SARS-CoV-2 bioinformatics Training

Bioinformatics Quality Control

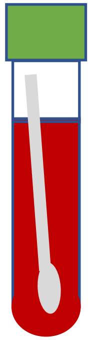
George Githinji

11th October 2020

KEMRI | Wellcome Trust

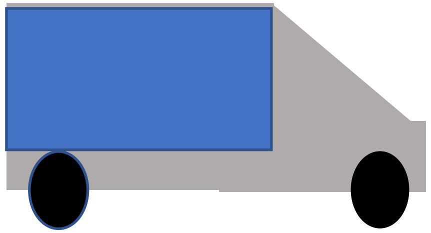
Sample preparation

Sample collection



- Time
- Proper sampling

Sample transportation



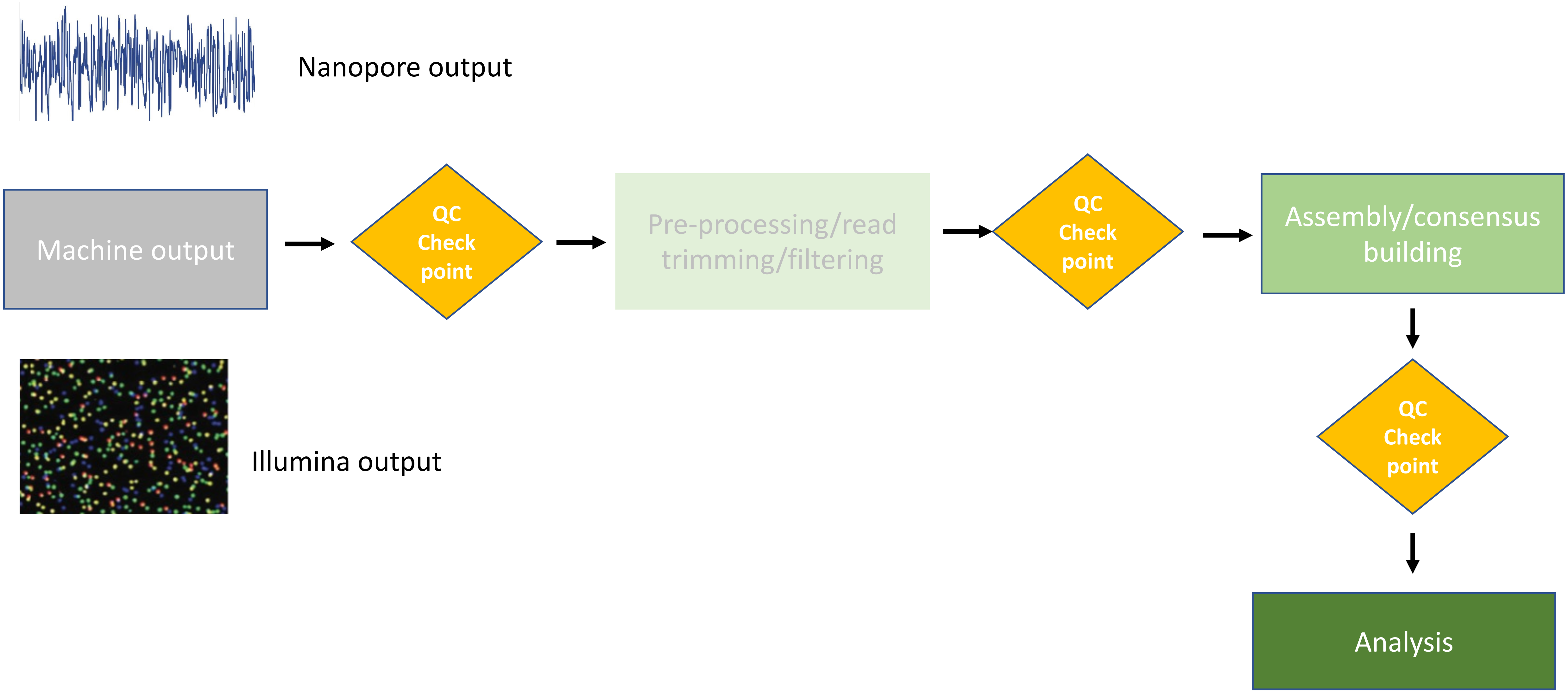
- Time
- Cold chain

Laboratory methods and storage

SARS-CoV-2 Detection Method

- Metagenomics approaches
- Amplicon based approaches
  - Pooled amplicon-based methods
- Sequencing platforms
  - ONT
  - Illumina

NGS process control and Quality check checkpoints



Nanopore output

Illumina output

Machine output

QC Check point

Pre-processing/read trimming/filtering

QC Check point

Assembly/consensus building

QC Check point

Analysis

What quality measures are we interested with?

- Degree of contamination
- Genome completeness
  - Proportion on non-N bases
- Sequence accuracy
  - Per base accuracy
  - Concensus accuracy

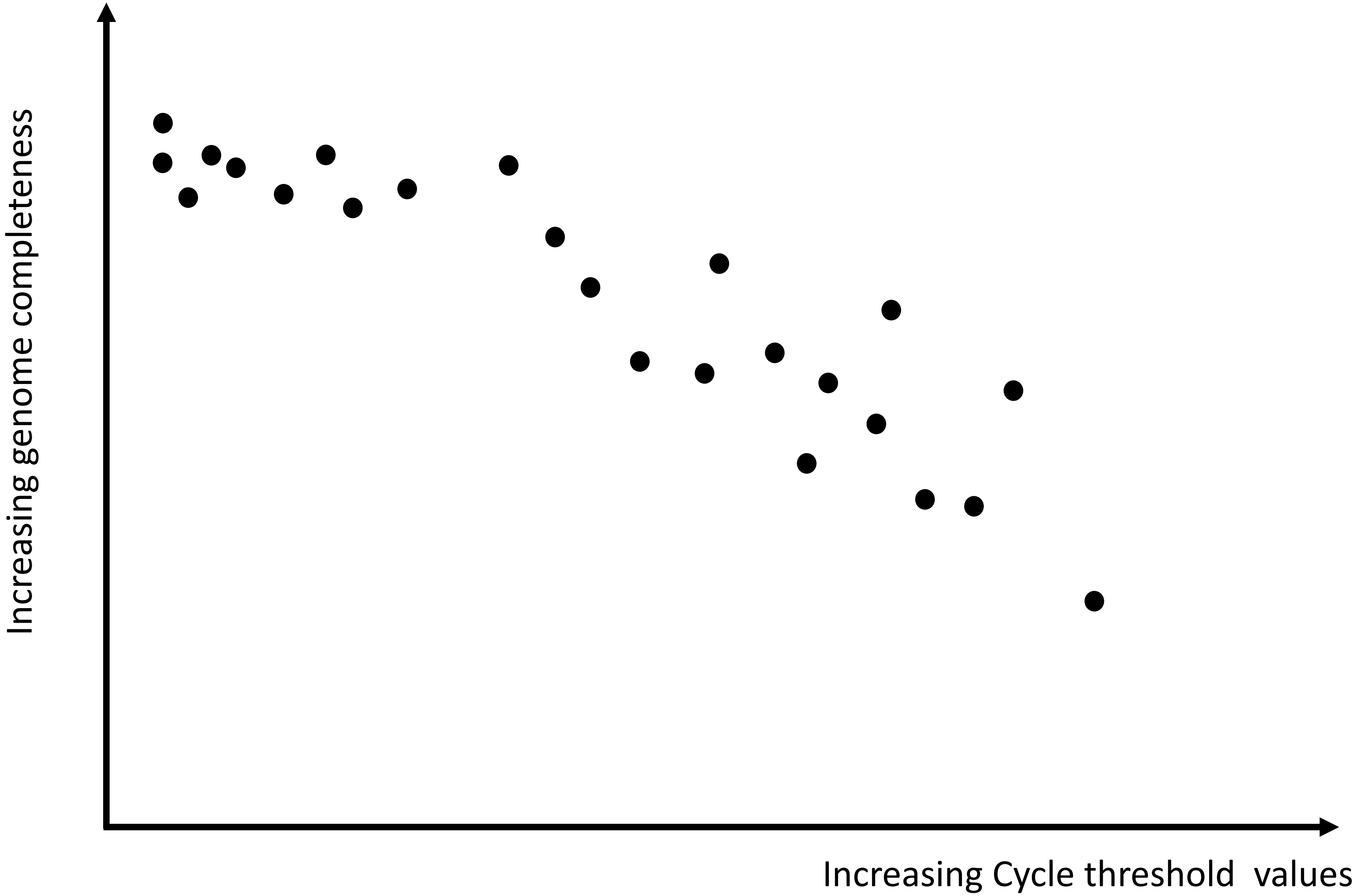
Why do we care?

- Contamination will read to misinterpretation of the results
  - For SARS-CoV-2 this might have serious consequences on policy

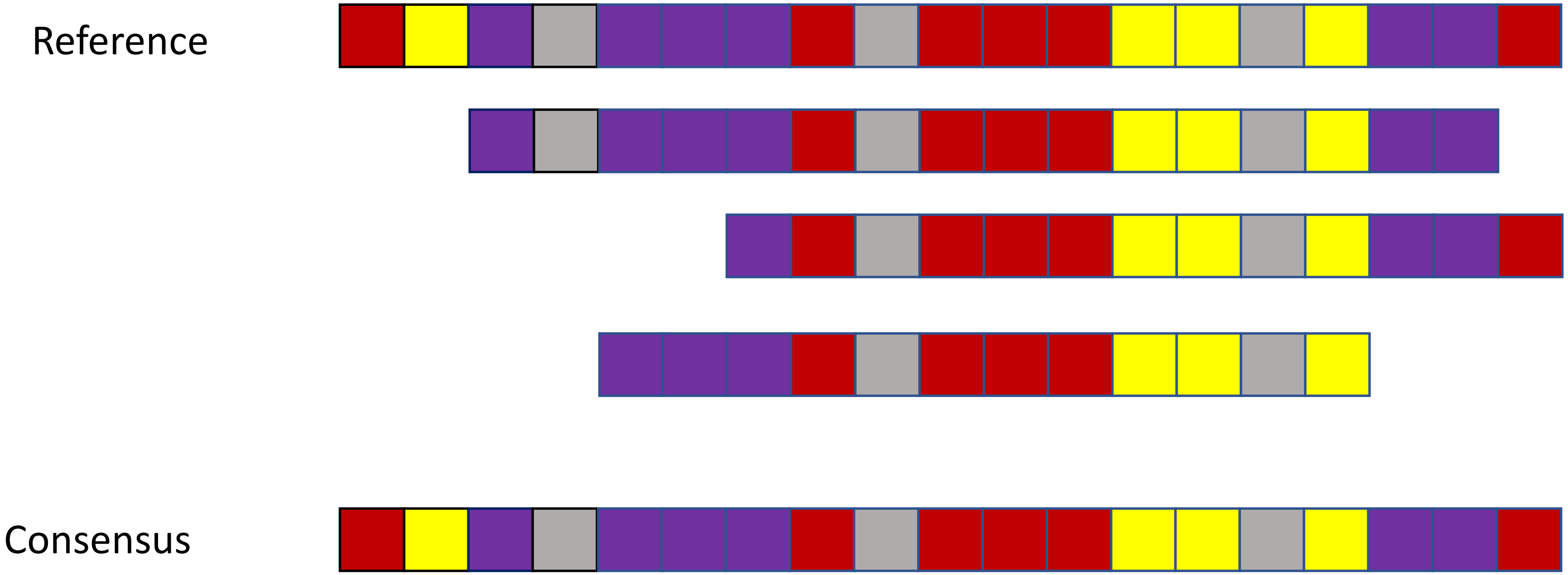
- Incomplete genomes are difficult to analyse
  - Lineage misassignment
  - Lack of phylogenetic signal

- Might be difficult to submit to public repositories
  - Genbank
  - GISAID

Viral load and genome completeness



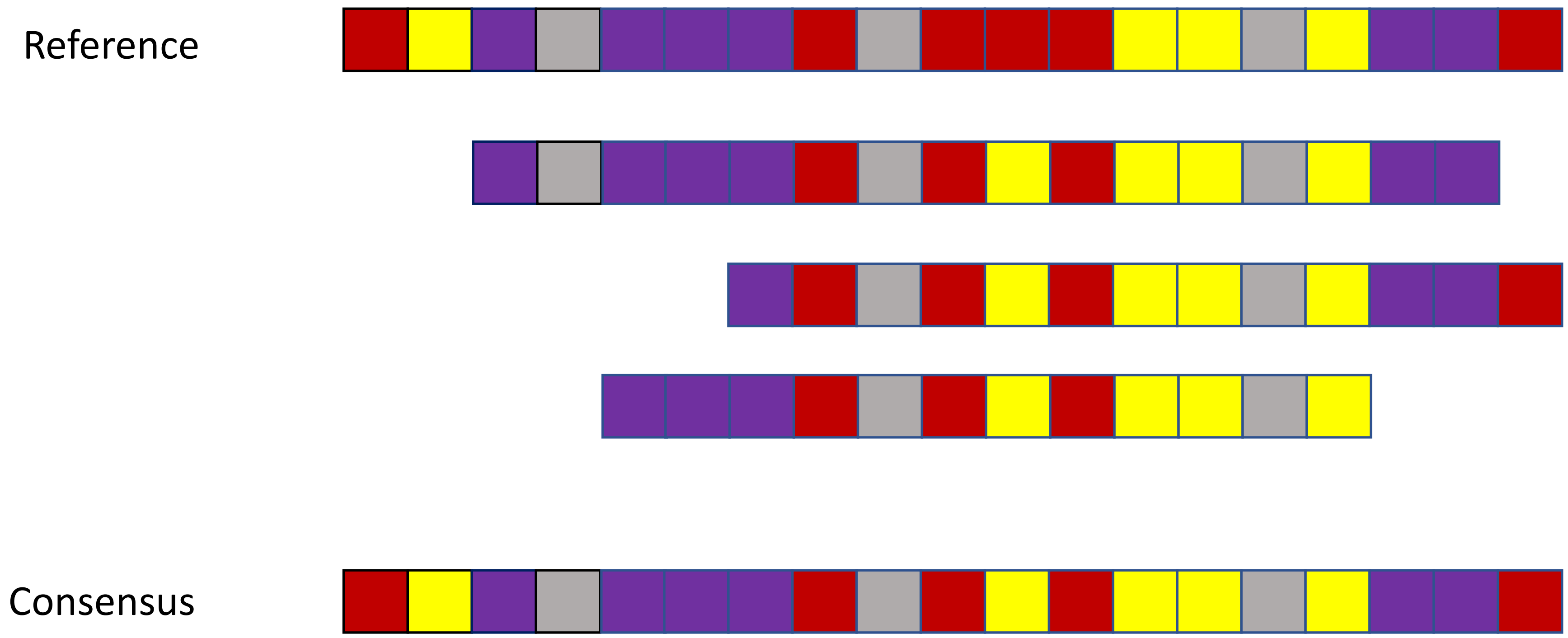
Accessing the accuracy of the genomes



Reference

Consensus

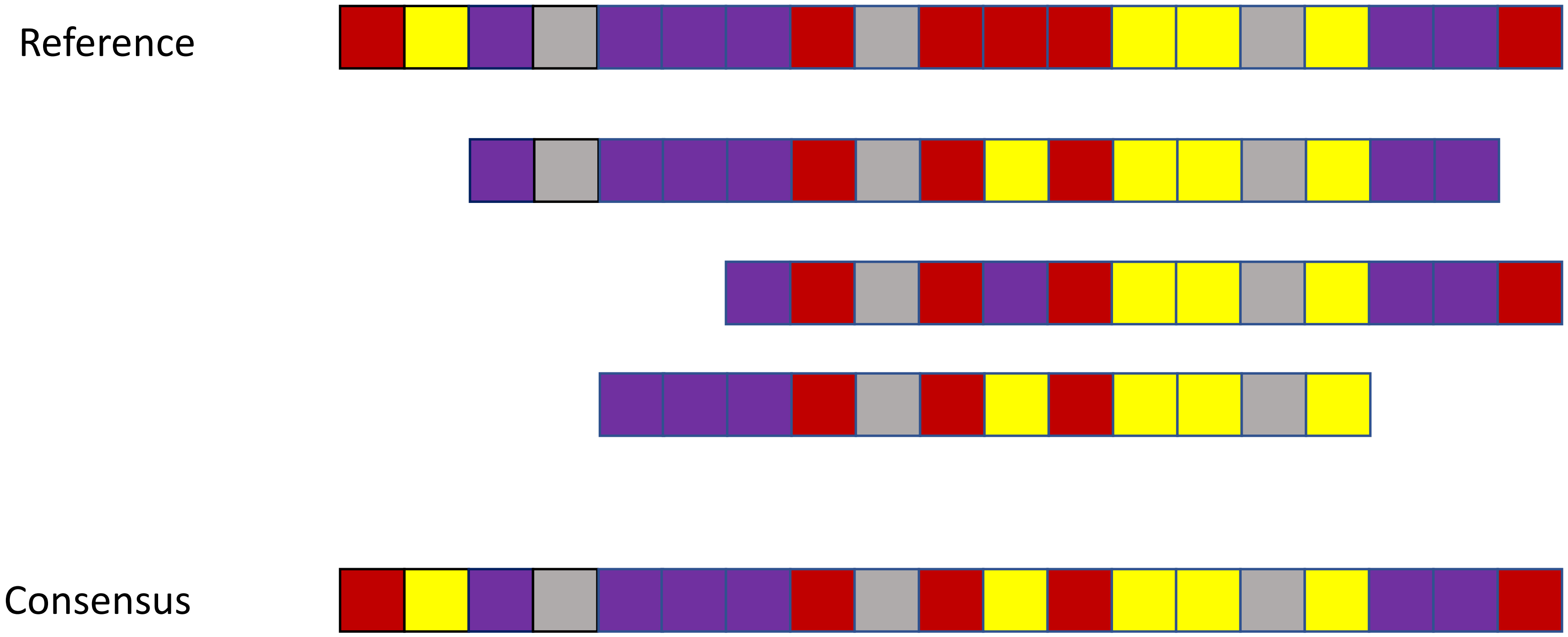
Reference Mismatch support



Reference

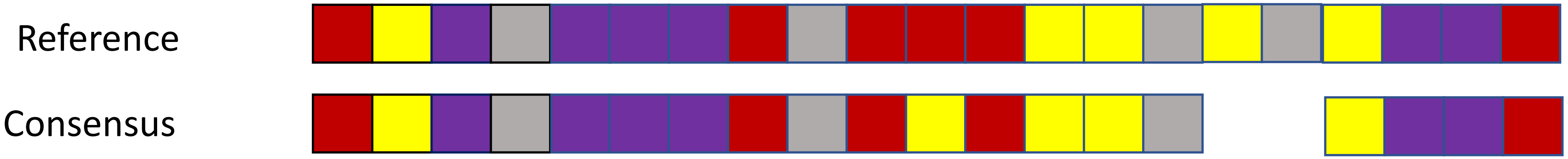
Consensus

Mixed positions



- Contamination?
- High Ct samples?
- Within host variation?

Frame-shifts



Reference

Consensus

Frame shift insertion

- Contamination?
- High Ct samples?
- Within host variation?

Sample contamination

Always include controls in your sequencing run

Negative control

- You don’t expect to see or assemble a genome from negative control

Positive control

- Will assist to troubleshoot in case of suspected contamination

Questions

Thank you