bioinformatics.ca

Canadian Bioinformatics Workshops

www.bioinformatics.ca

bioinformaticsdotca.github.io


Creative Commons
This page is available in the following languages:
Afrikaans български Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto
Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE)
Euskara Suomeksi français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu
Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik српски srpski (latinica) Sotho svenska
中文 華語 (台灣) isiZulu
creative commons
Attribution-Share Alike 2.5 Canada
You are free:
to Share — to copy, distribute and transmit the work
to Remix — to adapt the work
Free Cultural
APPROVED FOR
Works
Under the following conditions:
Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.
For any reuse or distribution, you must make clear to others the licence terms of this work.
Any of the above conditions can be waived if you get permission from the copyright holder.
The author's moral rights are retained in this licence.
Disclaimer
Your fair dealing and other rights are in no way affected by the above.
This is a human-readable summary of the Legal Code (the full licence) available in the following languages:
English French
Learn how to distribute your work using this licence

Contributions

- This slide deck includes material from John Tyson (BCCDC)

Module 4: Viral Pathogen Genomic Analysis

Jared Simpson

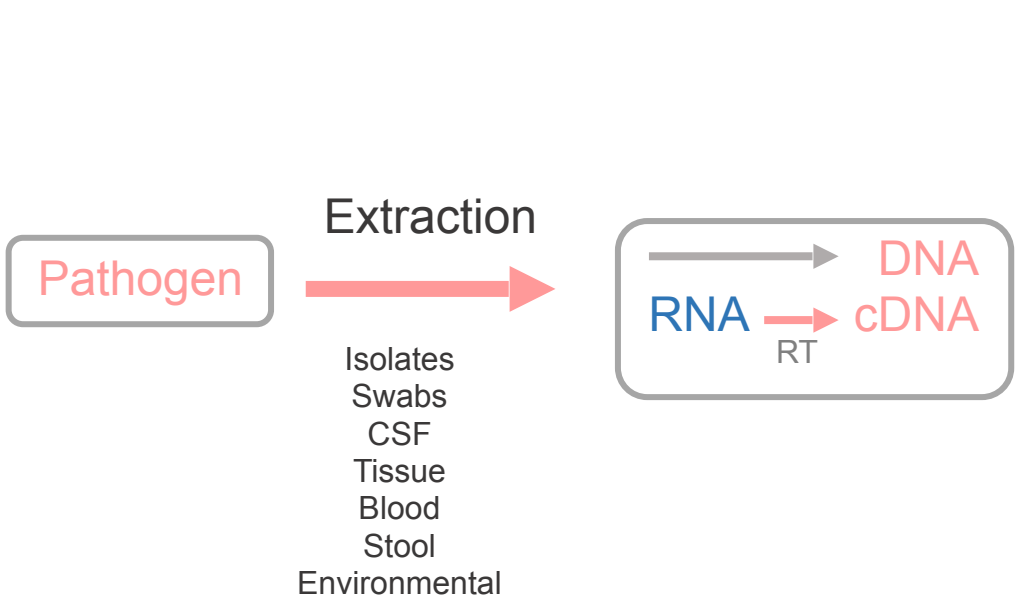
Infectious Disease Genomic Epidemiology

April 18-21, 2023

Learning Objectives

- By the end of this lecture, you will:
  - Understand the different approaches to sequence a viral genome
  - Know the key steps for analyzing amplicon data
  - Interpret the results of an amplicon-based analysis pipelines
  - Perform quality control on your data

Pathogen Genetic testing & typing



qPCR Diagnostic

Pathogen ID
Pathogen subtyping
Environmental Surveillance

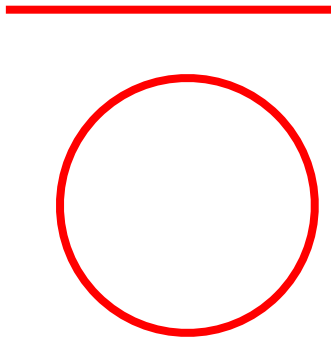
Sub-Genomic Target Sequencing

Agnostic pathogen ID e.g. 16S, ITS, COI, hsp70
Diagnostic confirmation
Variant Surveillance
Targeted AMR typing

WGS

Shotgun sequencing for outbreak investigation e.g. wgMLST, mobile elements
High-sensitivity target amplicon schemes e.g. *SARS-CoV-2*, FluA, Mpox

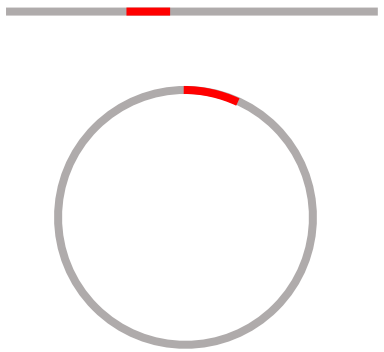
Isolate Sequencing



Sequencing

Identify Pathogen
(Whole genome or fragments)

Targeted Sequencing



PCR

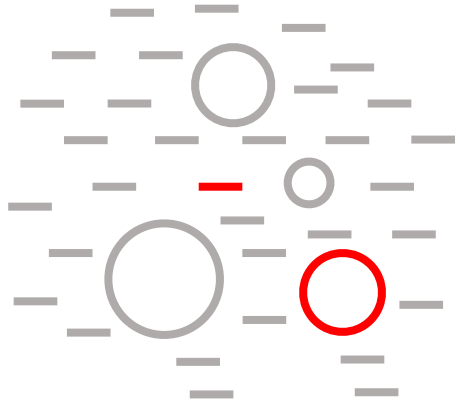
10^{5+} copies

Sequencing

Identify Pathogen
(e.g. 16S, ITS, COI or other genome fragment)

(Requires prior knowledge of Target)

Complex Mixture Sequencing

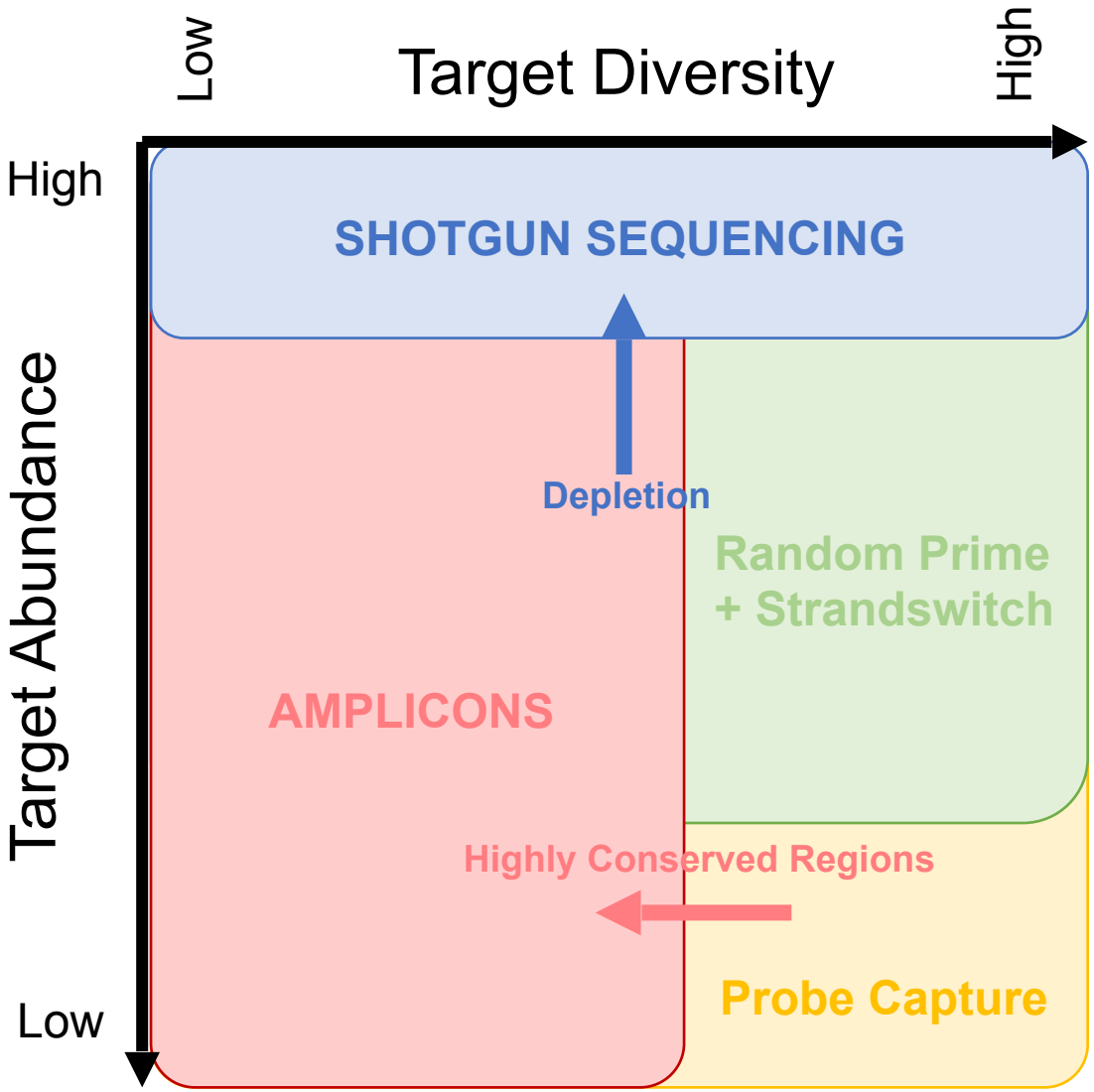


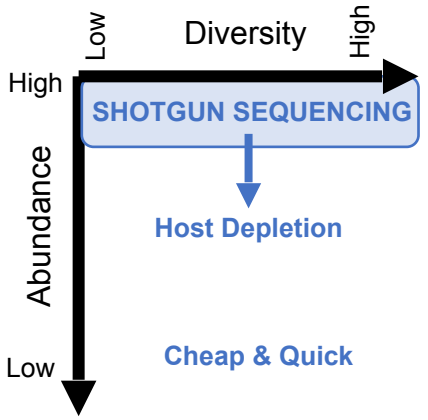
Capture or Depletion

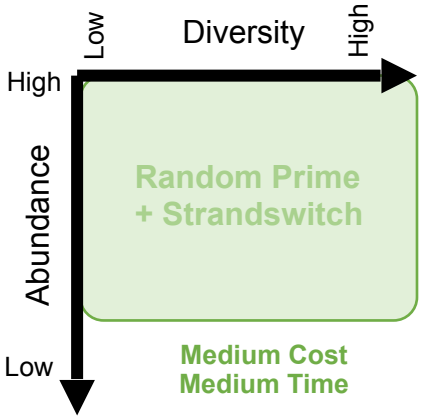
Shotgun Sequencing

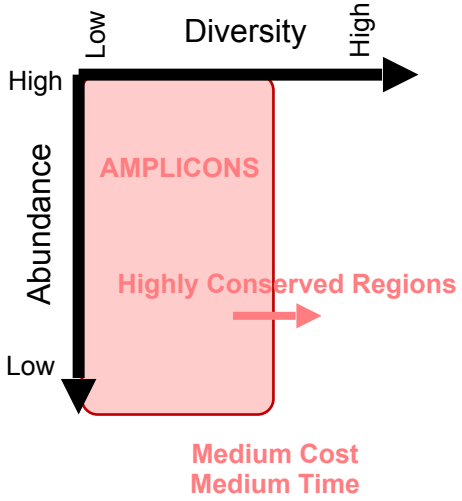
Sequencing

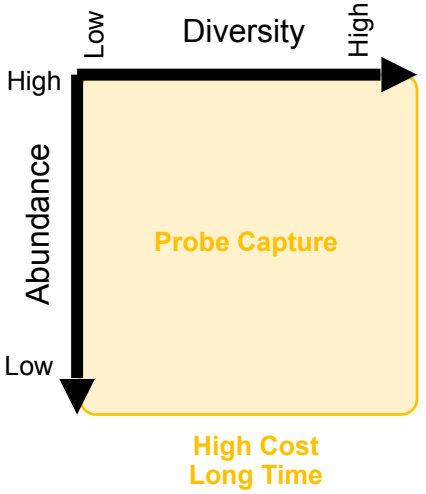
Identify Pathogen
(Whole genome or fragments)

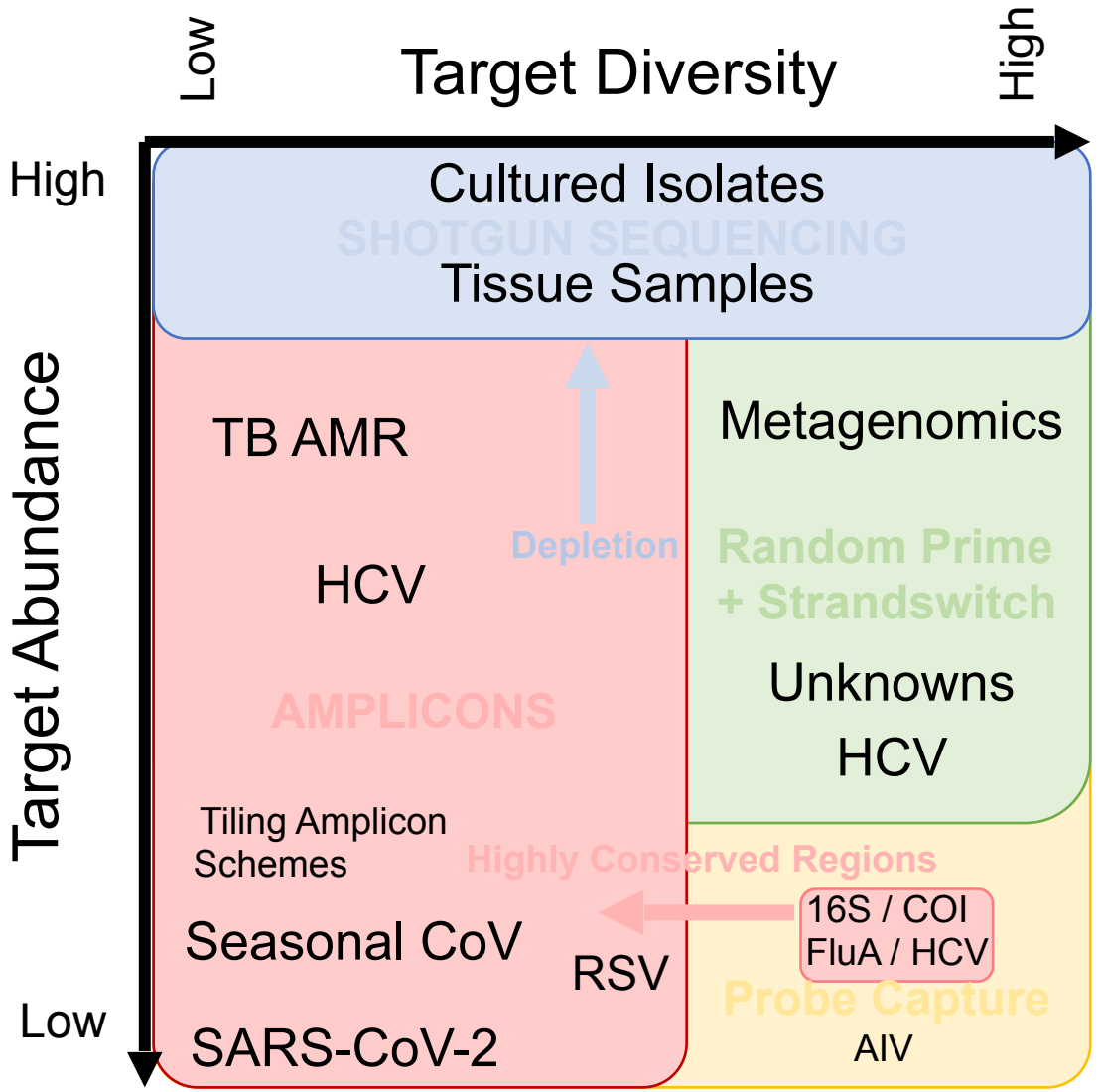

Low
Target Diversity
High
High
Target Abundance
Low
SHOTGUN SEQUENCING
Depletion
Random Prime
+ Strandswitch
AMPLICONS
Highly Conserved Regions
Probe Capture


Low
Diversity
High
High
Abundance
Low
SHOTGUN SEQUENCING
Host Depletion
Cheap & Quick


Low
Diversity
High
High
Abundance
Low
Random Prime
+ Strandswitch
Medium Cost
Medium Time


Low
Diversity
High
High
Abundance
Low
AMPLICONS
Highly Conserved Regions
Medium Cost
Medium Time

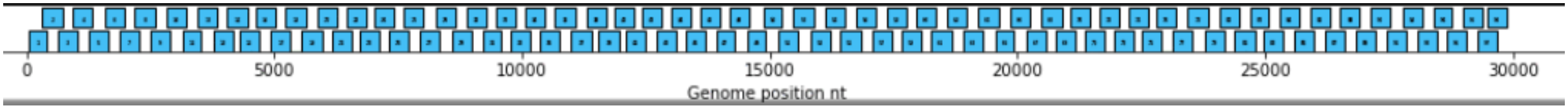

Low
Diversity
High
High
Abundance
Low
Probe Capture
High Cost
Long Time


Low
Target Diversity
High
High
Target Abundance
Low
Cultured Isolates
SHOTGUN SEQUENCING
Tissue Samples
TB AMR
HCV
AMPLICONS
Tiling Amplicon Schemes
Seasonal CoV
SARS-CoV-2
RSV
Depletion
Metagenomics
Random Prime + Strandswitch
Unknowns
HCV
Highly Conserved Regions
16S / COI
FluA / HCV
Probe Capture
AIV

The Different Sequencing Technologies

- There are a number of “Next Generation” sequencing technologies available:
  - Illumina short read length sequencers :- small to high capacity instruments, short read lengths (<300b)
  - PacBio long read length sequencers :- Medium capacity instruments, long read lengths (<50Kb CCS, ~20Kb HiSeq consensus)
  - Oxford Nanopore long read length sequencers :- Small to large capacity, Ultra long read lengths (250 to 2Mb+)

Tiled Amplicon Sequencing



Artic Protocol (V3)

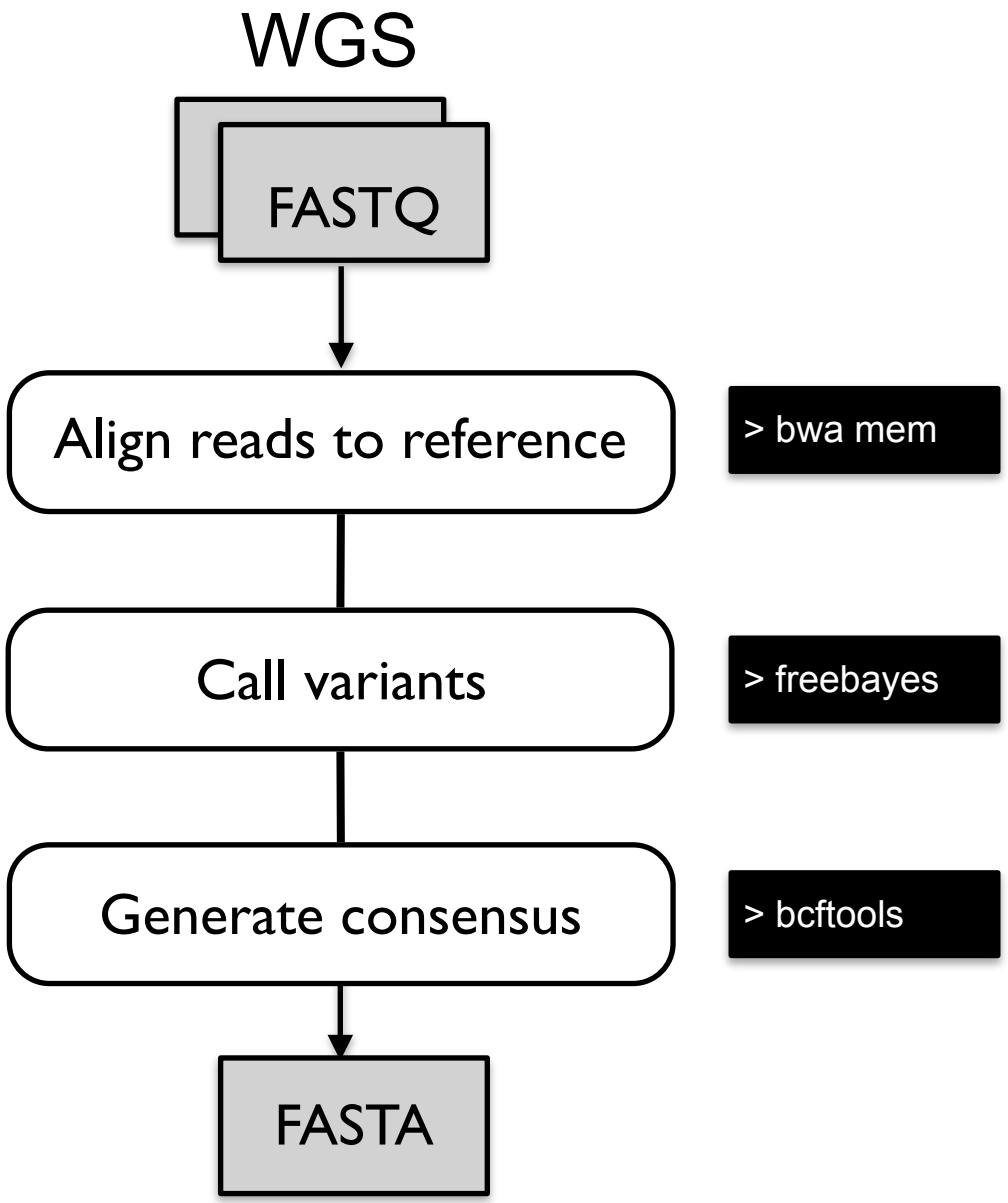
- Highly multiplexed PCR reaction
- 98 primer pairs over two primer pools
- 400 bp amplicons
- Balancing amplicons → sequencing extra deep to accommodate

Preprint 2020

Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore

John R Tyson,[1] Phillip James,[2] David Stoddart,[2] Natalie Sparks,[3] Arthur Wickenhagen,[4] Grant Hall,[5] Ji Hyun Choi,[6] Hope Lapointe,[6] Kimia Kamelian,[7] Andrew D Smith, Natalie Prystajecky,[7,8] Ian Goodfellow,[5] Sam J Wilson,[4] Richard Harrigan,[6] Terrance P Snutch,[1] Nicholas J Loman,[3] and Joshua Quick[3]

Analysis Pipelines



(This pipeline is intentionally simplified)

Read Mapping and Alignment

Typically reads are much shorter (100-200bp) than the reference genome (SARS-CoV-2: ~30,000bp)

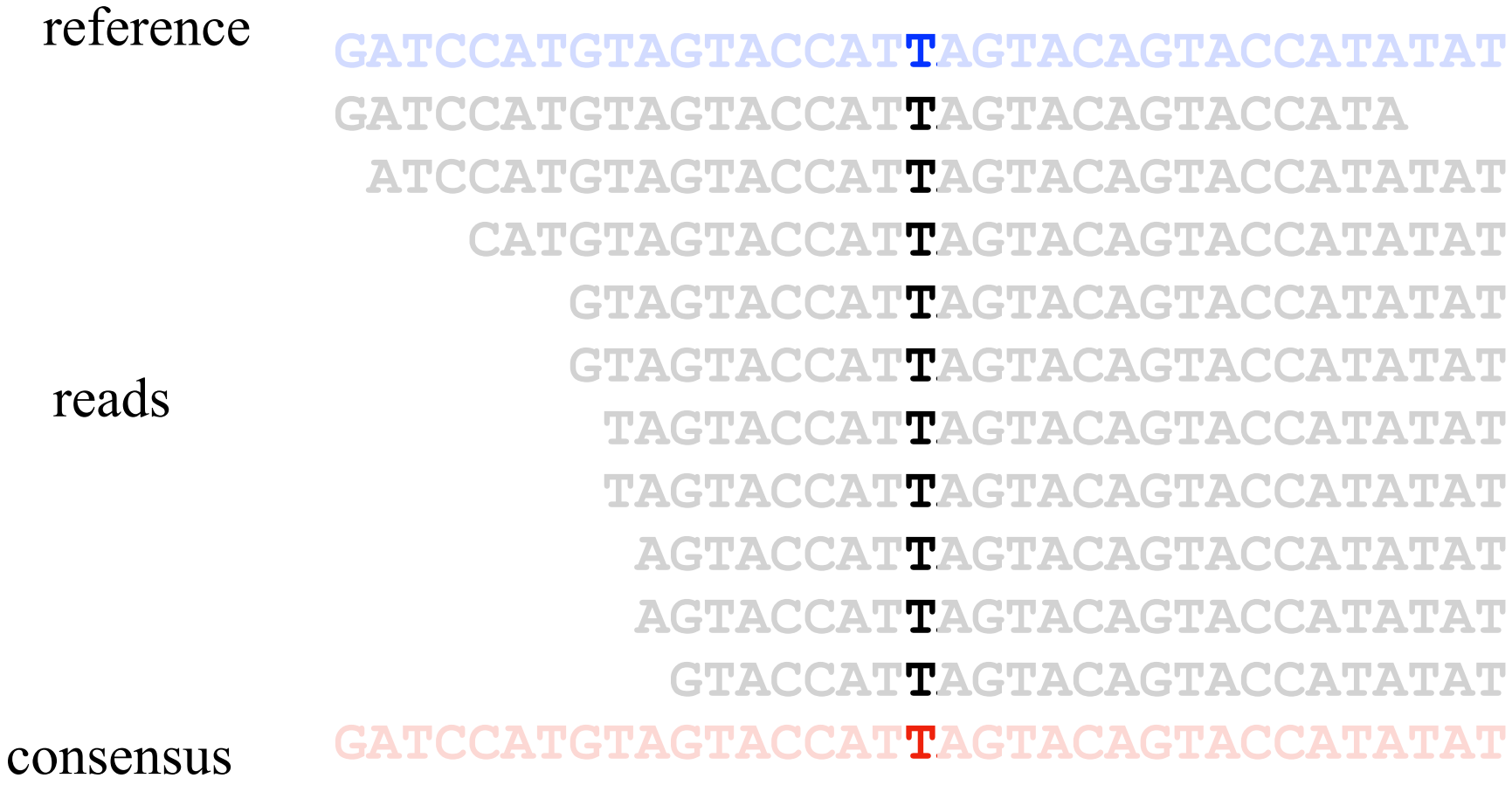
We need to find where the read *maps* to in the reference genome and how the bases in the read *align* to the reference

```
reference    GAGATACATGAGAGAGTATCTCGACTCTAGGCCGATACCATTGTA
                        ||||||| |||||||
read                    AGTATCTTGACTCTA
```

Key definitions:
mapping: the region in the reference that is most similar to the read
alignment: how the read lines up to the reference base-by-base

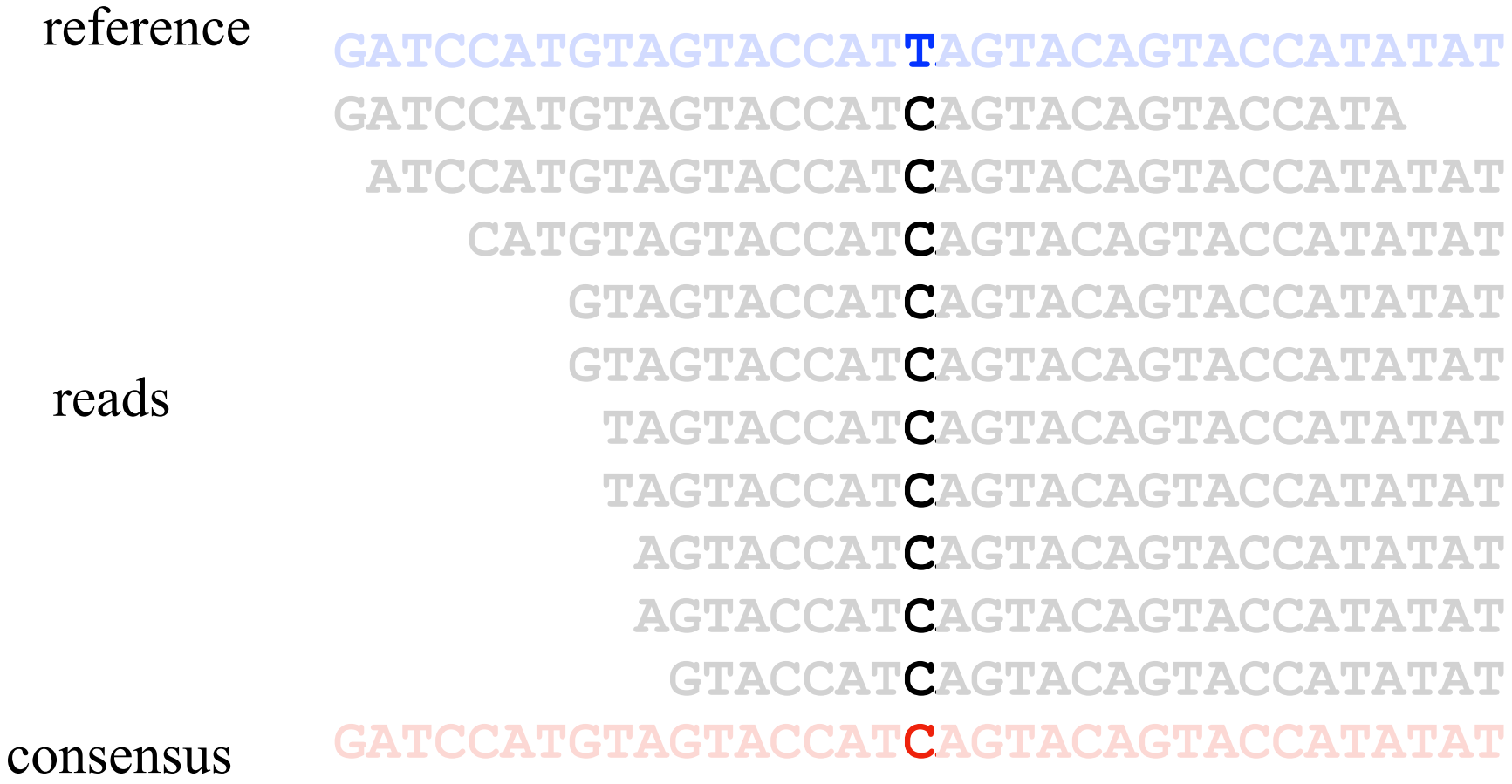
Consensus sequences

- Compare reads to the SARS-CoV-2 reference genome



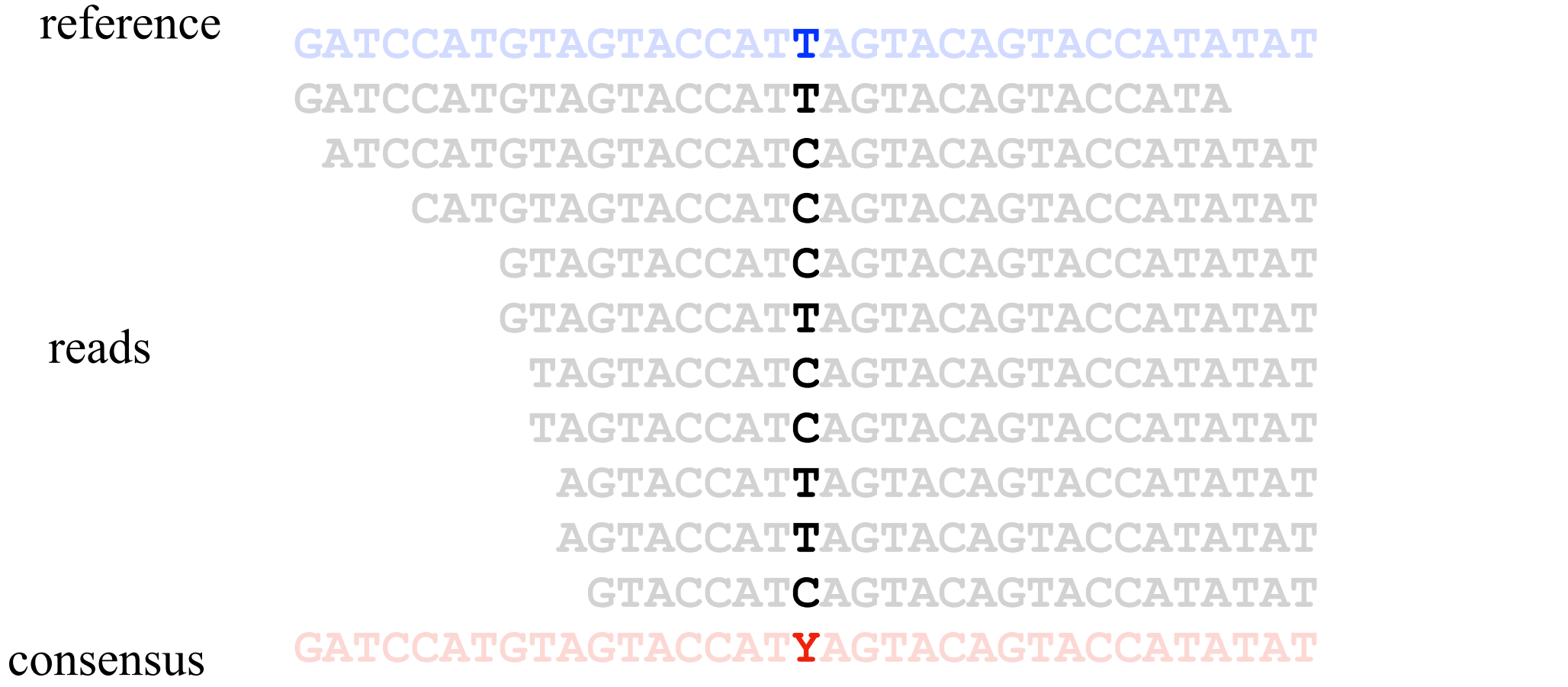
Consensus sequences

- *variants* or *mutations* are detected when the aligned reads differ from the reference



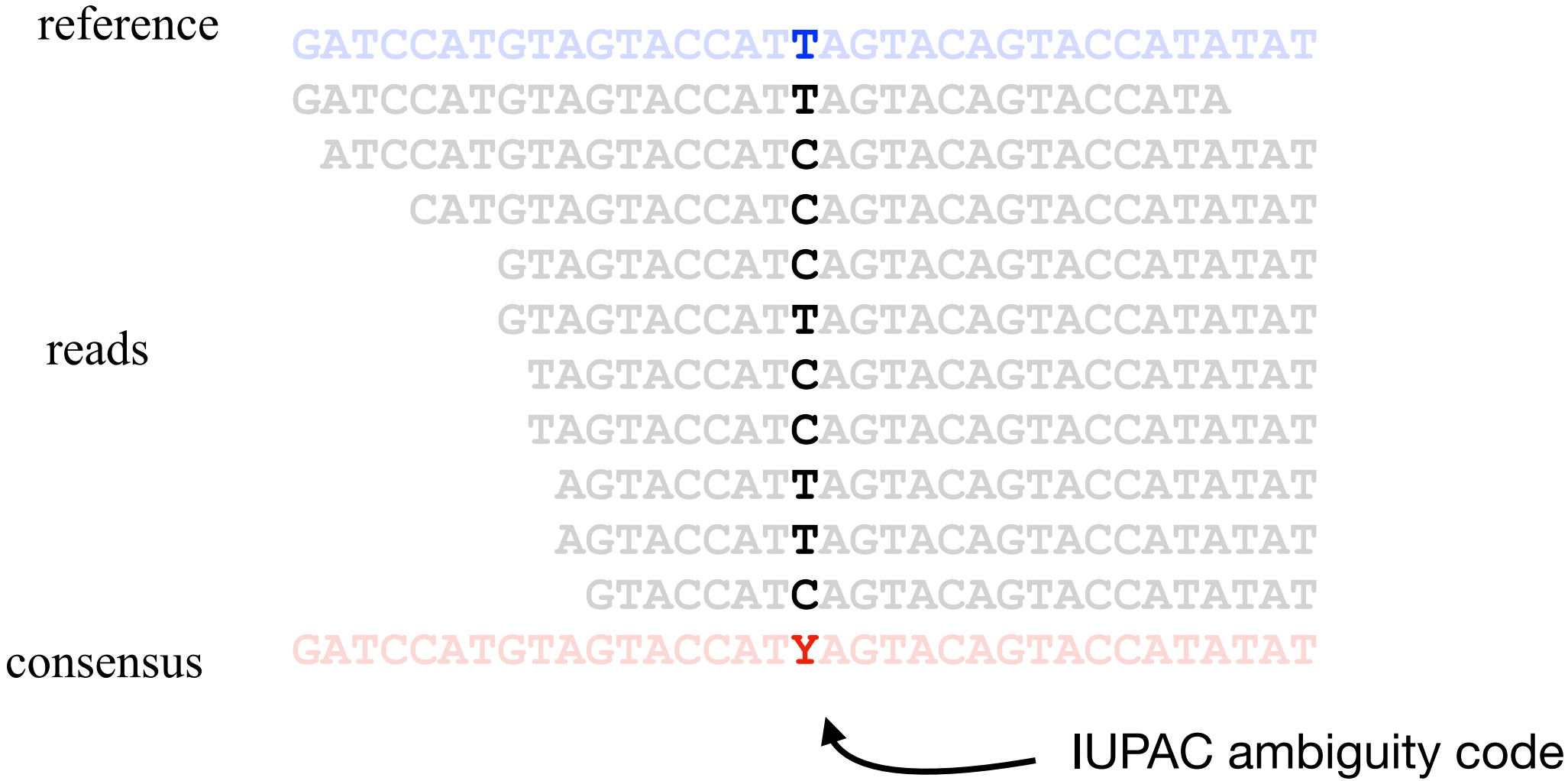
Consensus sequences

- *ambiguous* or *mixed* bases are detected when the aligned reads have evidence for more than one type of base

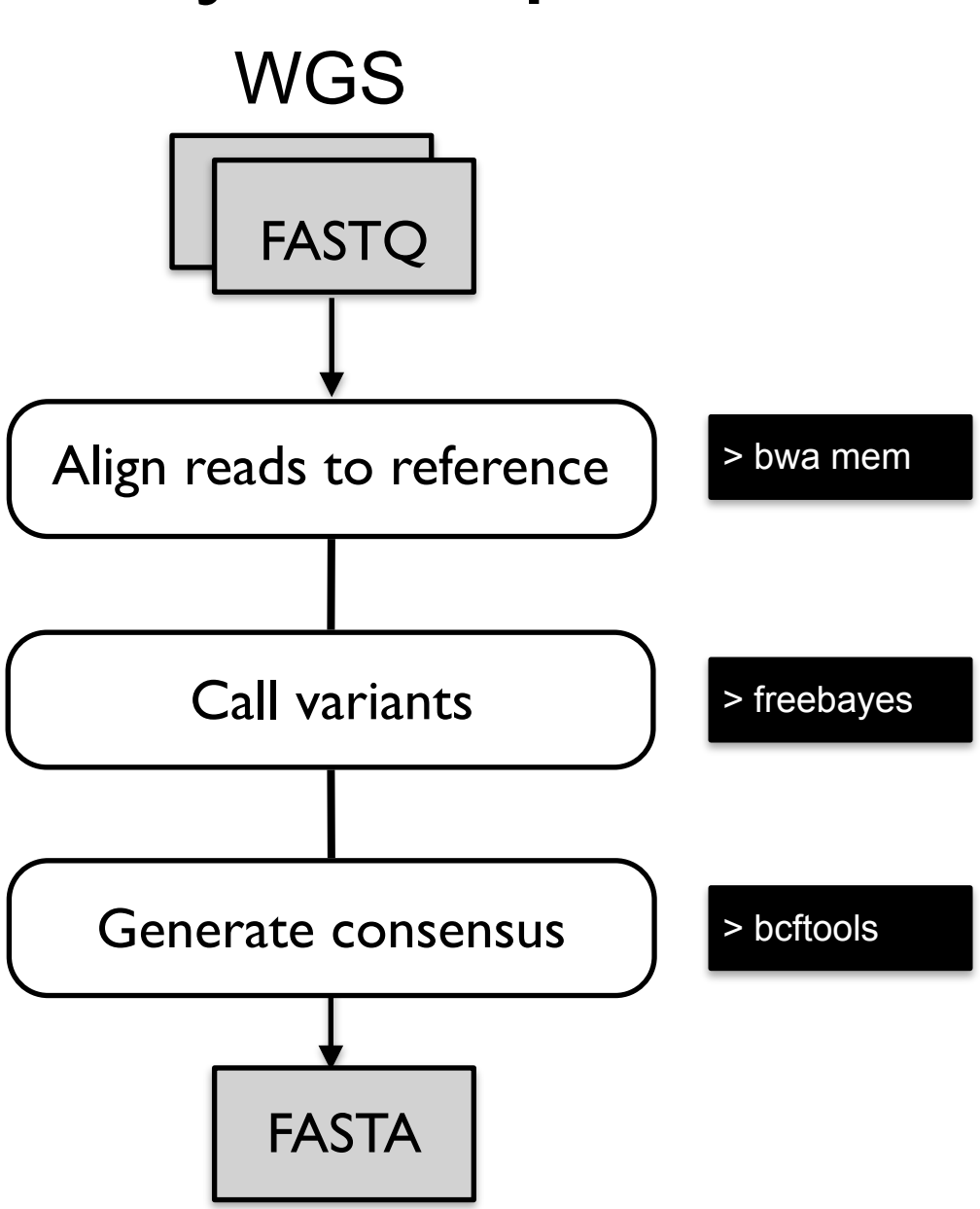


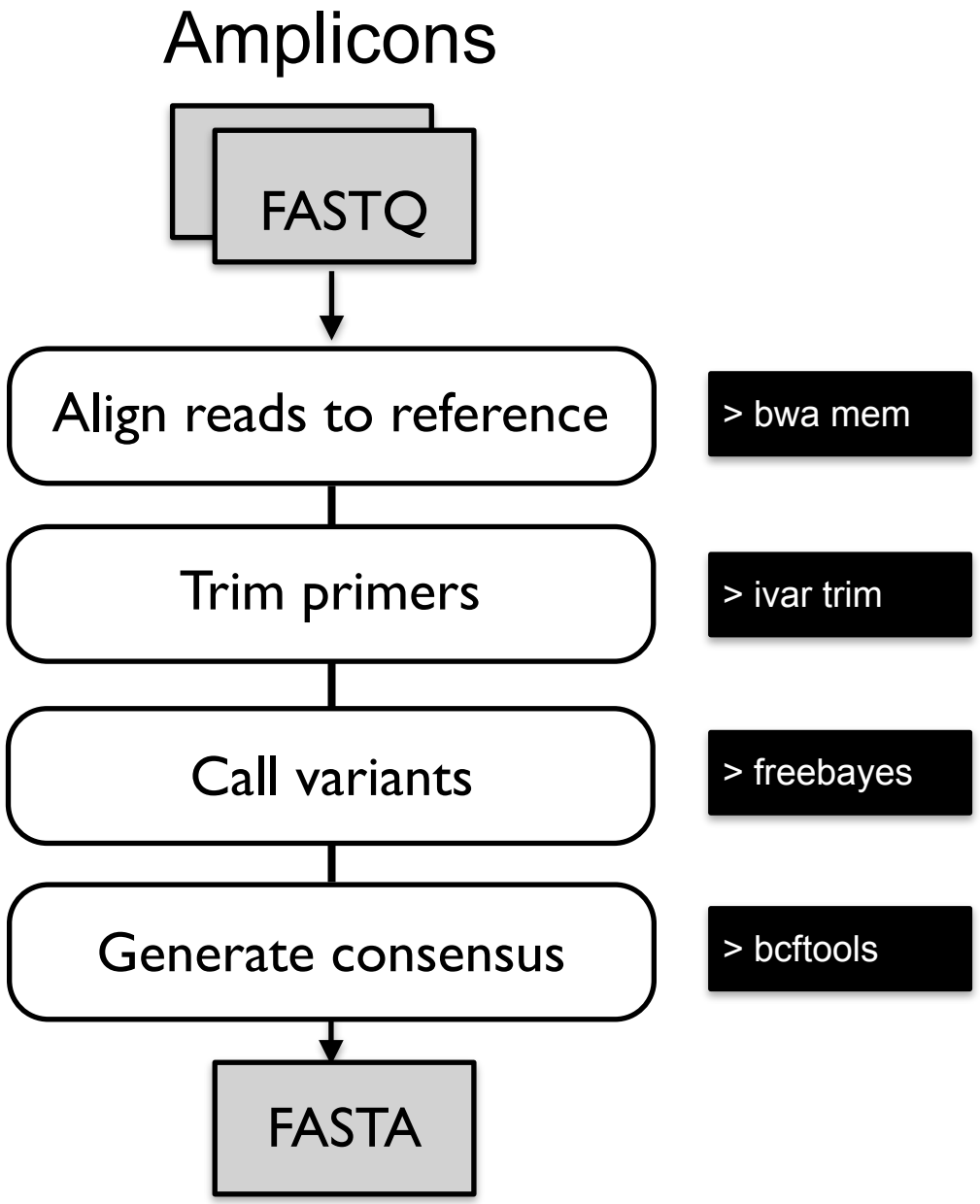
Consensus sequences

- Can you think of explanations for *why* this can happen?



Analysis Pipelines





Why do we need to trim primers?

Example Illumina read:

```
ERR5338522.5
ACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAACTA
+
FFFFF:FFF:FFFFFF:FFFFFFFFFFFFFFFFFFFFFFF,FFF,,FFFFFFFFFFFFFF:FFFFFFFFF:FFFFFFFFFFFFFFFFFFFF:FFFFFFFFF,FFFF:FFFF
```

Why do we need to trim primers?

Example Illumina read:

```
ERR5338522.5
ACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAACTA
+
FFFFF:FFF:FFFFFF:FFFFFFFFFFFFFFFFFFFFFFF,FFF,,FFFFFFFFFFFFFF:FFFFFFFFF:FFFFFFFFFFFFFFFFFFFF:FFFFFFFFF,FFFF:FFFF
```

ARTIC V3 primer highlighted in red

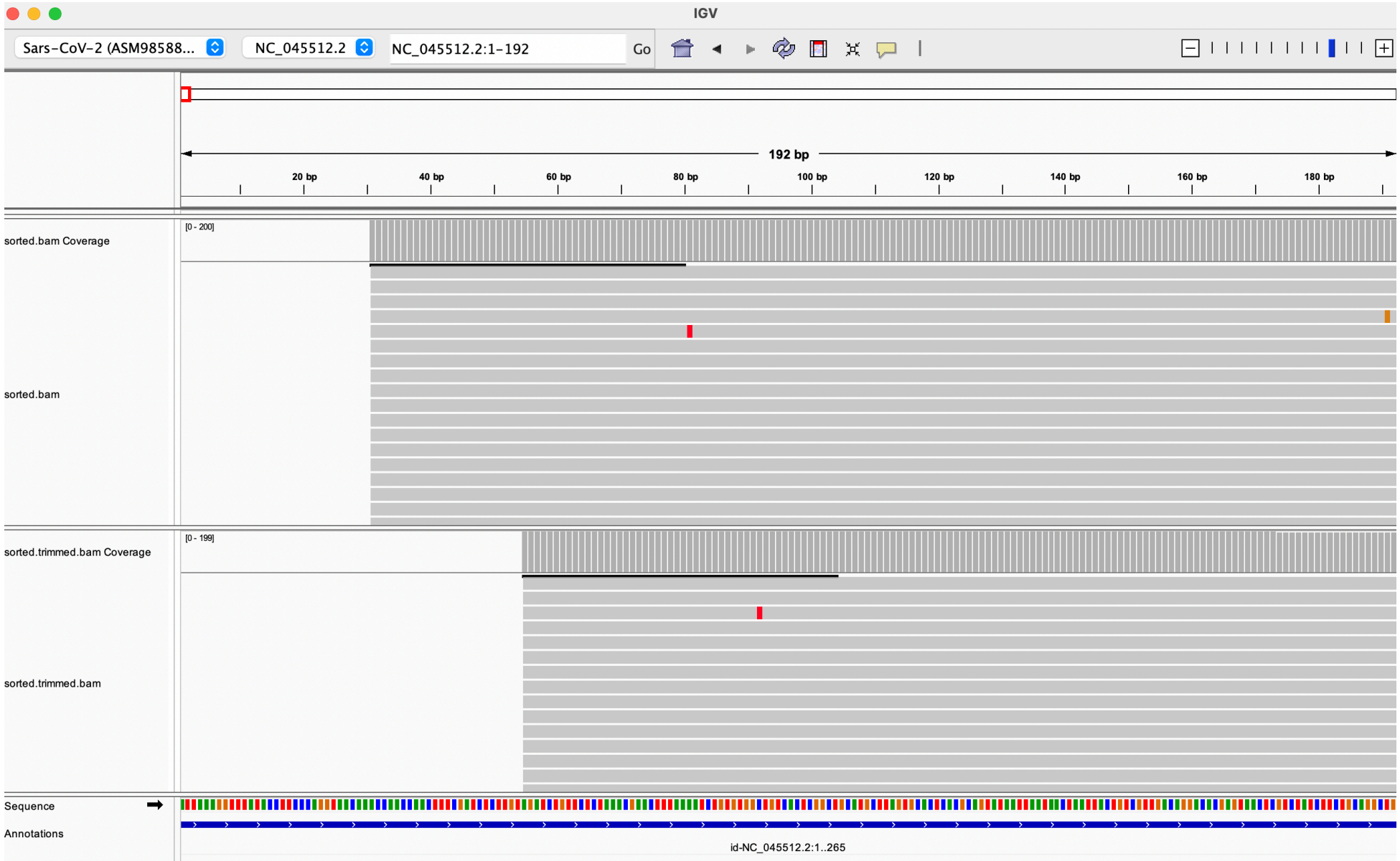
Primer sequence is always identical to reference and cannot be used to detect mutations

Unless removed from the reads we will over-call reference alleles

Why do we need to trim primers?

Before trimming:

After trimming:



How do we quality check our results?

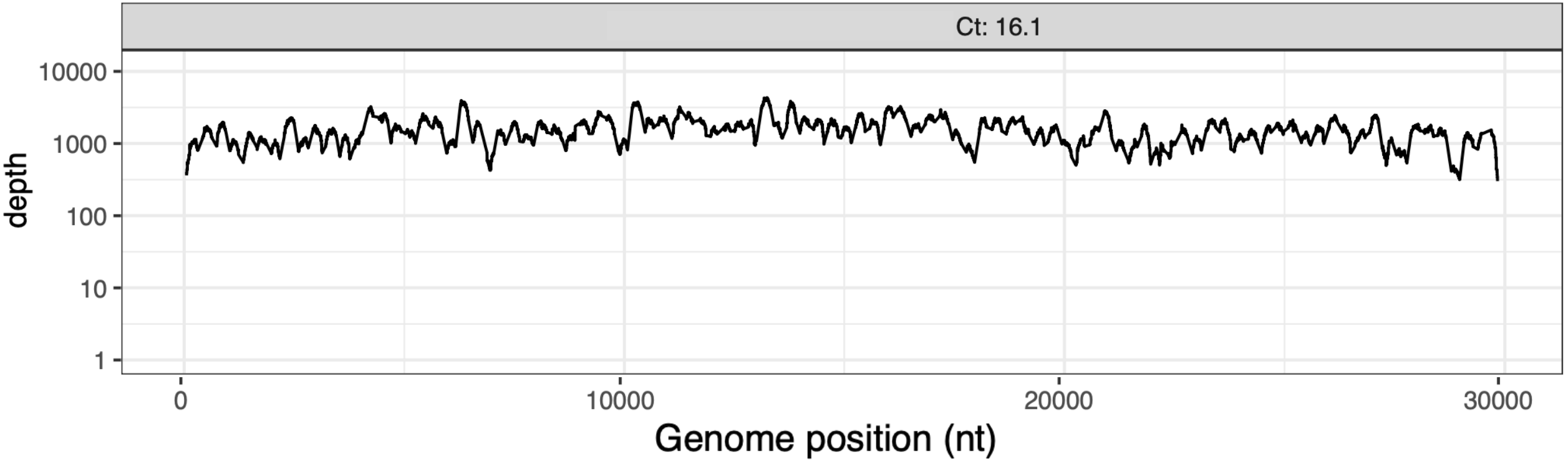
- Was the genome successfully sequenced?
  - Virus abundance varies sample-to-sample, storage conditions and extraction methods may differ in ability to recover complete genomes
- Is the genome sequence accurate?
  - Low abundance samples may have amplification artifacts, low coverage may lead to consensus errors
- Is the sequencing run contaminated?
  - Sequencing protocols are highly multiplexed, many amplification cycles increases risk of cross-contamination on the plate

What do we mean by quality?

- **Was the genome successfully sequenced?**
  - Virus abundance varies sample-to-sample, storage conditions and extraction methods may differ in ability to recover complete genomes
- Is the genome sequence accurate?
  - Low abundance samples may have amplification artifacts, low coverage may lead to consensus errors
- Is the sequencing run contaminated?
  - Sequencing protocols are highly multiplexed, many amplification cycles increases risk of cross-contamination on the plate

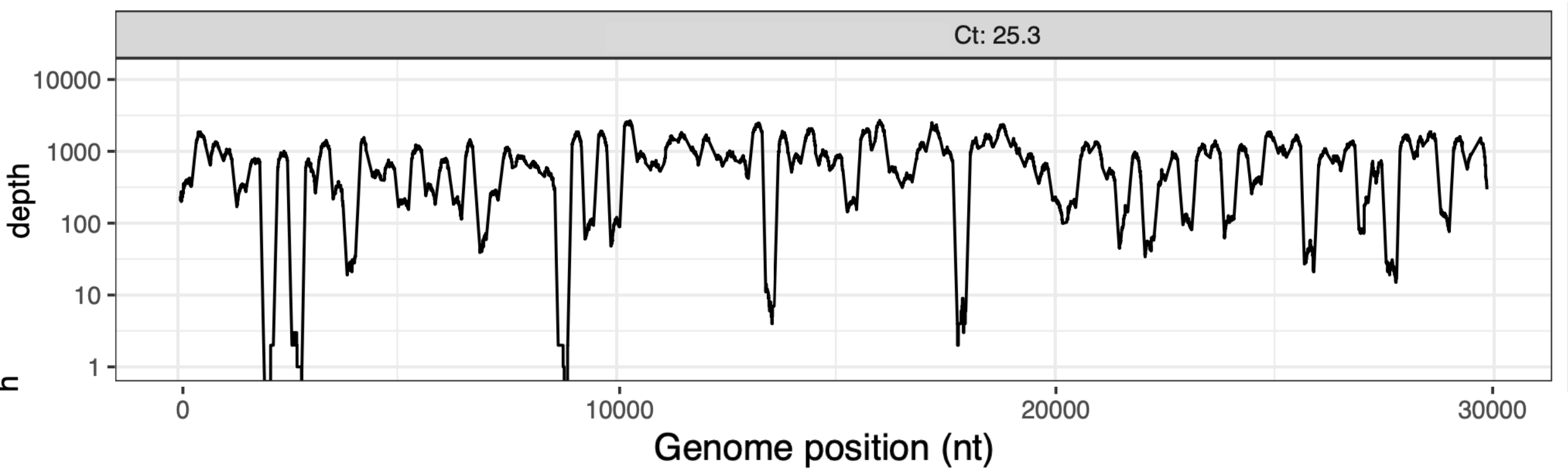
Genome Coverage

Low Ct sample, ideal coverage profile



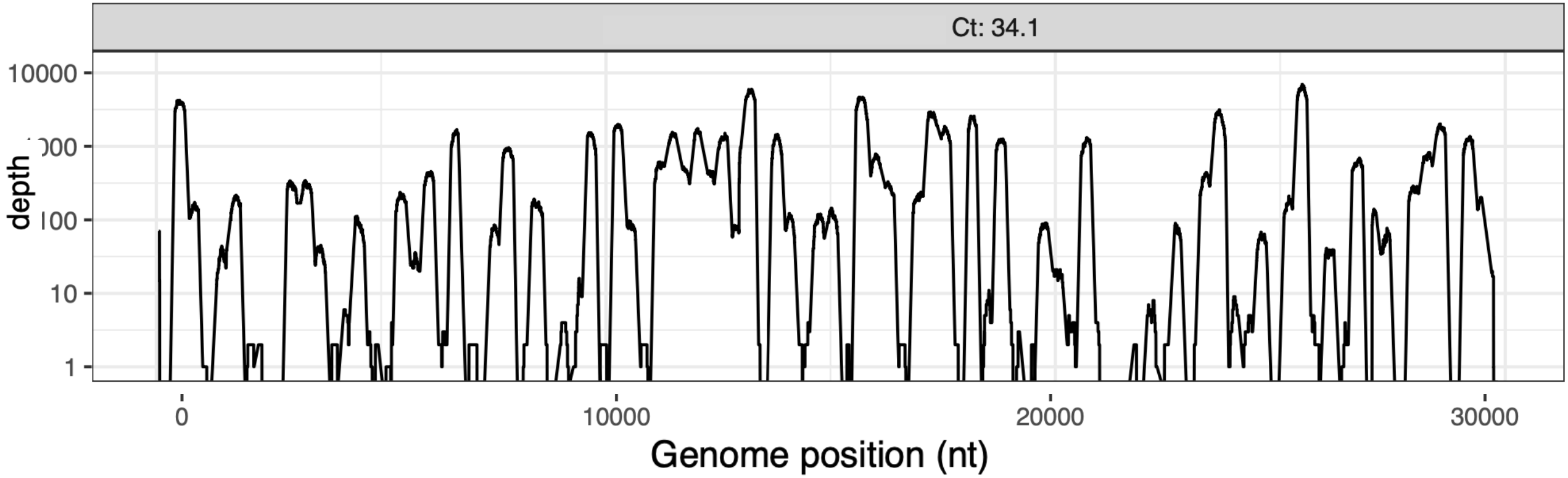
Genome Coverage

Moderate Ct sample, some dropouts



Genome Coverage

High Ct sample, many dropouts



Genome Completeness

- The bioinformatics pipelines take the mapped reads and output a consensus sequence
  - Positions with insufficient coverage are masked with “N” bases
- Genome completeness is defined as the proportion of non-N bases

```
>sampleA
ACGGNNNACA
```

Genome completeness = 7/10 = 0.7

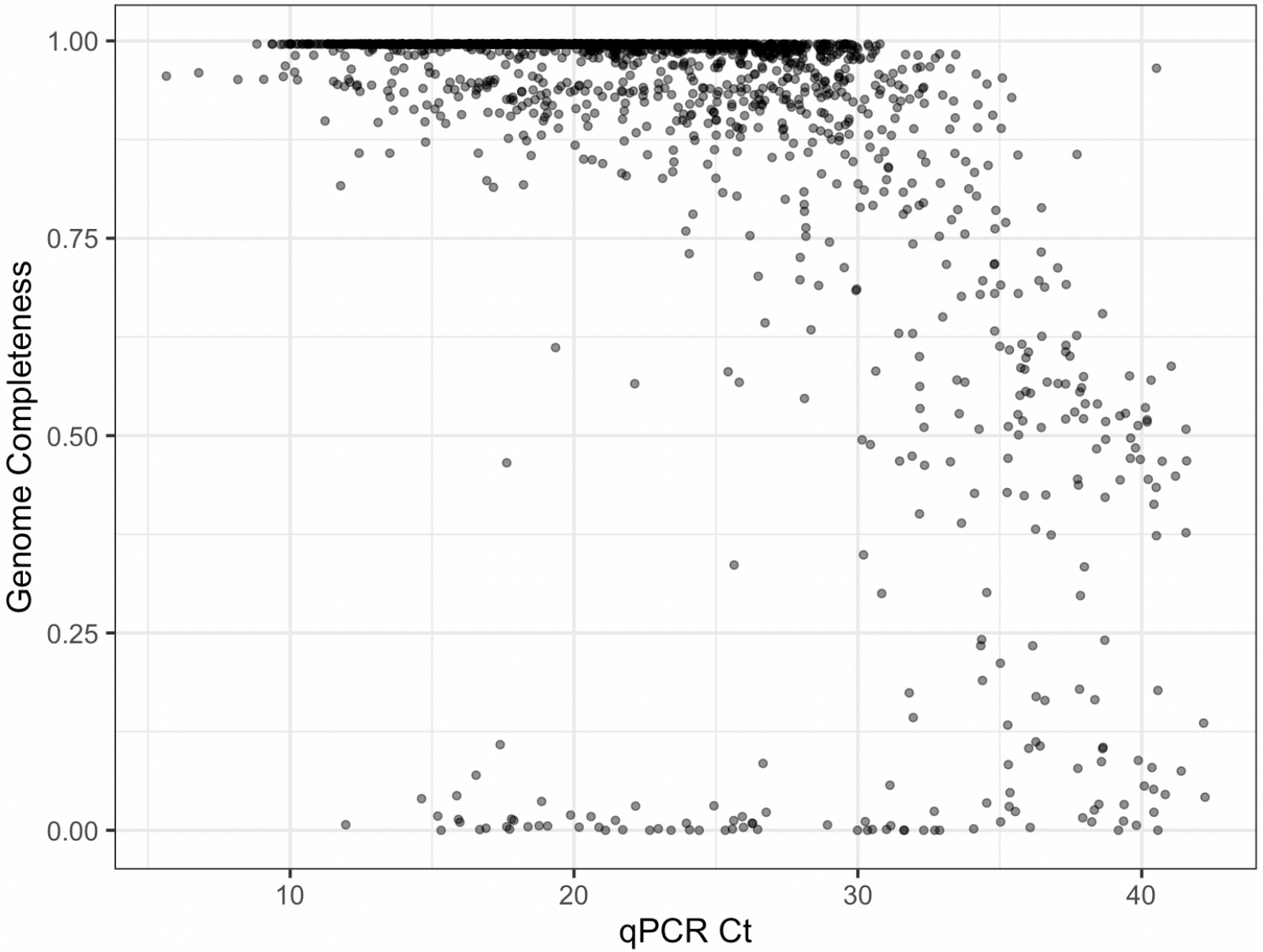
```
>sampleB
ACGGANTACA
```

Genome completeness = 9/10 = 0.9

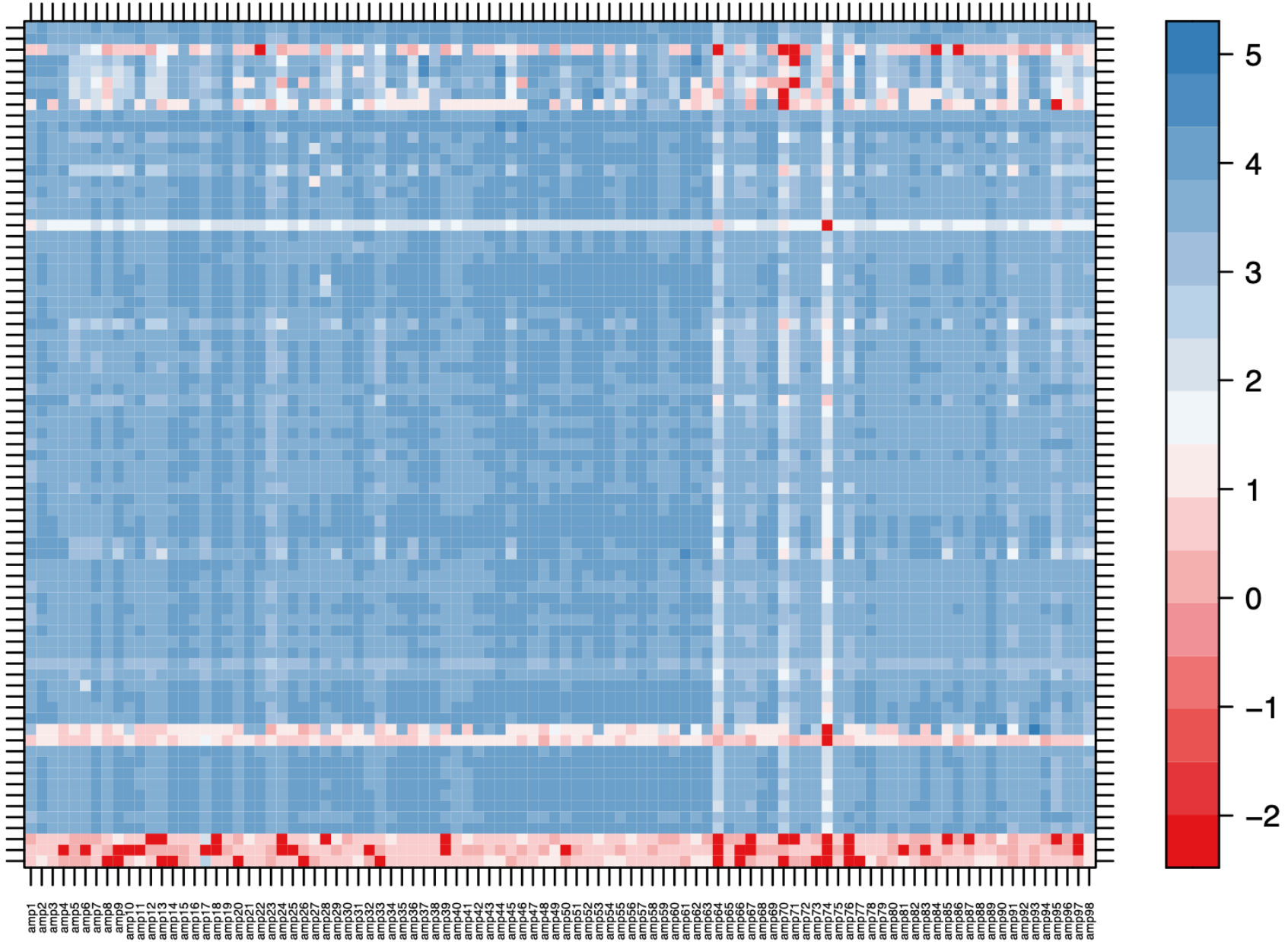
Genome Completeness

- Incomplete genomes are harder to analyze (e.g. may cause problems for multiple sequence alignment or building a tree)
- Many projects use genome completeness as a primary criteria for submission to public repositories
  - For CanCOGeN genomes with completeness >90% are submitted to GISAID

Relationship between Ct and completeness



Genome Coverage QC

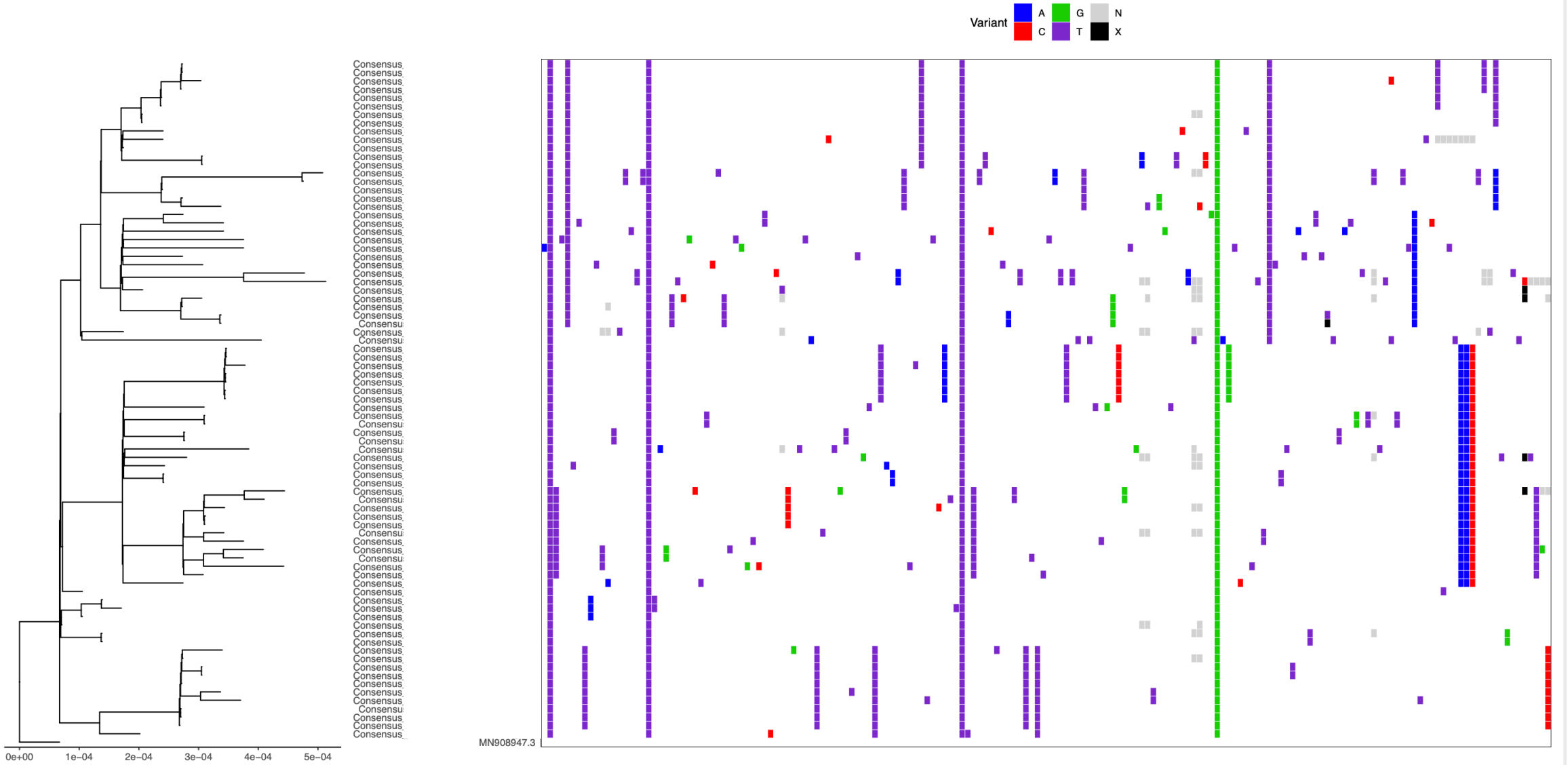


Richard de Borja

What do we mean by quality?

- Was the genome successfully sequenced?
  - Virus abundance varies sample-to-sample, storage conditions and extraction methods may differ in ability to recover complete genomes
- **Is the genome sequence accurate?**
  - Low abundance samples may have amplification artifacts, low coverage may lead to consensus errors
- Is the sequencing run contaminated?
  - Sequencing protocols are highly multiplexed, many amplification cycles increases risk of cross-contamination on the plate

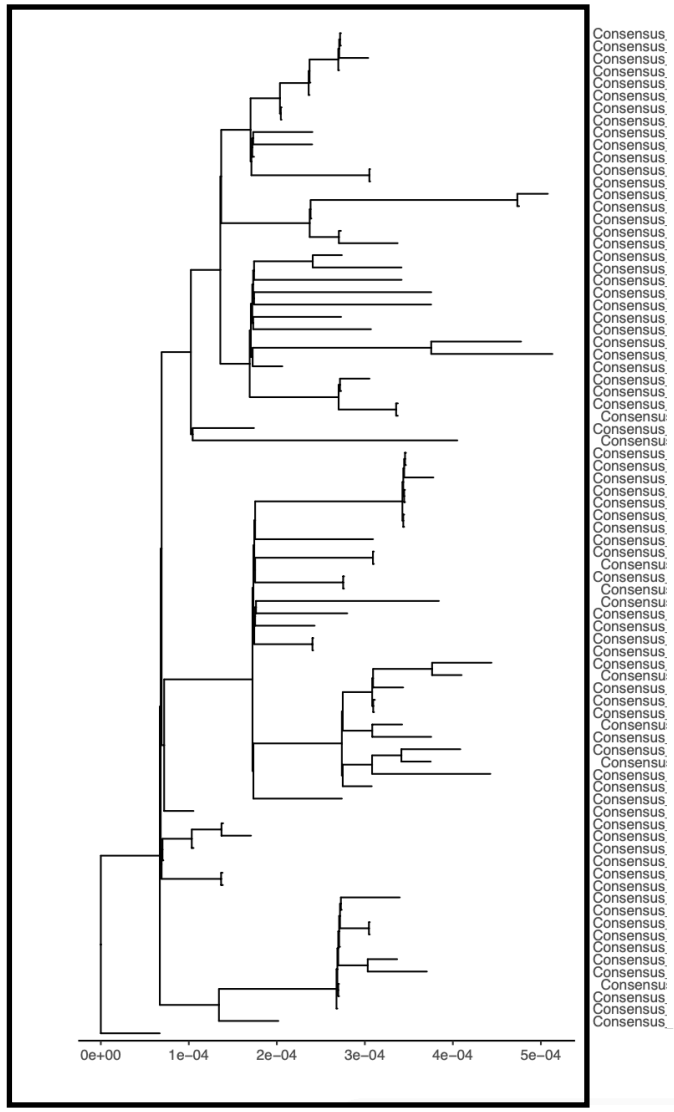
Assessing variants - ncov-tools tree/snps

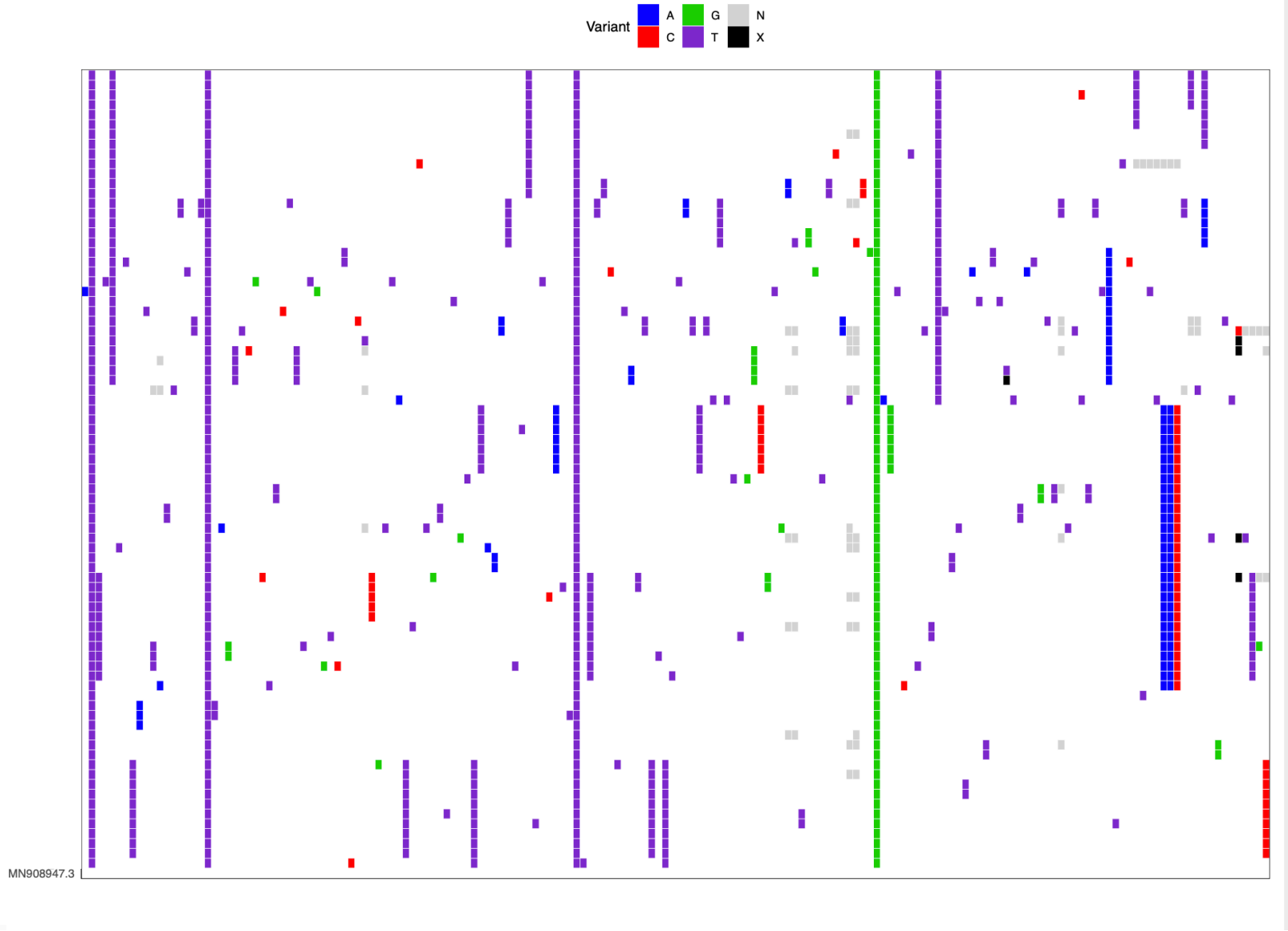


This plot was inspired by work from Mads Albertsen

Assessing variants - ncov-tools tree/snps

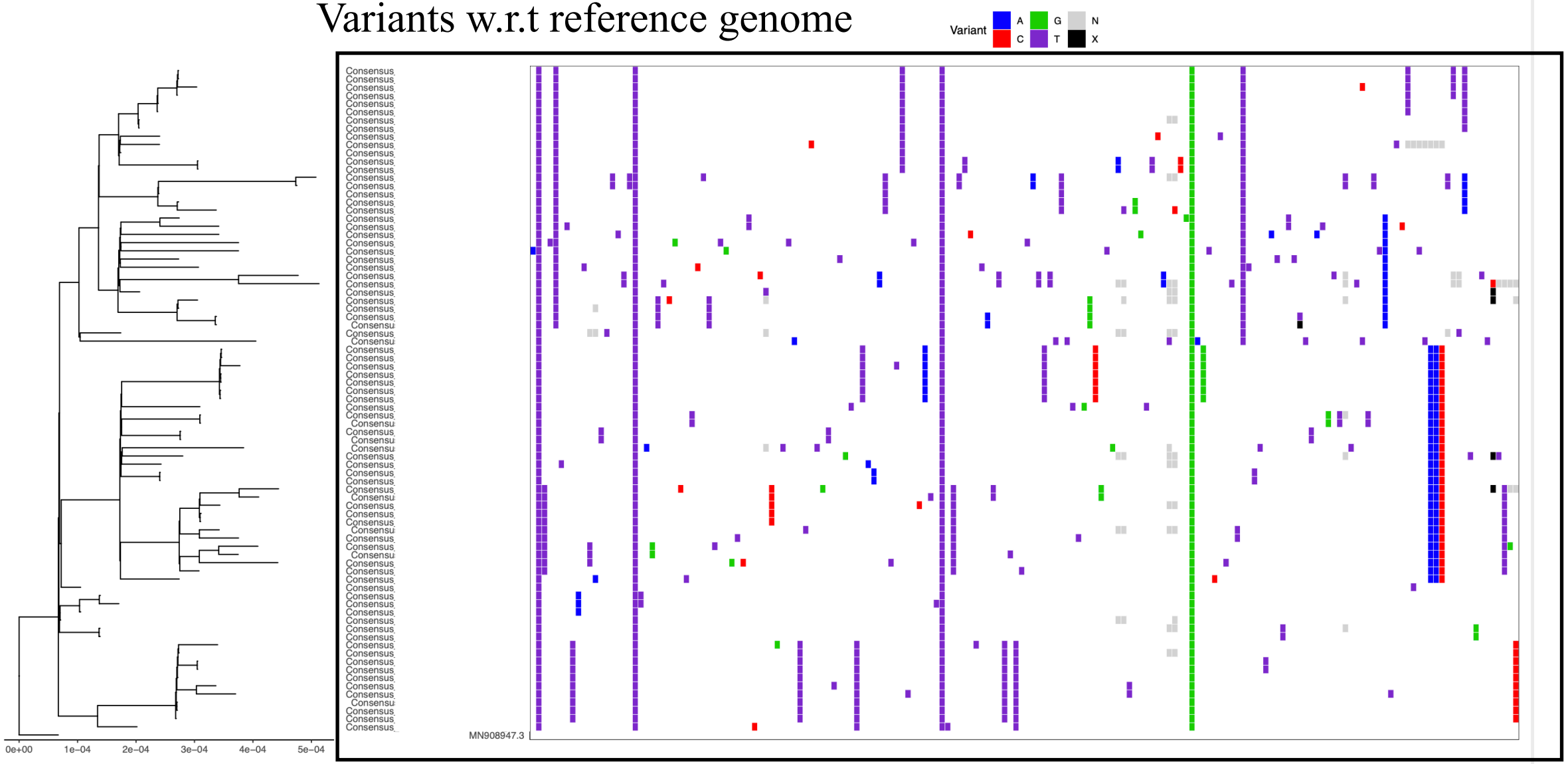
tree built by MAFFT+iqtree



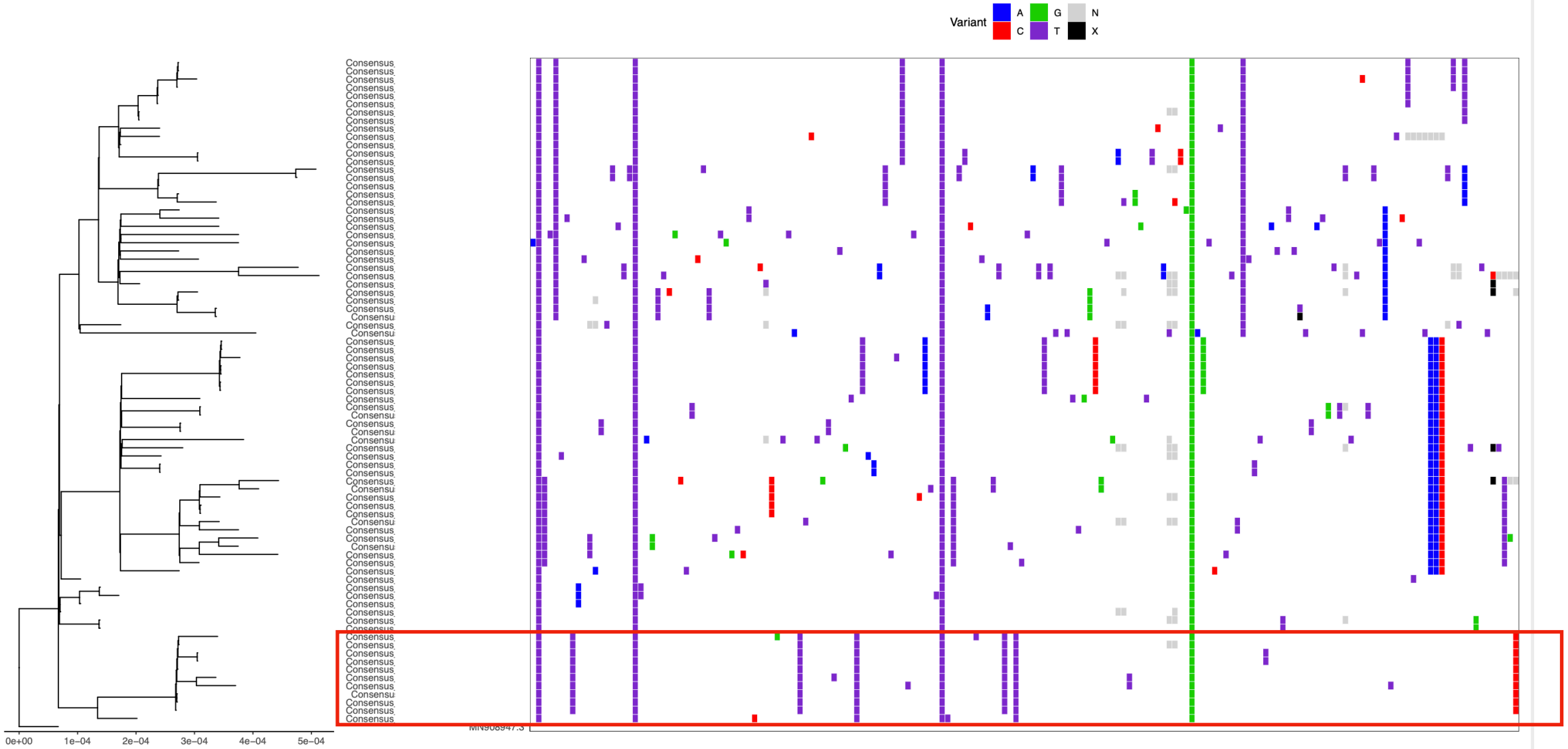


Assessing variants - ncov-tools tree/snps

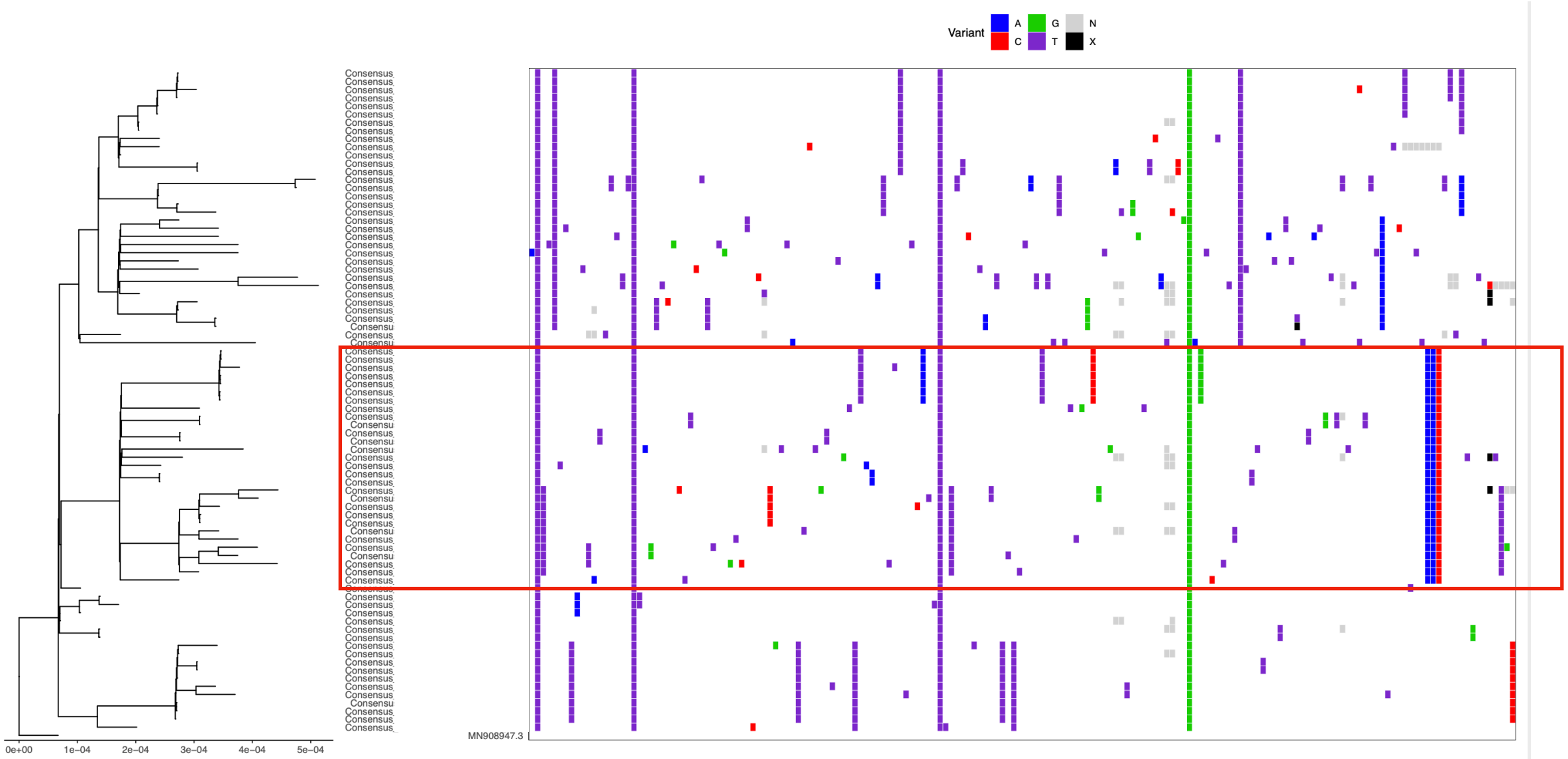
Variants w.r.t reference genome



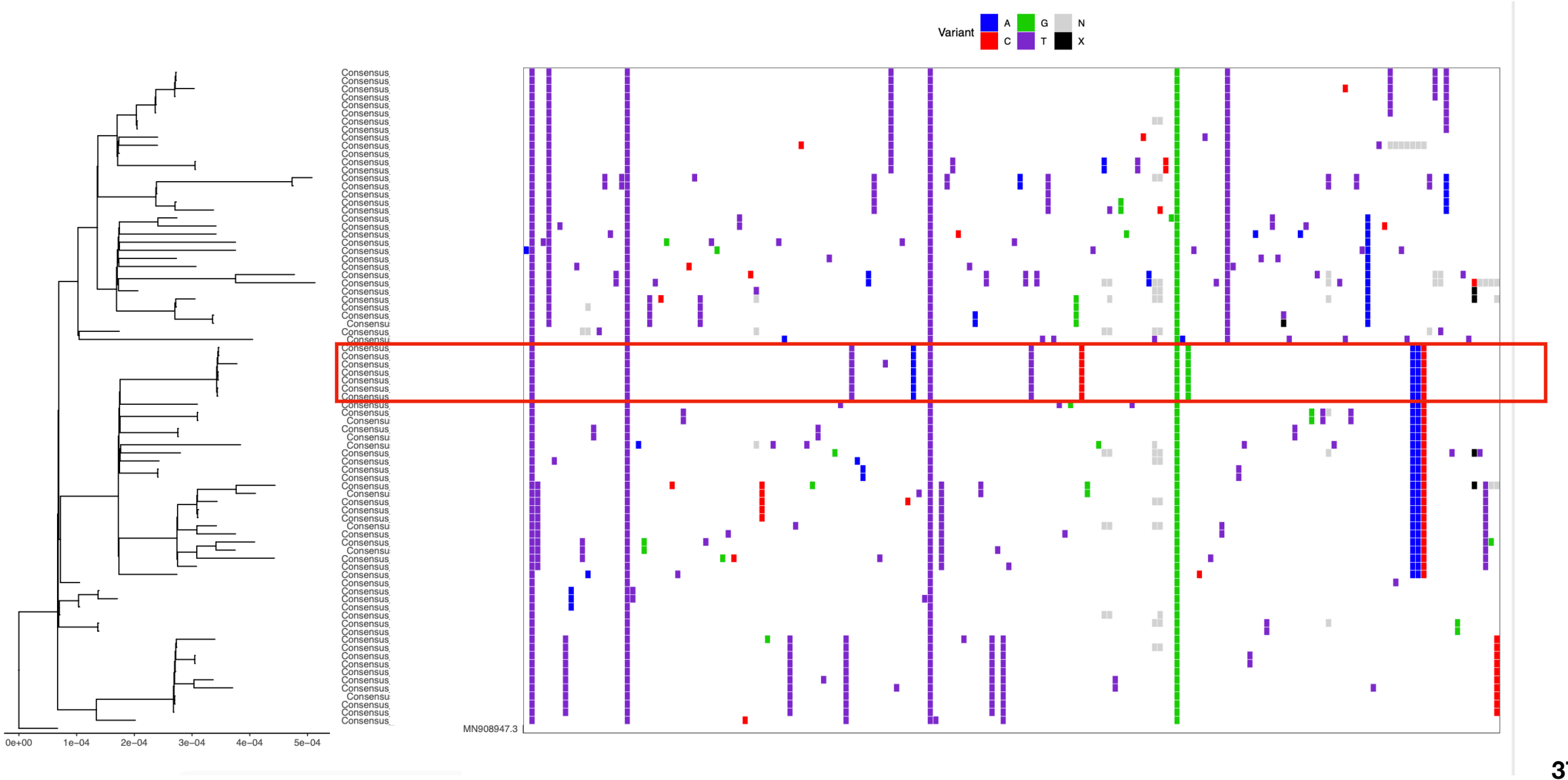
Assessing variants - ncov-tools tree/snps



Assessing variants - ncov-tools tree/snps



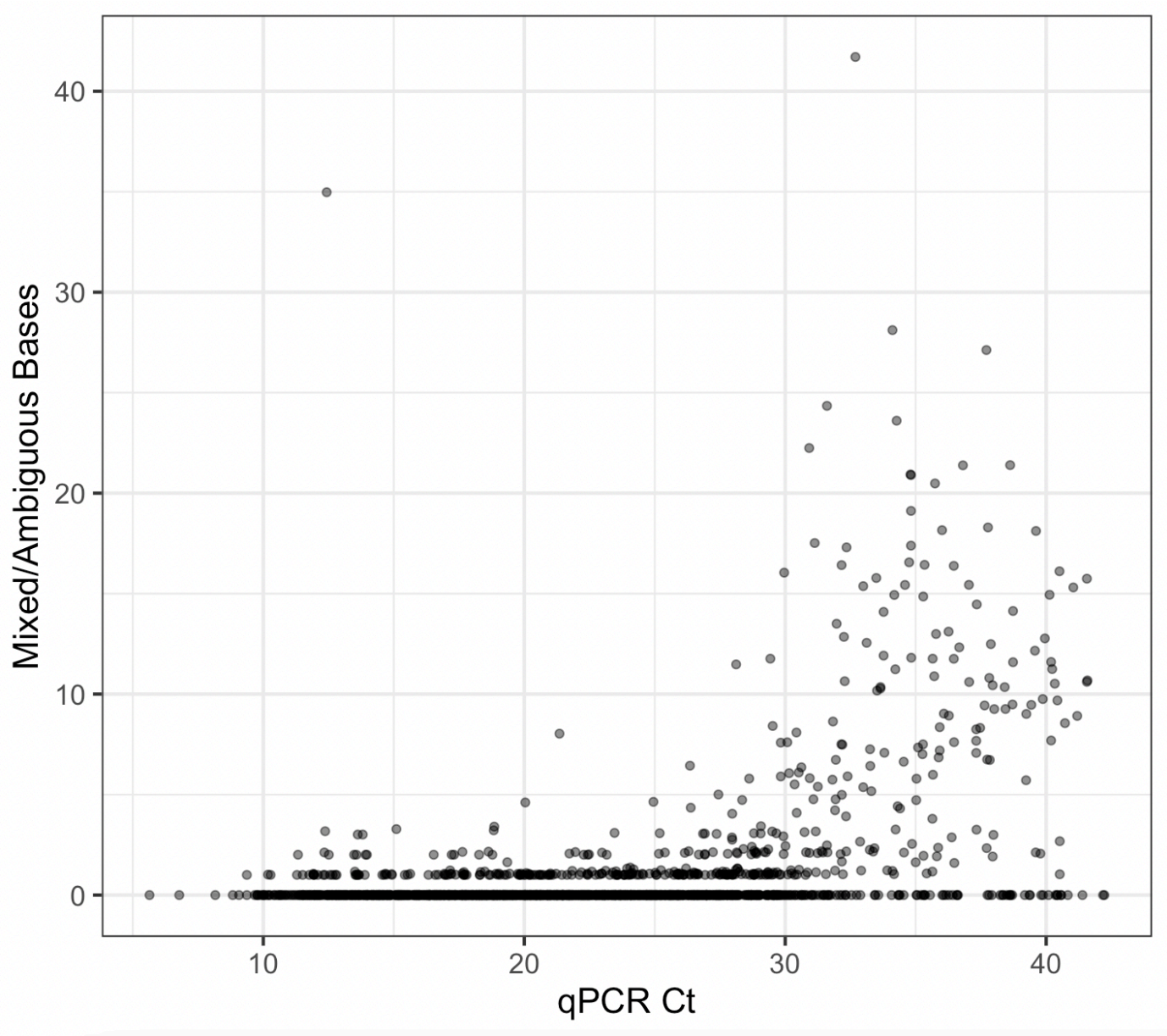
Assessing variants - ncov-tools tree/snps



Variant QC - Mixed/Ambiguous Bases

Number of observed ambiguous bases increases for high Ct samples

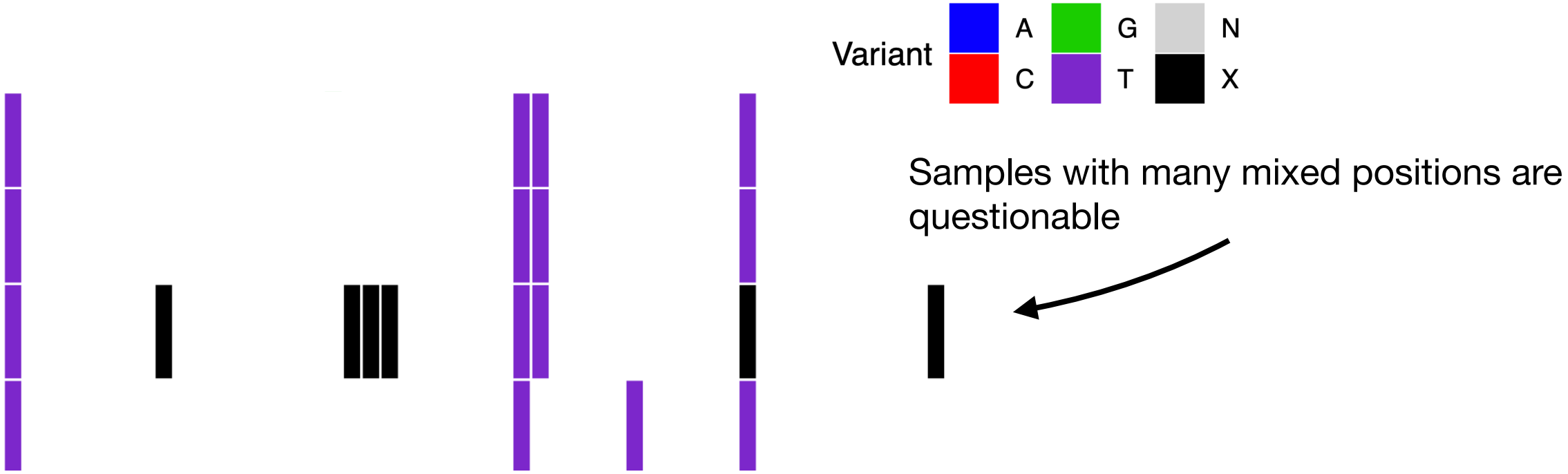
Possible RNA edits, RT or PCR errors from low-template samples



Variant QC - Mixed/Ambiguous Bases

Ambiguous bases (IUPAC symbols) are displayed using black rectangles

We QC-fail samples that have 5 or more ambiguous positions

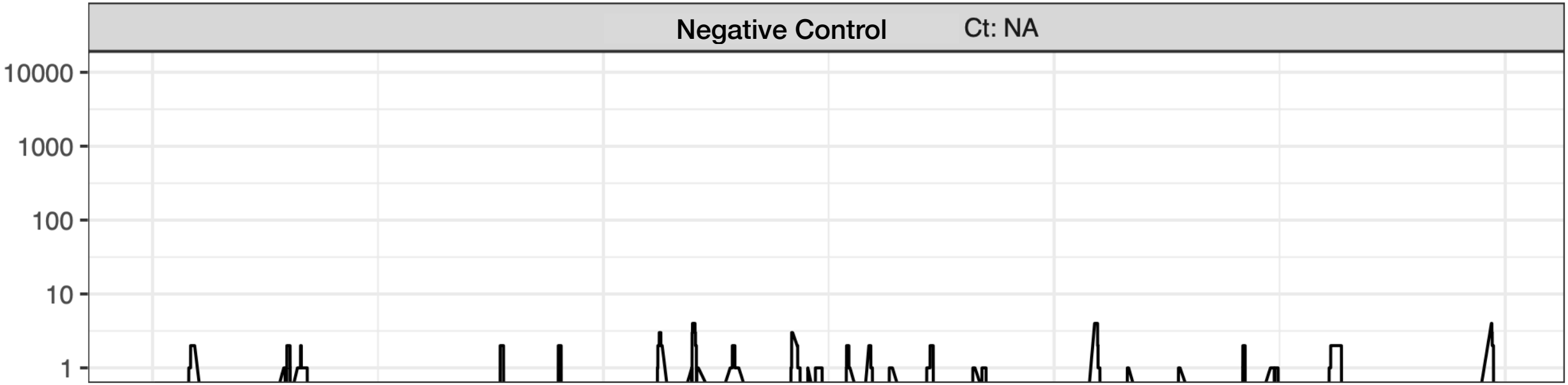


What do we mean by quality?

- Was the genome successfully sequenced?
  - Virus abundance varies sample-to-sample, storage conditions and extraction methods may differ in ability to recover complete genomes
- Is the genome sequence accurate?
  - Low abundance samples may have amplification artifacts, low coverage may lead to consensus errors
- **Is the sequencing run contaminated?**
  - Sequencing protocols are highly multiplexed, many amplification cycles increases risk of cross-contamination on the plate

Contamination - Assess by Negative Control

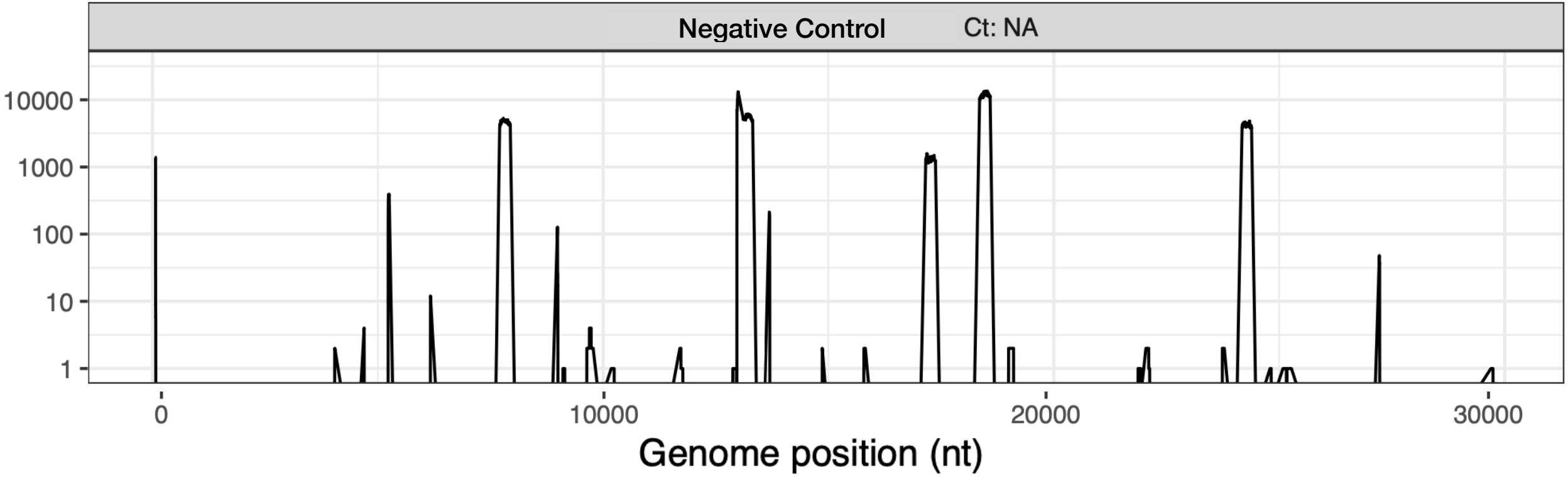
Ideal: very few reads mapped to genome



Contamination - Assess by Negative Control

Entire amplicons present in negative control

Recommendation: discard run, or mask these regions



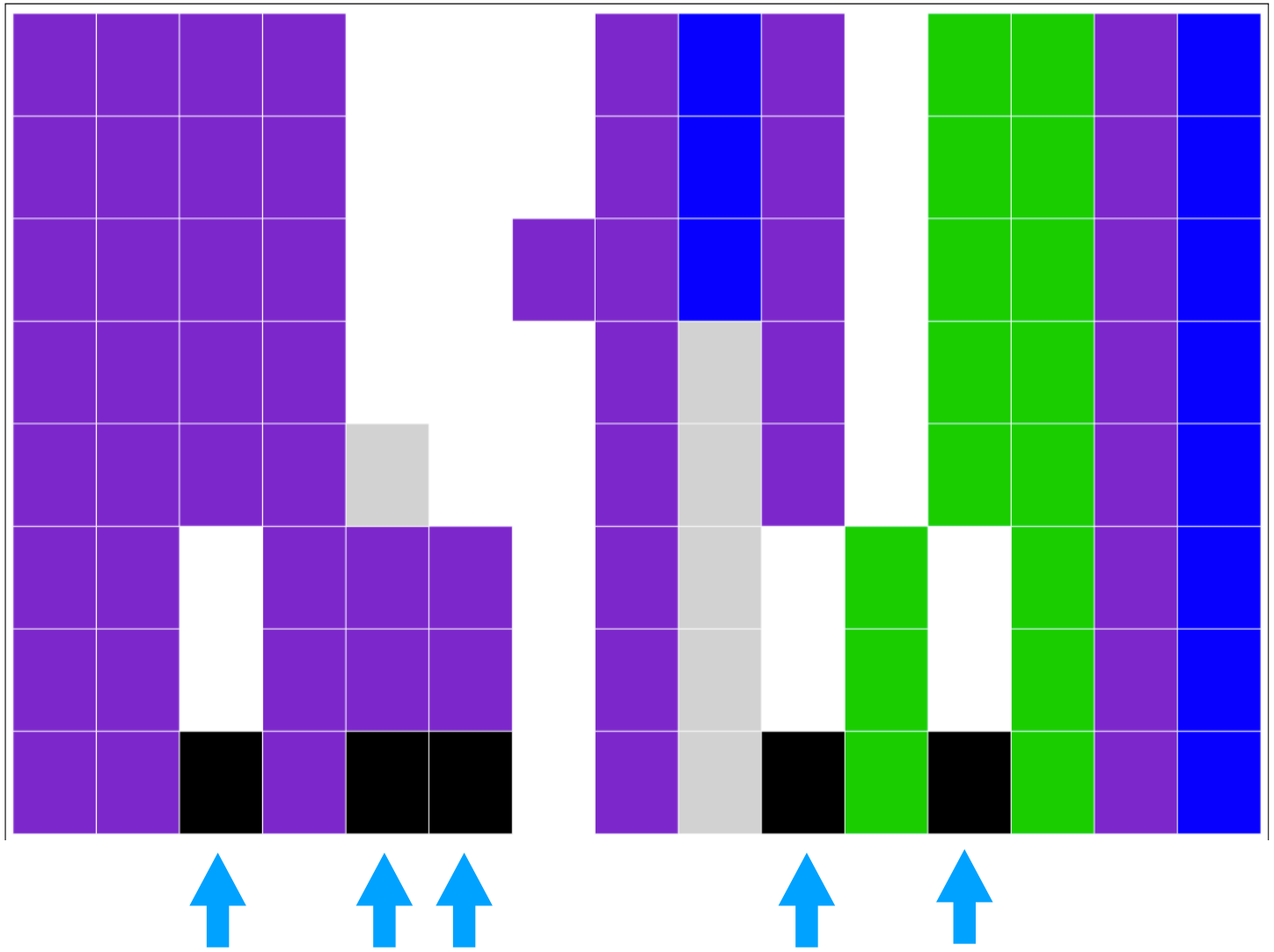
Sample contamination

SampleC has evidence of alleles from both sampleA and sampleB
Recommendation: discard sample after investigation

sampleA

sampleB

sampleC



We are on a Coffee Break & Networking Session

Workshop Sponsors: