Doherty Institute
THE UNIVERSITY OF MELBOURNE
The Royal Melbourne Hospital

A joint venture between The University of Melbourne and The Royal Melbourne Hospital

Lecture 5 - Quality control for consensus sequences

Dr Kristy Horan

Objectives

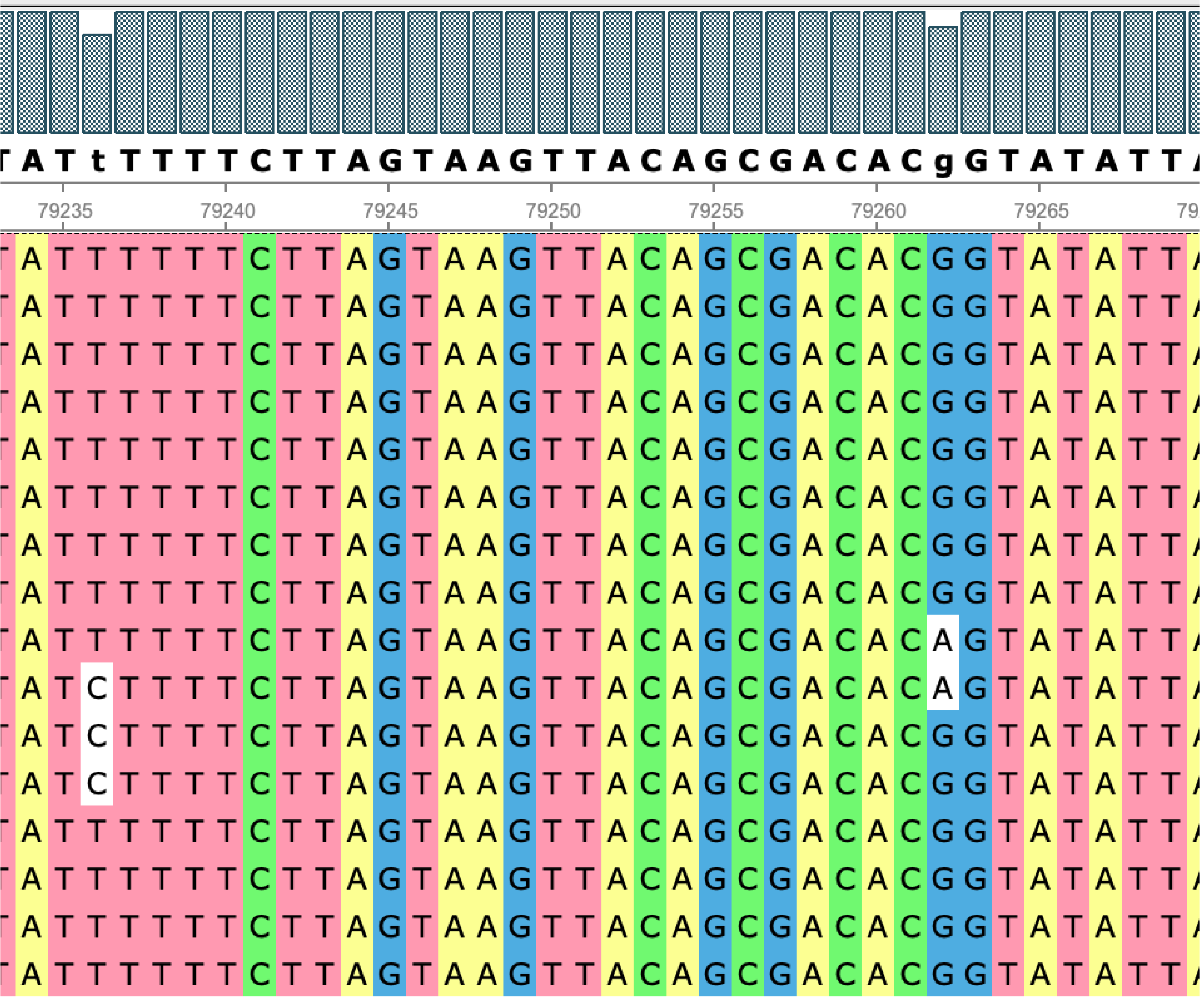
- Understand common quality control metrics used on viral consensus sequences

Refresher: Consensus Assembly

The output of the process of generating a **single representative sequence** from a mapped set of reads to a reference sequence. In other cases, from a set of similar sequences that are also multiple sequence aligned.

Consensus sequence

Majority nucleotide
"CONSENSUS"



```
TATtTTTTCTTAGTAAGTTACAGCGACACgGTATATTA
  79235     79240     79245     79250     79255     79260     79265     79
TATTTTTTCTTAGTAAGTTACAGCGACACGGTATATTA
TATTTTTTCTTAGTAAGTTACAGCGACACGGTATATTA
TATTTTTTCTTAGTAAGTTACAGCGACACGGTATATTA
TATTTTTTCTTAGTAAGTTACAGCGACACGGTATATTA
TATTTTTTCTTAGTAAGTTACAGCGACACGGTATATTA
TATTTTTTCTTAGTAAGTTACAGCGACACGGTATATTA
TATTTTTTCTTAGTAAGTTACAGCGACACGGTATATTA
TATTTTTTCTTAGTAAGTTACAGCGACACGGTATATTA
TATTTTTTCTTAGTAAGTTACAGCGACACAGTATATTA
TATCTTTTCTTAGTAAGTTACAGCGACACAGTATATTA
TATCTTTTCTTAGTAAGTTACAGCGACACGGTATATTA
TATCTTTTCTTAGTAAGTTACAGCGACACGGTATATTA
TATTTTTTCTTAGTAAGTTACAGCGACACGGTATATTA
TATTTTTTCTTAGTAAGTTACAGCGACACGGTATATTA
TATTTTTTCTTAGTAAGTTACAGCGACACGGTATATTA
TATTTTTTCTTAGTAAGTTACAGCGACACGGTATATTA
```

Minor differences

Minor differences

Challenges in Genome Assembly

Pre-assembly

- Read Filtering: Exclude low-quality reads and adapter sequences to improve assembly accuracy.
- Error Correction: Use tools for error correction, such as k-mer-based correction or consensus-based correction
- Thresholds: minimum frequency threshold, minimum depth to call consensus

Challenges in Genome Assembly

Alignment or read mapping

- Alignment-based Filtering - Align reads and/or consensus sequences to a reference genome to identify and remove erroneous reads or artifacts or gaps
- Variant Calling - Identify and filter out variants that are likely sequencing errors rather than true genetic variation.
- Depth of Coverage Analysis - Assess the depth of coverage across the genome to identify regions with low coverage orpotential errors.

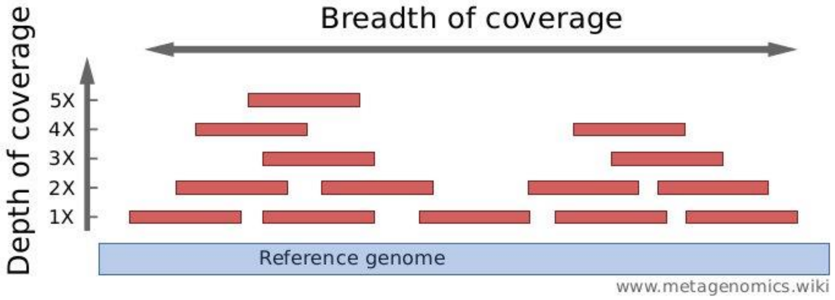
Organism Specific considerations

Genome size: How many base pairs is it?
Genome organization: How many open reading frames (ORF)? What is their orientation and is it correct?
Repeat regions: Are there known repeat regions? What are their positions? If reads don't span this region (i.e., region covered by single reads) the assembly or consensus sequence over these regions should not be trusted.
Low or high GC content areas: Are there genomic regions with low or high GC content? (more relevant to bacteria but important to keep in the back of your mind.

Other QC metrics to consider

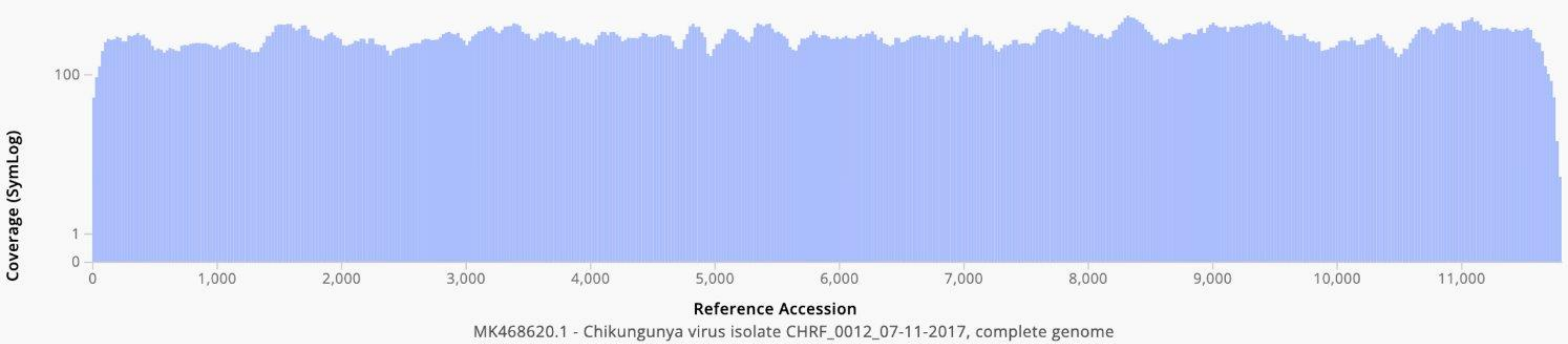
- **% Genome Called** - Refers to the percentage of the genome meeting thresholds for calling consensus bases. The closer this number is to 100%, the better.
  - Note: amplicon schemes do not cover the complete genome, these must be factored in.
- **Number of SNPs** - High number of SNPs could indicate issues such as wrong reference selection
- **Clustering of SNPs** – if SNPs cluster in one region (compared to spread across genome)
- **Ambiguous bases** - If multiple sequencing reads support *more* than one nucleotide at a given site, those sites will be designated with an IUPAC ambiguity code.
- **Mapped reads** - Refers to the total number of reads that mapped to the reference genome.

Coverage plots

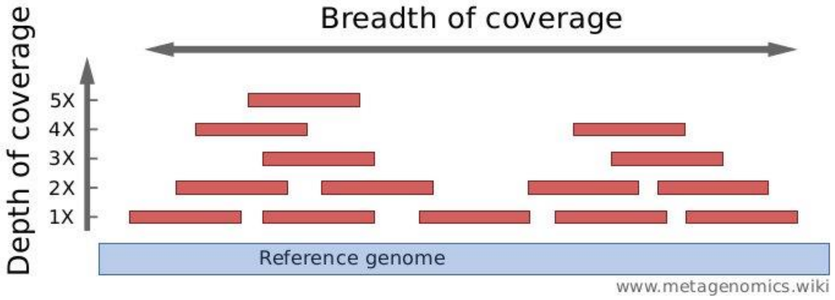


www.metagenomics.wiki

- **Coverage plot –** Coverage plots are good for looking at the breadth and depth of the assembled genome



Coverage plots



- Depth:
  - Per-base coverage is the average number of times a base of a genome is sequenced. It is often expressed as 1X, 2X, 50X times coverage.
- Breadth
  - The percentage of bases of a reference genome that are covered with a certain depth.
- Percent Coverage > X (not often used)
  - Combines depth and breadth greater than X
  - The breadth (%) greater than X, e.g. 89% genome covered at > 100X

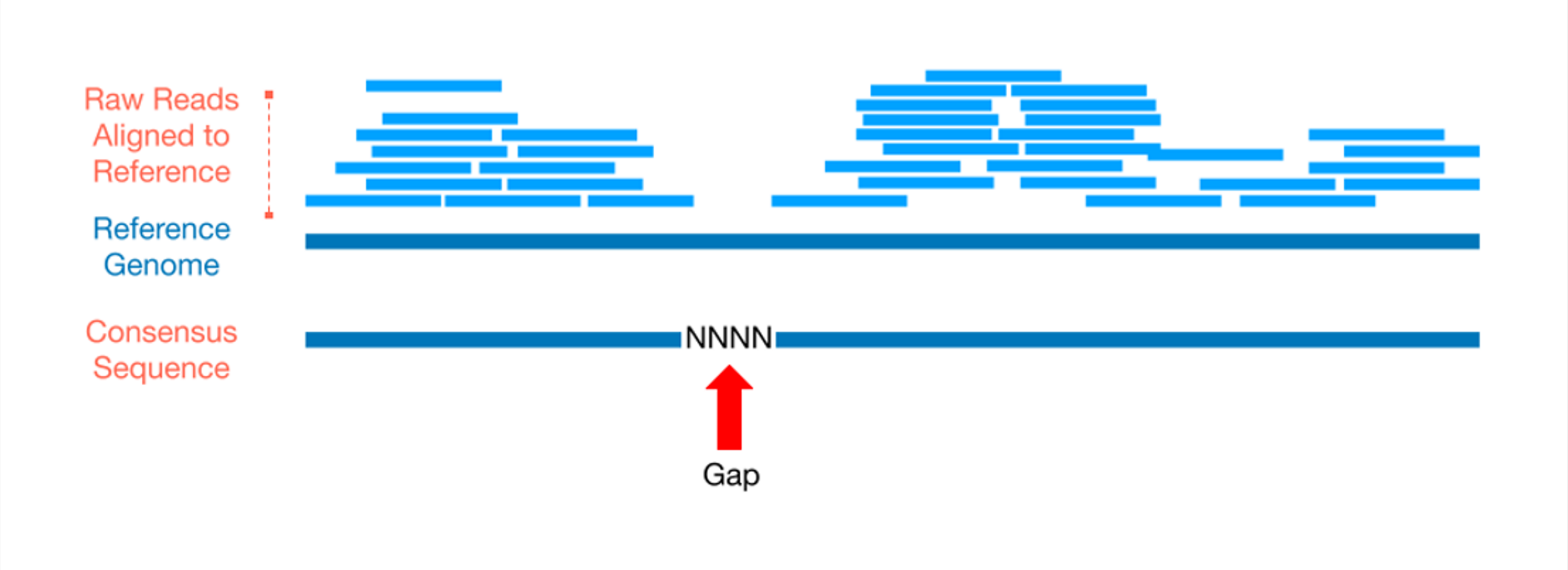
Thresholds – task dependent

Use different thresholds for different objectives

- Speciation or MPXV detection
  - genome coverage of >50%
- Phylogenetics – Whole genome
  - Recommended: >80-90% genome coverage
- Clade/Lineage analysis
  - Use Nextclade metrics

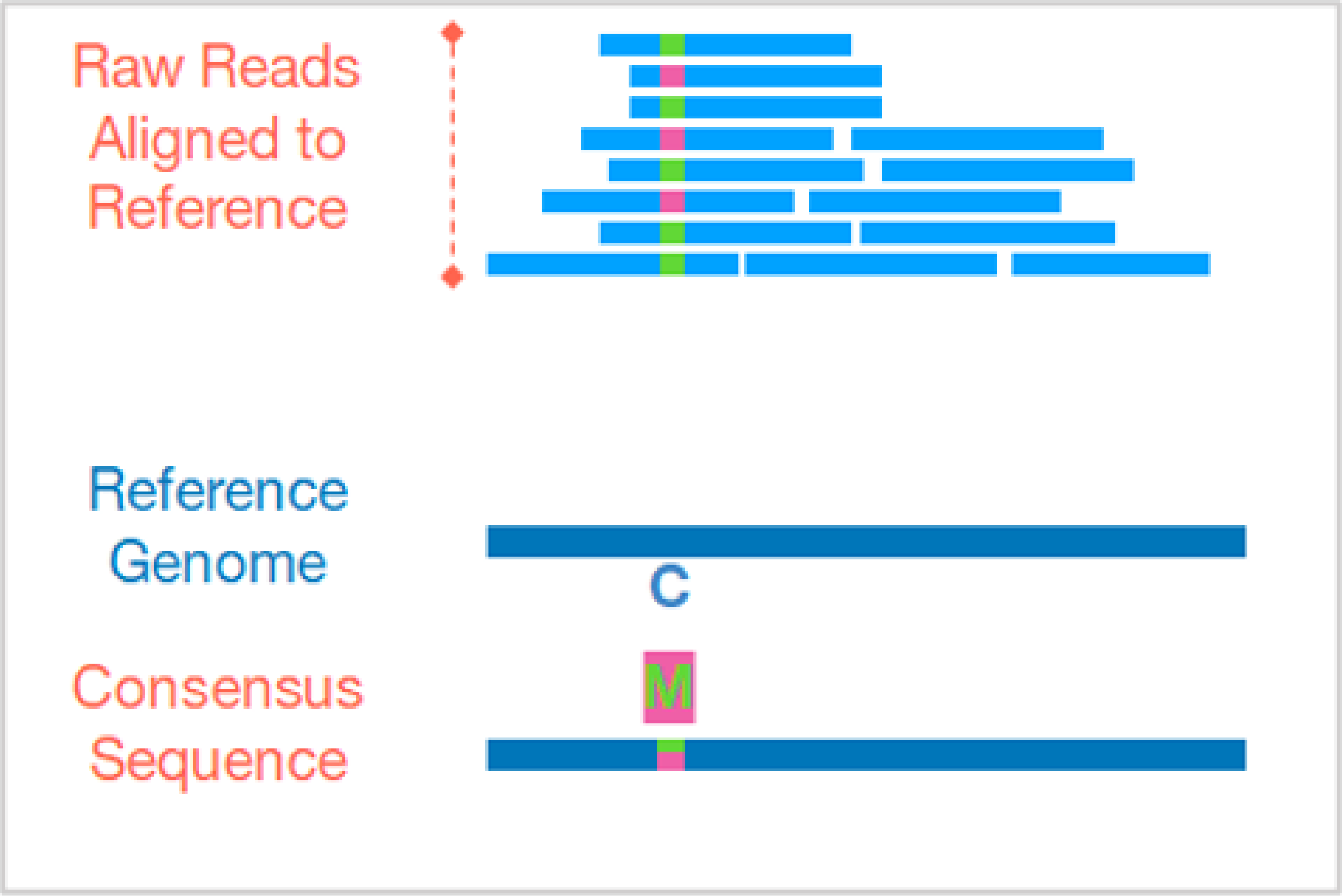
Check Read Alignments

MPXV Pipeline uses 20 reads as minimum coverage



How to tell the difference between gaps and deletions?

Check Read Alignments



Ambiguous bases can indicate a mixture of multiple viruses or in some viruses (SC2) recombination

Questions?