

A Bioinformatics Pipeline for Characterizing SARS-CoV-2 Viral Stocks

P. Ford Combs, PhD; Nikhita Puthuveetil, MS; Amy L. Reese, MS; David Yarmosh, MS; Marco Riojas, PhD; John Bagnoli, BS; Briana Benton, BS; Jonathan L. Jacobs, PhD
ATCC, Manassas, VA 20110

Introduction

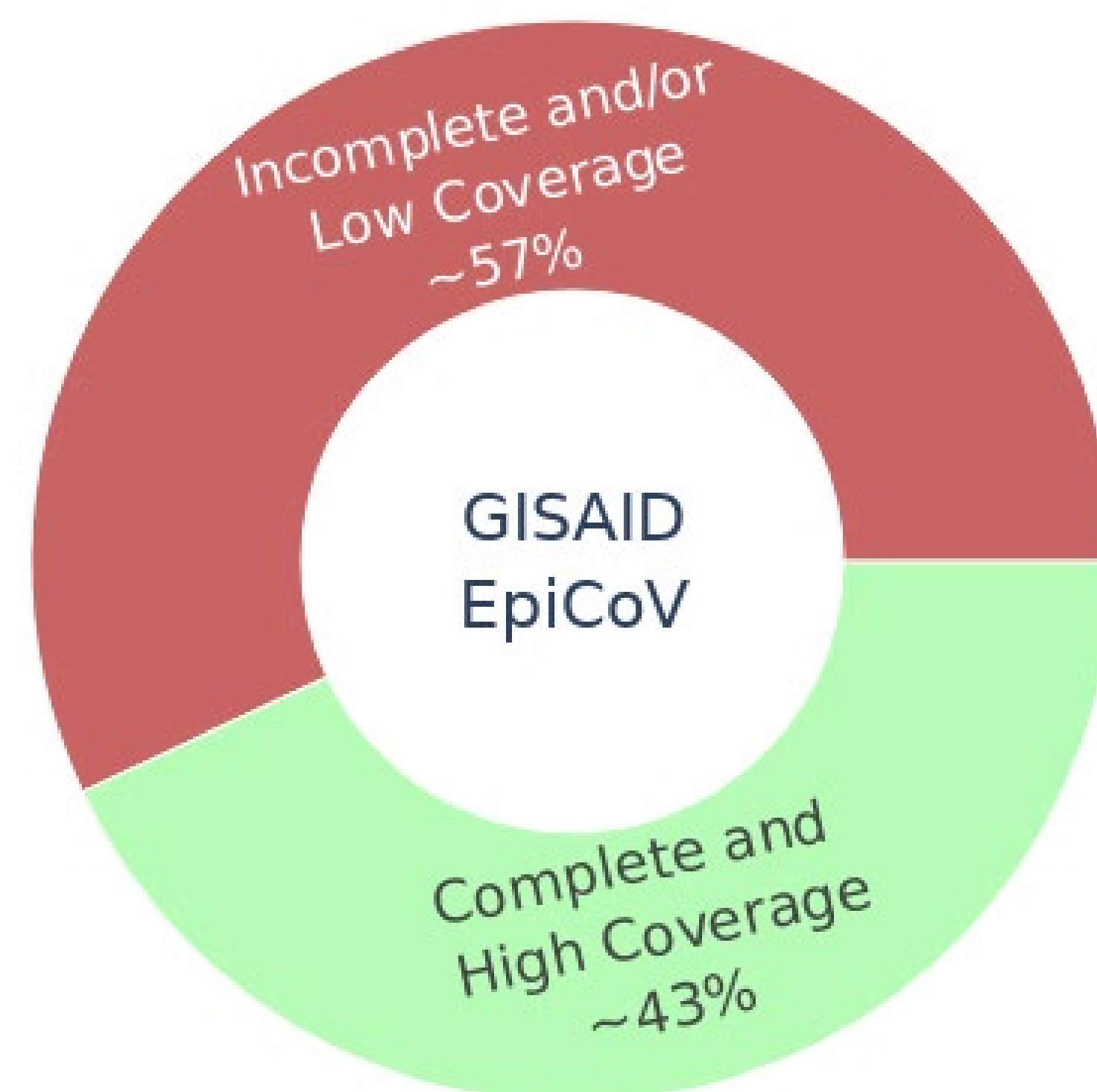


Figure 1: As of 13 September 2022, GISAID contained 13,081,886 SARS-CoV-2 sequences of which only 43% were marked as complete with high coverage, leaving 57% sequences to be incomplete and/or low coverage.

Vaccine and therapeutic efficacy differ between SARS-CoV-2 variants; therefore, it is critical to accurately determine the identity of viral stocks. While clinical isolates are often linked to sequences found in the EpiCoV database of GISAID, many of these sequences are incomplete and/or contain long stretches of ambiguous nucleotides (Ns), creating difficulties for NGS analysis (Figure 1). To address this issue, the Sequencing and Bioinformatics Center (SBC) of the American Type Culture Collection (ATCC) has developed a pipeline to verify the genomic contents of viral stocks (Figure 2).

Determining Lineage

Since the initial SARS-CoV-2 outbreak, many variants have emerged and the EpiCoV database has swollen with sequences, many of which contain long stretches of Ns. When a stretch of Ns is longer than the length of a read, reads cannot be mapped to that region and variants cannot be called. If lineage-defining mutations occur in these regions, then the variant identity of a stock cannot be verified.

To address this issue, the SBC pipeline analyzes the sample reference sequence (SRS) for long stretches of Ns and replaces them with the associated nucleotides from the ancestral sequence (AS), MN908947.3. The resulting sample consensus sequence (SCS) can be used for analysis or as the query in a GISAID AudacityInstant search for a more complete reference to be used in a rerun of the pipeline.

Pipeline Overview

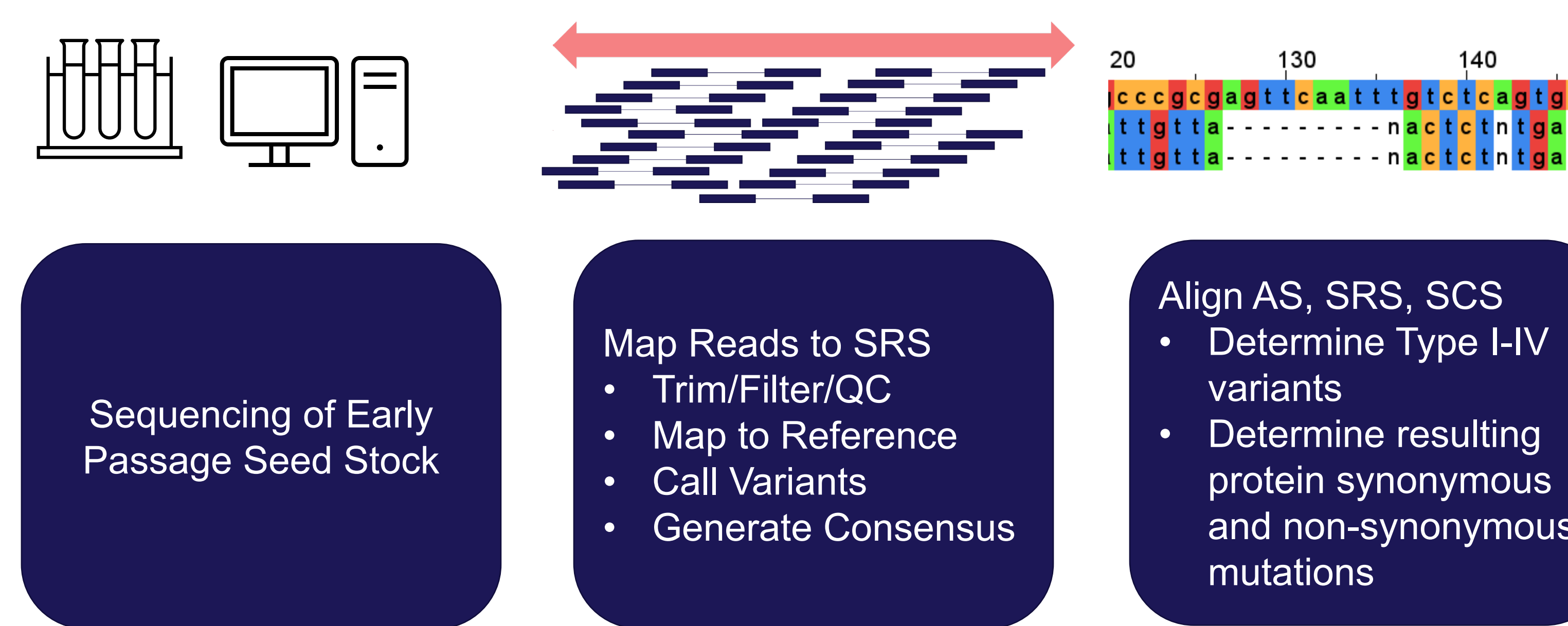


Figure 2: The pipeline begins with early passage seed stock, limiting the amount of mutation due to passaging. Next, the reads are mapped to the SRS, variants are called, and the SCS is generated. The AS, SRS, and SCS are then aligned and variant type and resulting amino acid mutation are determined.

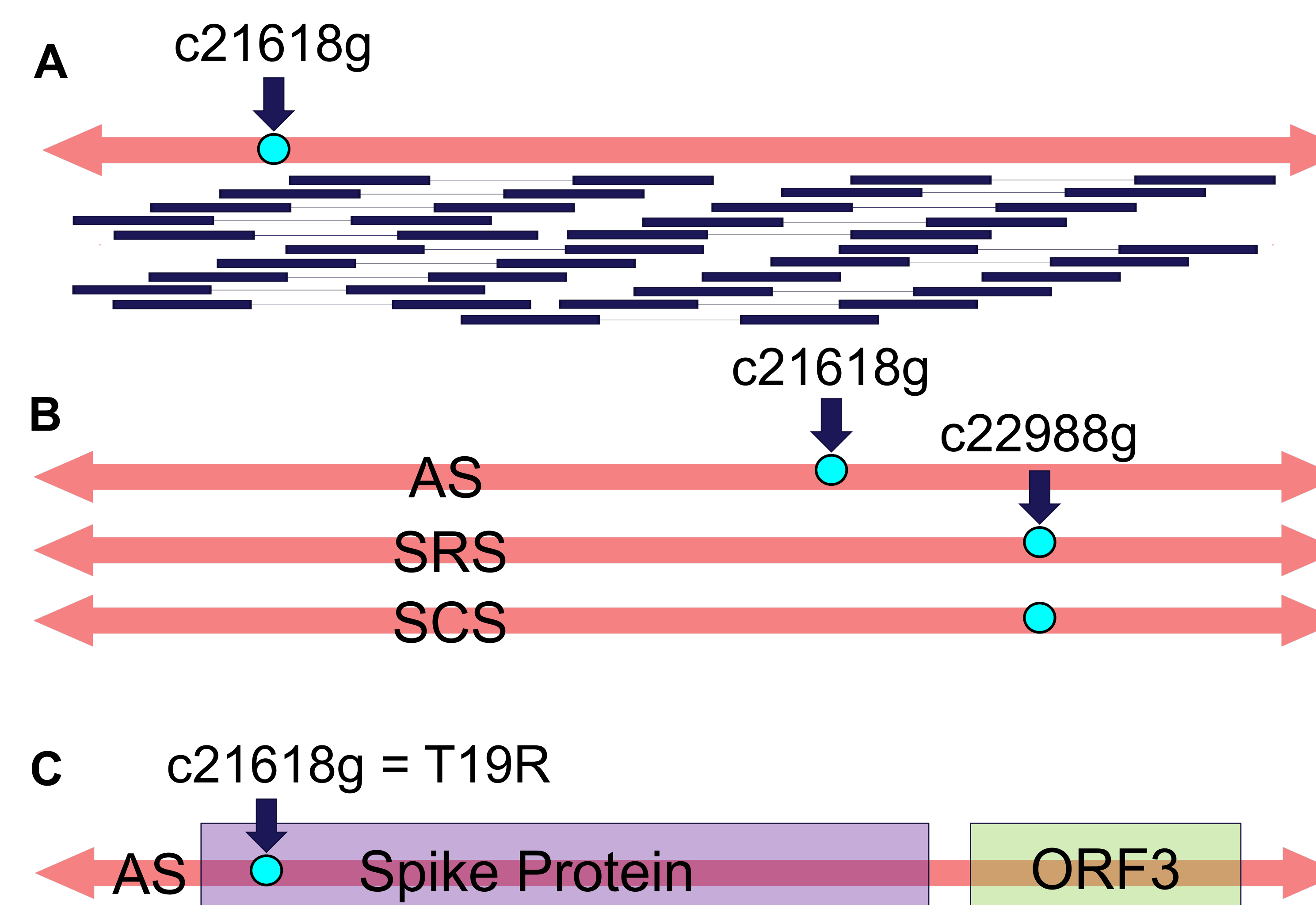


Figure 3: (A) Demonstrates a variant called by mapping reads to the SRS. (B) Shows how a variant called between the SCS and the SRS can differ from its associated position in the AS. (C) Shows how the relative position in the ancestral strain can be used to determine the amino acid mutation caused by the variant, without requiring the annotation of the SCS.

Variant Type

The multiple sequence alignment (MSA) of the AS, SRS, and SCS allows for the classification of variants by type. Here, type is one of five possible permutations of agreement at a position in the MSA and is an indication of the presence of potential selective pressures. Examples of the different types are shown in Table 1, with the exclusion of the fifth, where all nucleotides are identical and there is no variant.

Table 1: Example of types of variants found in MSA. Type I variants occur where the SRS and SCS are the same, but differ from the AS. These variants are expected and do not indicate the potential presence of selective processes. Types II, III, and IV are all potential indicators of selection, because they indicate a shift of the sample away from the SRS.

Variant Type	I	II	III	IV
AS	A	A	A	A
SRS	T	A	T	T
SCS	T	T	A	C

Conclusion

SARS-CoV-2 lineages are determined by the presence of characteristic nucleotide and amino acid mutations. These mutations can arise due to environmental or laboratory-based selective pressures and their existence determines the phenotype of the variant. Because mutations affect therapeutic and vaccine efficacy, it is critical to verify the mutations present in a viral stock.

The SBC's pipeline assigns each variant in the SCS a variant type (Table 1), which helps identify the presence of laboratory-based selective pressures. Furthermore, the pipeline accurately computes the resulting amino acid mutations (Figure 2). Together, these steps allow for the verification of identity of the sample.