**Supplemental Table S1: Glossary of Next-Generation Sequencing (NGS) and Bioinformatics Pipeline–Related Terminologies.**

|  |  |
| --- | --- |
| Terminology | Description |
| BAM | Compressed (binary) format of a SAM file intended for faster random access (search) of aligned and unaligned sequences and its related metadata. Compression enables smaller file size and storage efficiency which makes it popular over the SAM format. This is a default output of many alignment and post-alignment softwares used in bioinformatics pipeline and commonly used as an input by many variant callers\*. |
| FASTQ | It is a *de facto*, human readable, file format that stores nucleotide sequences and corresponding quality (PHRED) scores for each nucleotide as an ASCII encoded character.4 This is commonly used for storing unaligned short sequence reads after the steps of base calling and is a typical starting point for NGS bioinformatics pipeline. |
| PHRED Score | It is a per base (nucleotide) quality score that is defined as an estimated probability for a called base to be incorrect (erroneous call). Mathematically, it is expressed as4  where *Q* is Phred quality score, *Pe* is the probability for an erroneous base call. The *Pe* is typically generated by the base calling software which is sequence instrument specific. Therefore, *Q* values in isolation cannot be used to compare sequence quality across different sequencing platforms. |
| SAM | Stands for Sequence Alignment/Map format. It is a human readable (text file) file format specification for storing information on aligned sequence. This is a default output of many alignment softwares used in bioinformatics pipeline. Given the large file size and slower random access, BAM format is preferred for routine bioinformatics data processing. This format is helpful for technical troubleshooting when manual review of the stored information is necessary\*. |
| Variant - horizontally complex | When two or more sequence alterations are present on the same read in close proximity such that they may represent a single complex variant. These variants are frequently represented as deletion-insertions and may result in ambiguous sequence description or HGVS nomenclature. |
| Variant - Left-aligned | If there are multiple potential VCF entries of the same allele length that represent the same variant, then left-alignment refers to the VCF entry with the smallest base position. The base position is typically represented in genomic coordinate for a given primary assembly (eg, GRCh38) and represents the most 5’ position.5 |
| Variant - Normalized | A normalized variant must be parsimonious as well as left-aligned.5 |
| Variant - Parsimony | If there are more than one way to represent the same variant in a VCF file, parsimony refers to the representation with the shortest possible allele length5  (positive and non-zero length). |
| Variant - vertically complex | A vertically complex variant occurs when three or more alleles are represented by different sequence reads, typically with or uncommonly without a reference (normal) allele, at the same genomic coordinate or set of coordinates. |
| VCF | Variant Call Format is a versioned, text-file (human readable) specification for storing sequence variant calls. The file contains meta-information containing various details of the variant calling process and definition of headers and format tags, a header line and data lines. Each data line represents a sequence variant defined using a combination of chromosome, position, reference allele, and alternate allele†. |

\* Sequence Alignment/Map Format Specification, https://samtools.github.io/hts-specs/SAMv1.pdf, last accessed 9/28/17.

†The Variant Call Format (VCF) Version 4.2 Specification, https://samtools.github.io/hts-specs/VCFv4.2.pdf, last accessed 9/28/17.