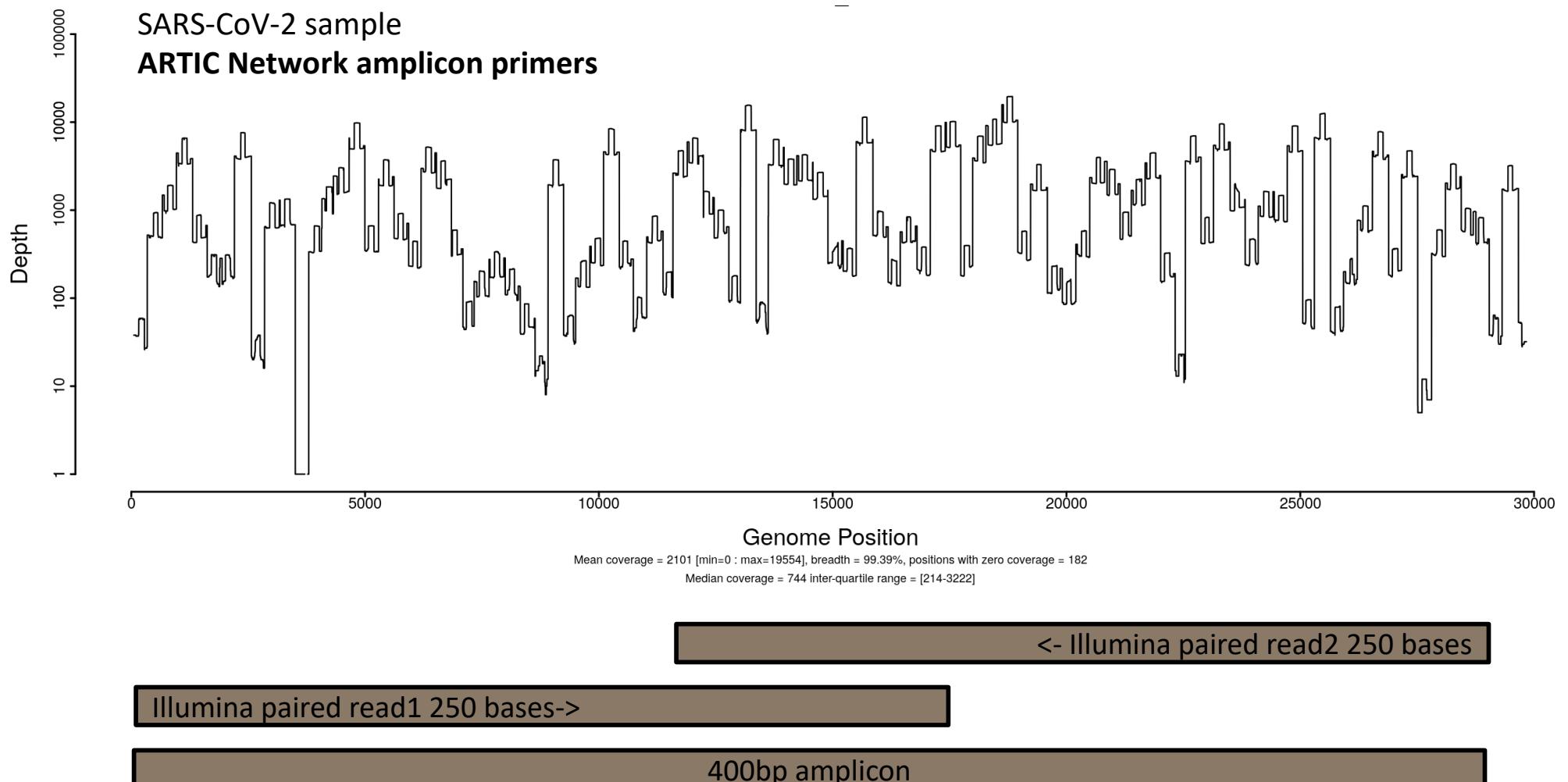
- Sometimes there is just not enough data present
 - Lower read trimming threshold
 - Just use the raw reads
 - Will be noisy
- Re-run the sample
 - Perhaps it was a bad run
 - Combine run data
- PCR amplification
- Bait capture

Louping ill virus sample



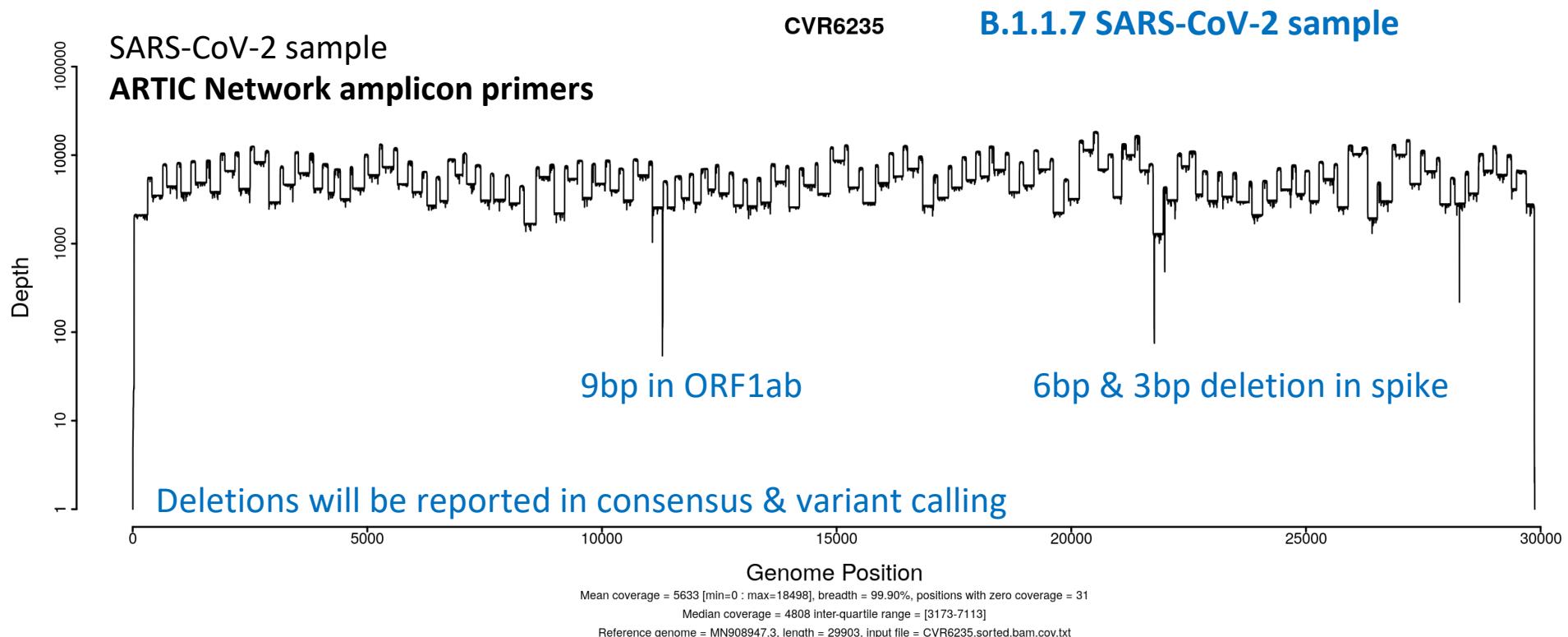
Coverage plots – amplicons

- Amplicon data can give step like plots



Coverage plots – deletions

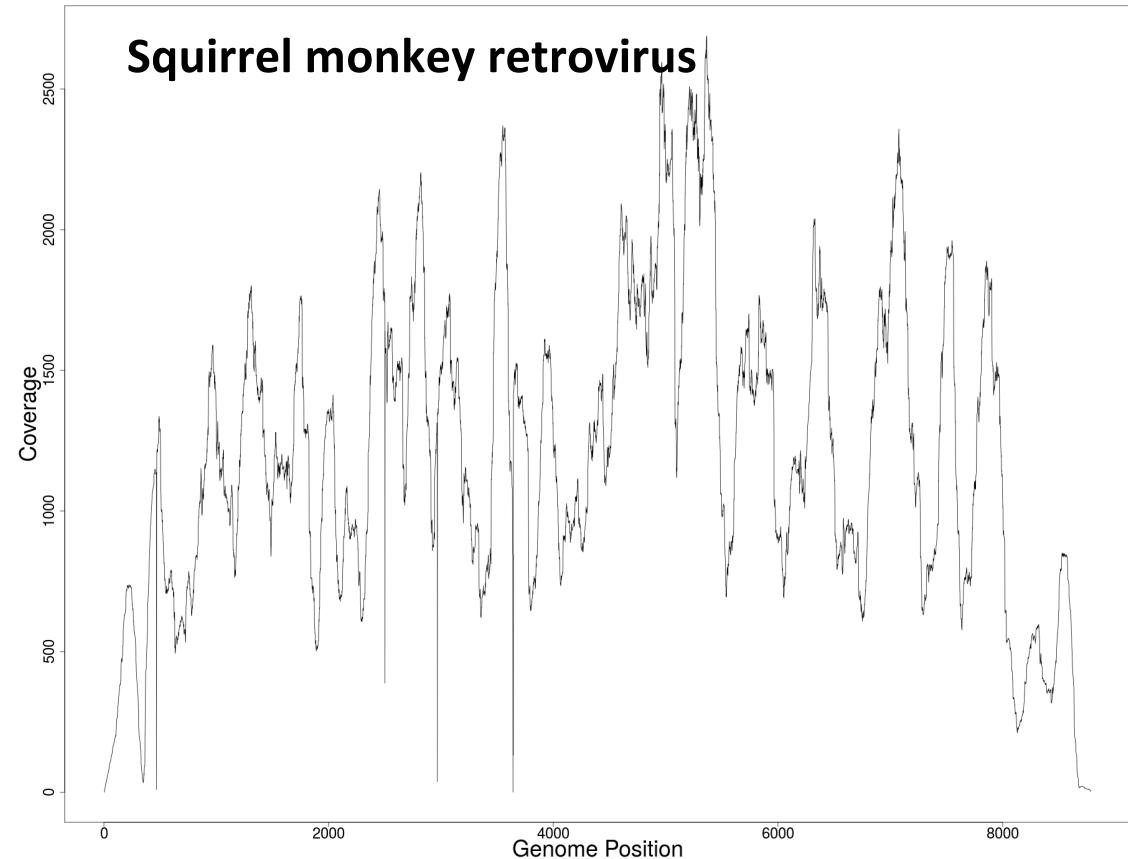
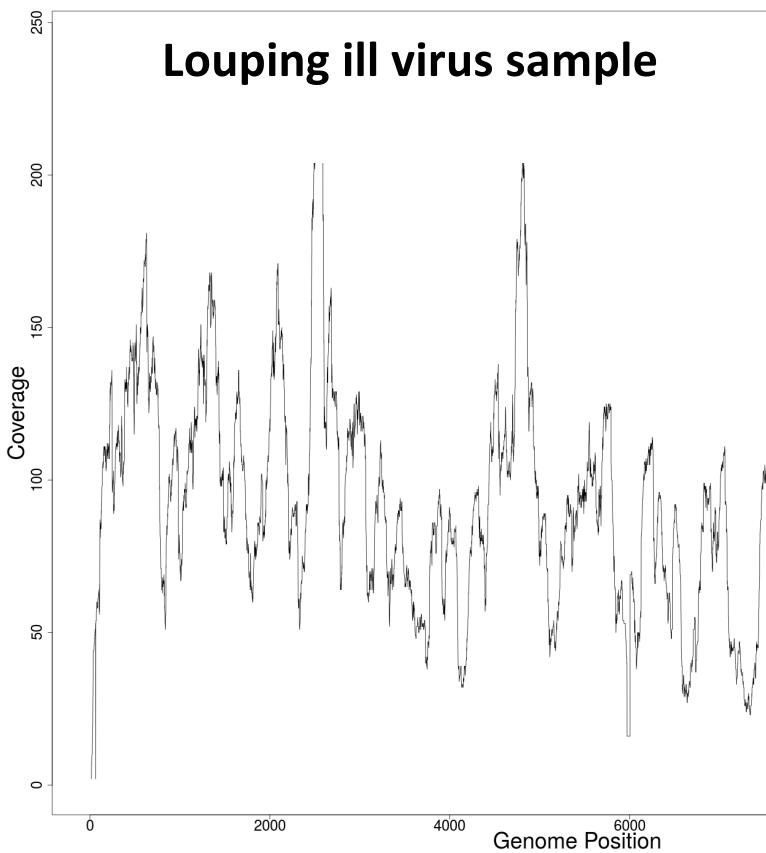
- Sudden drops in coverage at a small number of sites can indicate deletions with respect to the reference



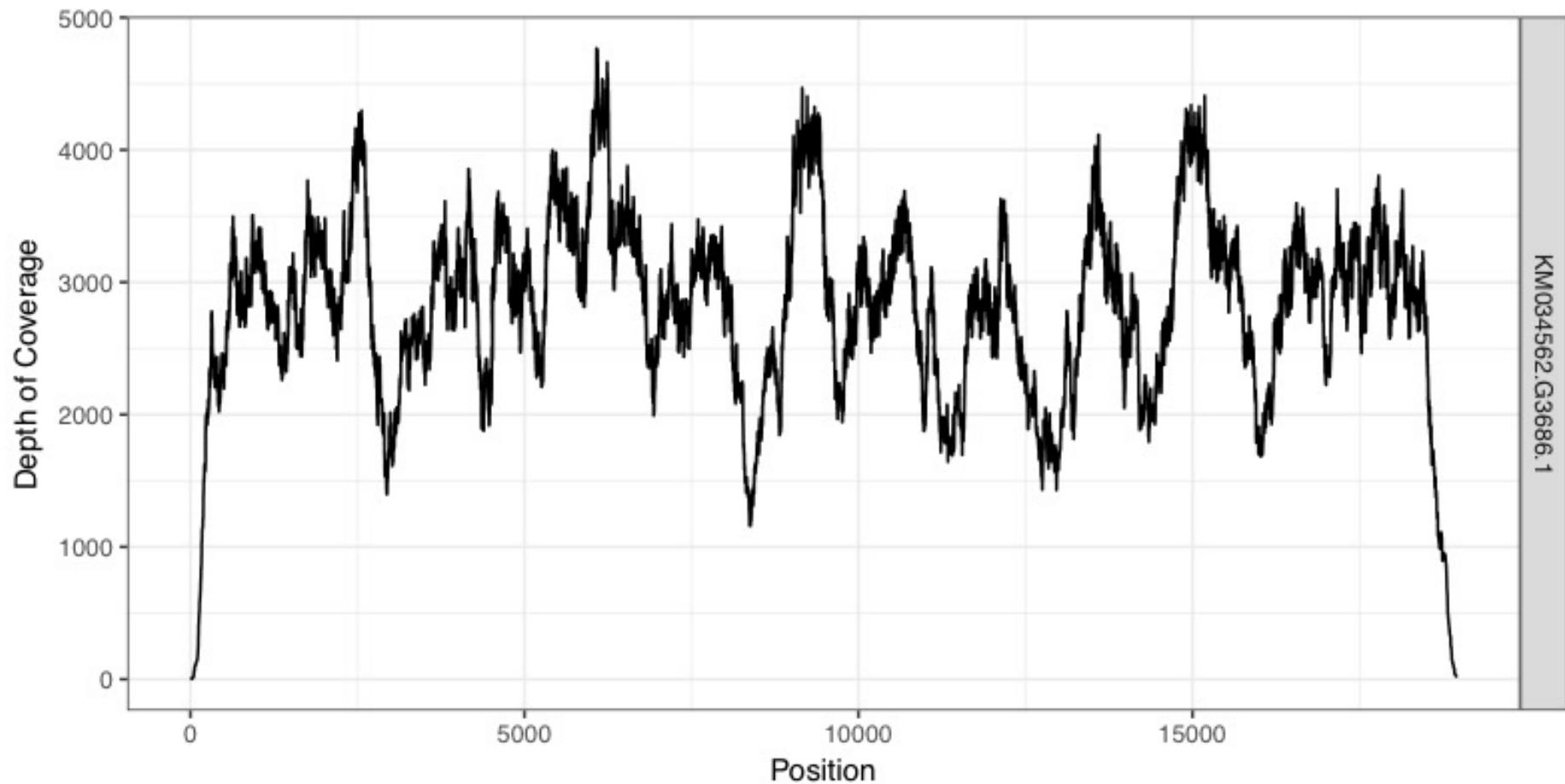
This is a log plot and noisy minion data – the deleted sites still have some coverage but this is nothing compared to the other sites

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)**



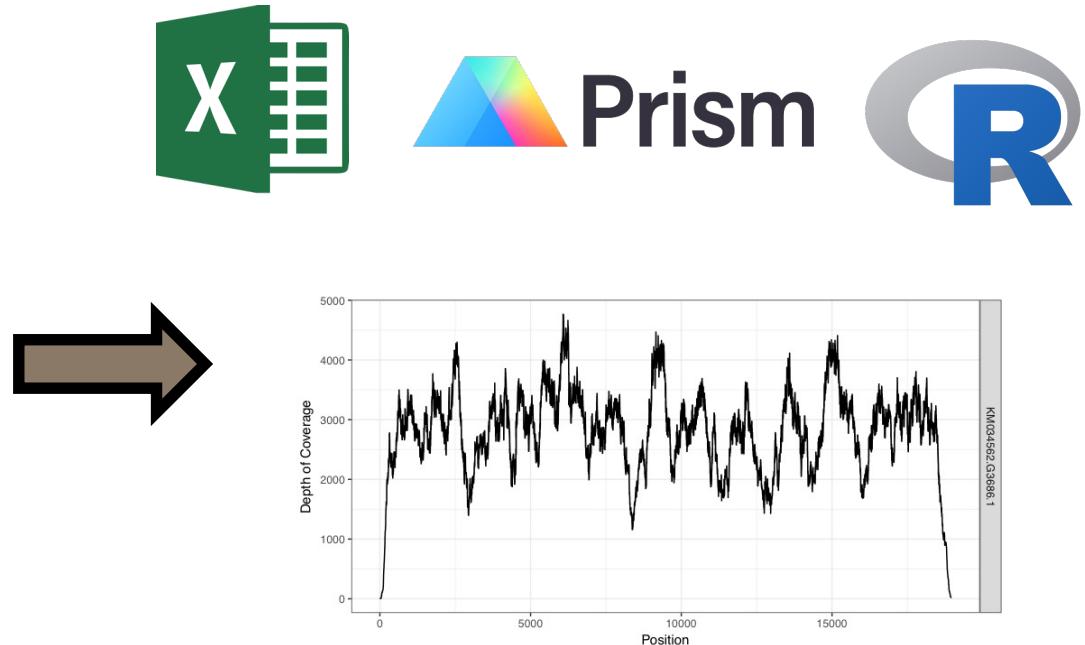
How do you create coverage plots?



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
...

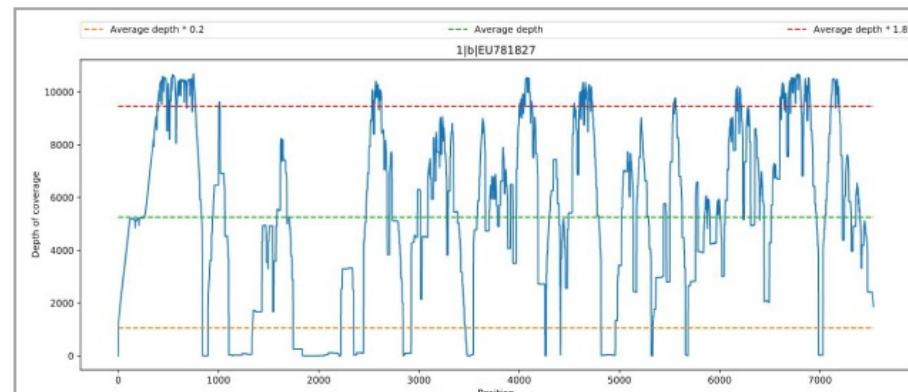


All chromosome will be reported in turn

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

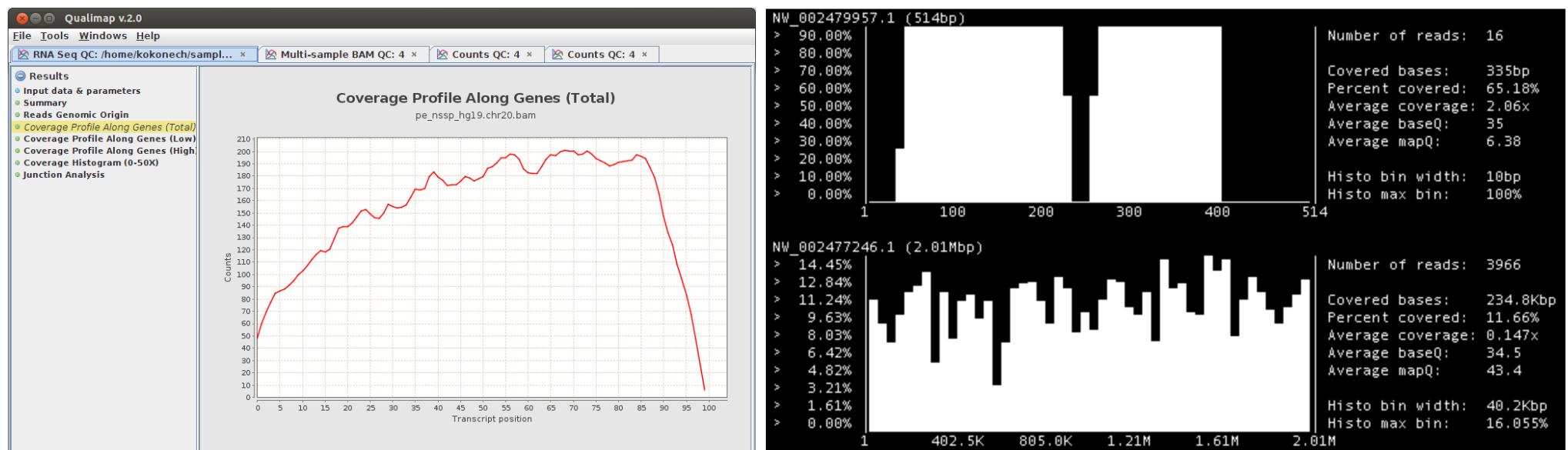
- weeSam is program that can give you information on breadth and depth of coverage as well as generate a coverage plot automatically
- weeSAM --bam 1a.bam --html 1a
- 1a_html_results/1a.html

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

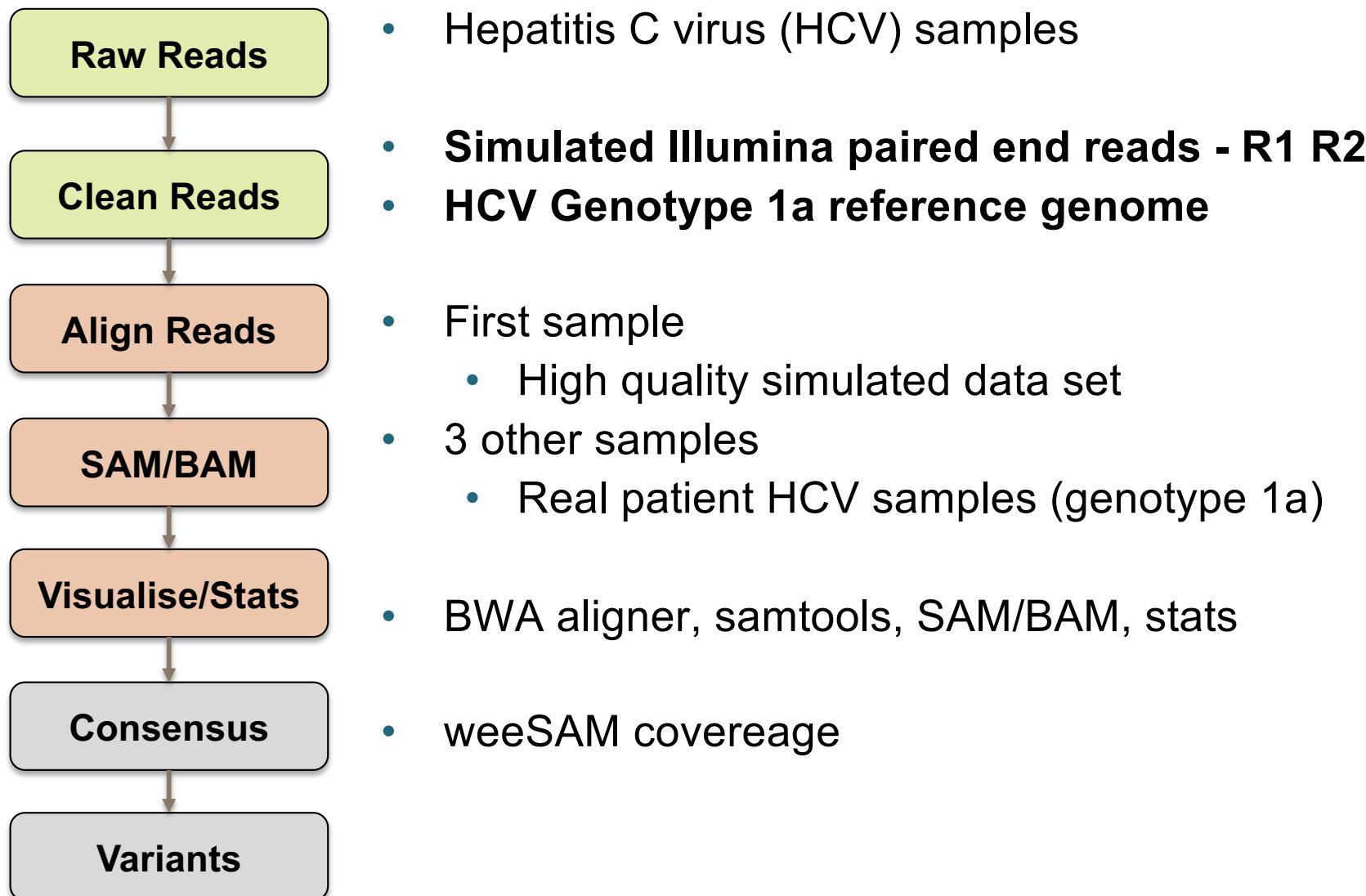


Other tools

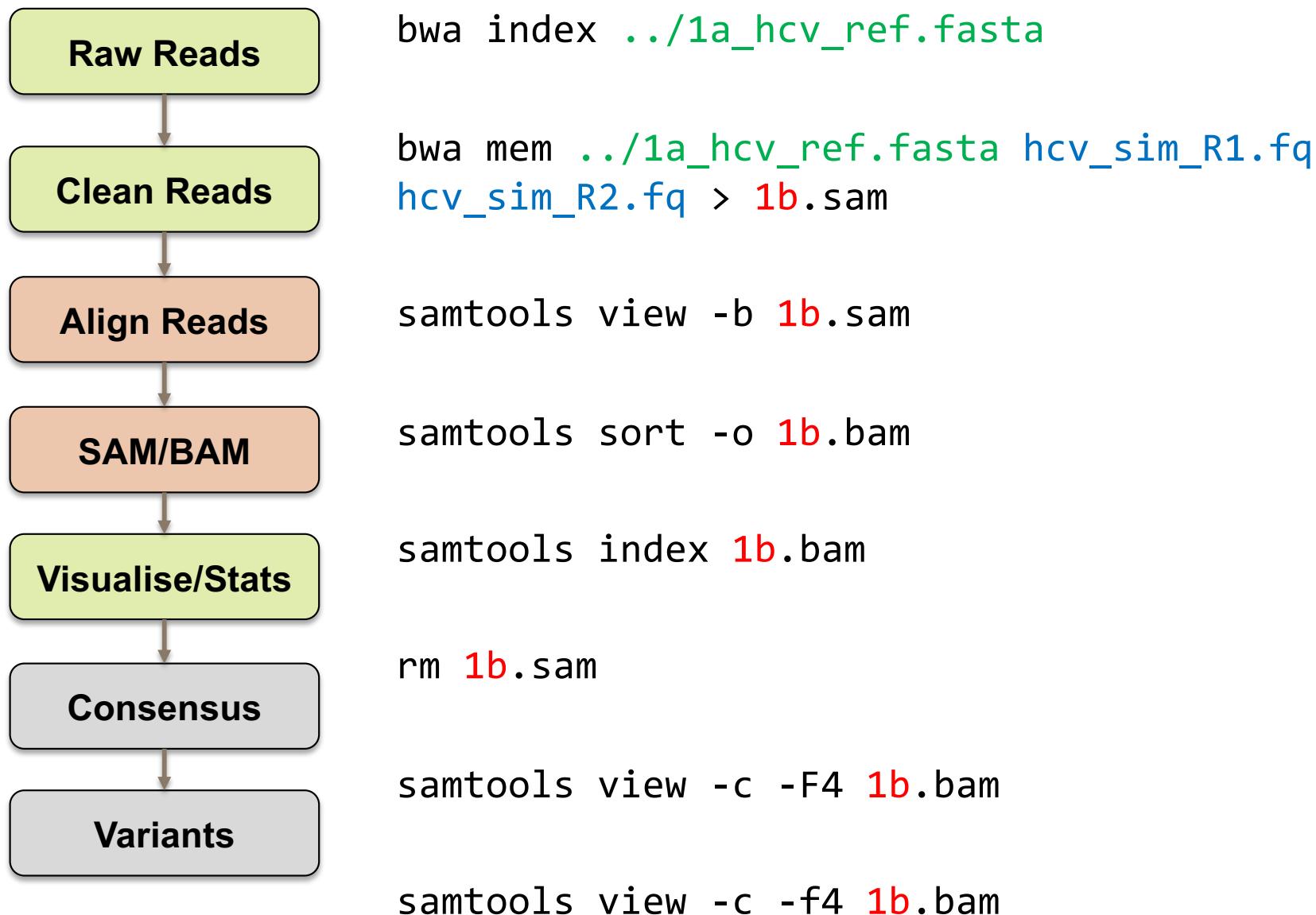
- Qualimap: <http://qualimap.conesalab.org>
- bamCov – <https://github.com/fbreitwieser/bamcov>



Practical



Practical – HCV_SIM commands – adapt for another sample

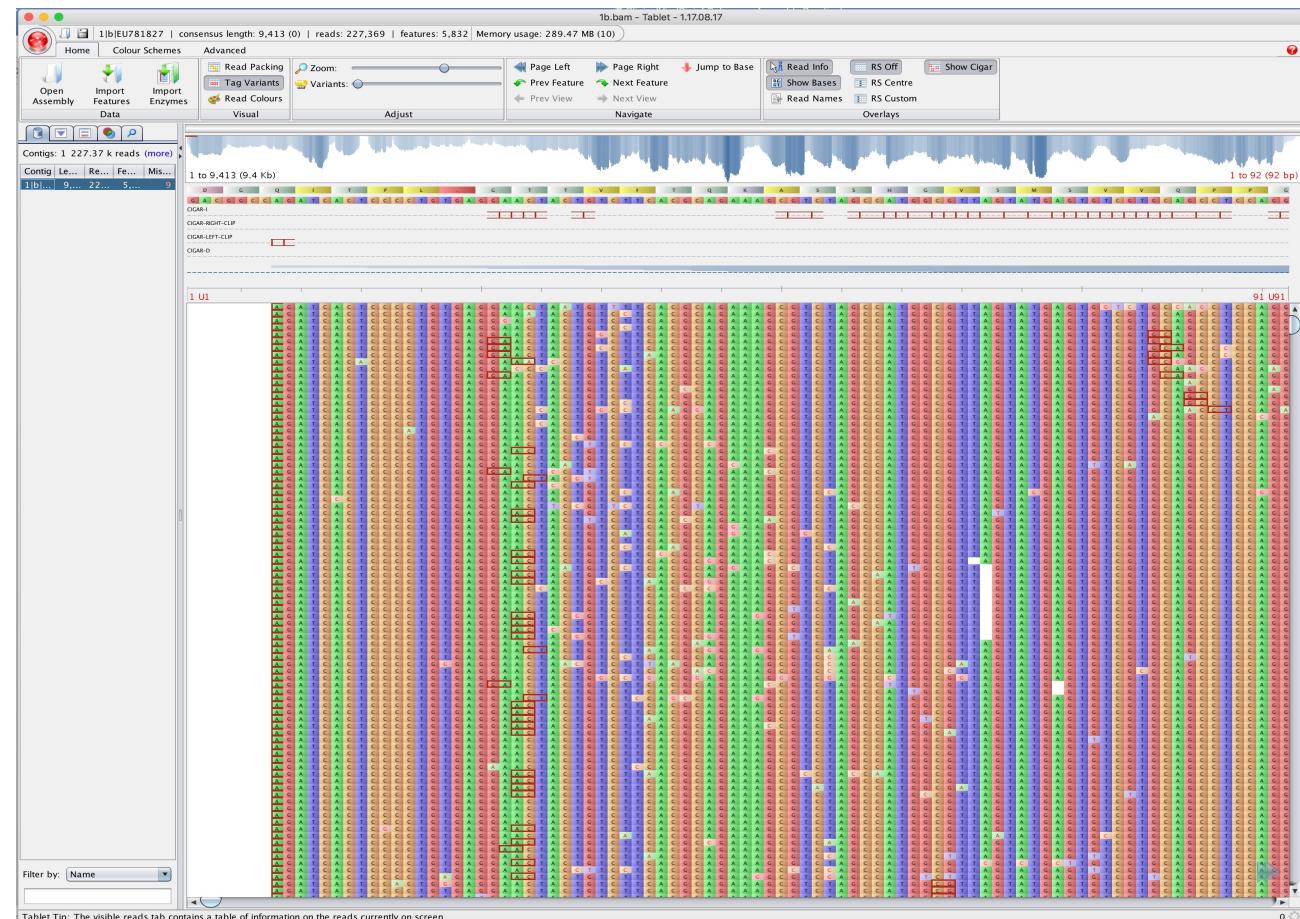


The End ... Tablet

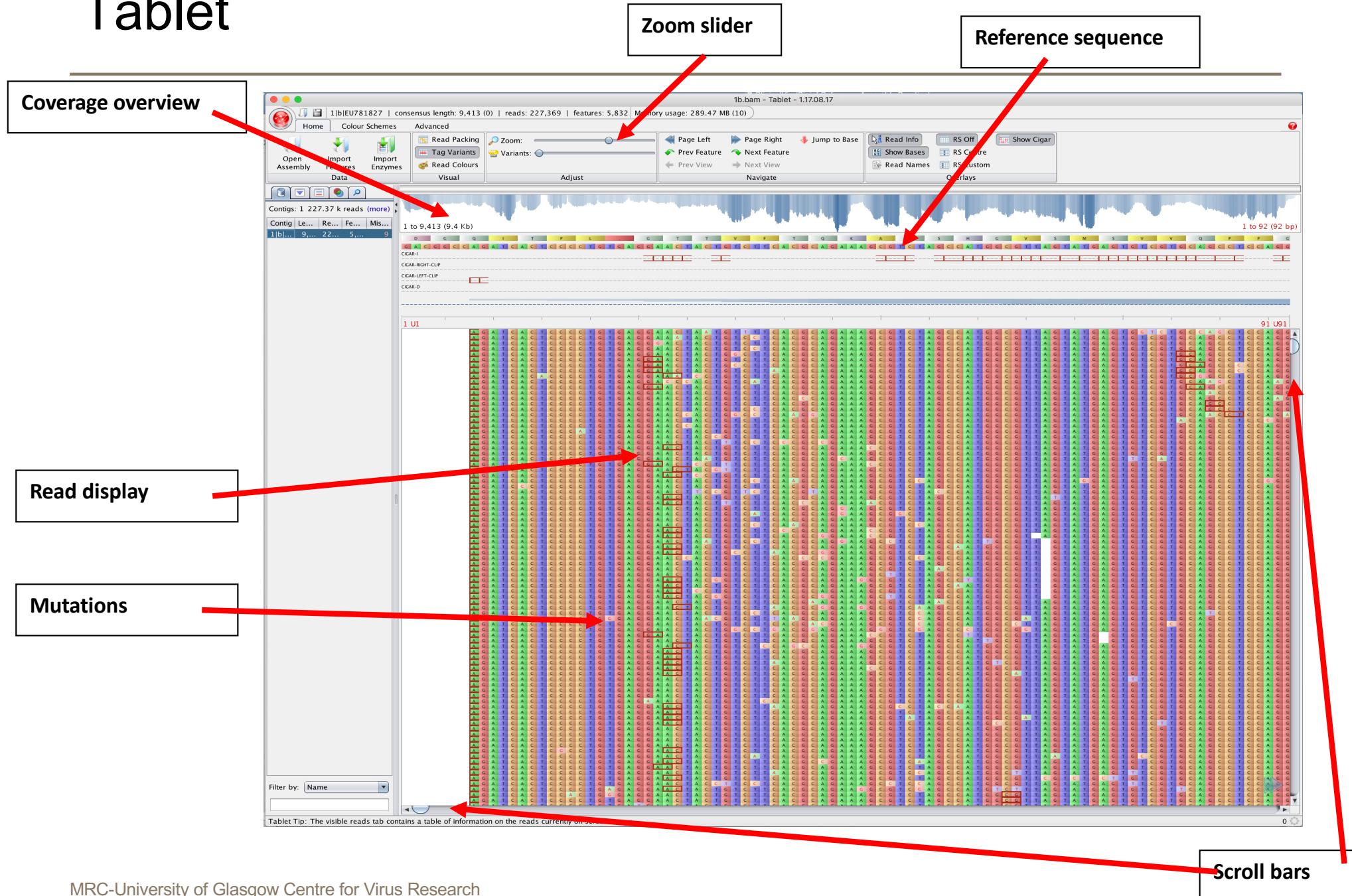
- Tablet demo if time later on

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



Tablet

