Guideline-driven nomenclature for variant analysis

In this Technical Note, we outline the guidelines behind the nomenclature convention employed in Alamut™ Visual Plus and how to apply them. We also explain how Alamut™ Visual Plus applies these guidelines to ensure that variant annotation follows the universally applied standards for variant analysis and interpretation.

HGVS Nomenclature

What is it?

The Human Genome Variation Society (HGVS) nomenclature standard was developed to prevent the misinterpretation of variants in DNA, RNA, and protein sequences. The HGVS nomenclature standard is used worldwide and is authorized by the Human Genome Organisation (HUGO).[1,2]

HGVS follow recognized standards for the nomenclature of DNA and RNA nucleotides, the genetic code, amino acid descriptions, and cytogenetic band position in chromosomes.[3]

- DNA- and RNA-level nomenclature is based on *Nomenclature for Incompletely Specified Bases in Nucleic Acid Sequences* guidelines
- Protein-level nomenclature is based on *Nomenclature and Symbolism for Amino Acids and Peptides* guidelines, and additionally uses "Ter" and "*" to indicate a translation termination (stop) codon
- Nomenclature for numerical and structural chromosomal changes detected using microscopic and cytogenetic techniques is based on the *International System for Human Cytogenomic Nomenclature (ISCN)*

How does it work?

The HGVS recommendations for sequence variant nomenclature state that the format of a complete variant description should include the reference sequence followed by the description and then the predicted consequence in parentheses. E.g. NM-004006.2:c.4375C>T p.(Arg1459*) (Figure 1).

Reference sequence

The HGVS recommendations for sequence variant nomenclature state that a complete variant description should begin with the reference sequence.[1] The reference sequence accession number begins with a two-letter abbreviation (Table 1), followed by a multi-digit number and version number.

Abbreviation	Reference sequence based on a:
NC	Chromosome
NG	Gene or genomic region
LRG	Locus Reference Genomic sequence: Gene or genomic region, used in a diagnostic setting
NM	Protein-coding RNA (mRNA)
NR	Non-protein-coding RNA
NP	Protein (amino acid) sequence

Table 1 | Meaning of the two-letter abbreviation at the beginning of a reference sequence accession number.

HGVS General Terminology Recommendations[1]

Mutation or polymorphism	Variant, change, allelic variant *Can be used for cancer tissue:* Mutation load and tumor mutation burden
Pathogenic	Affects function, disease-associated, phenotype-associated

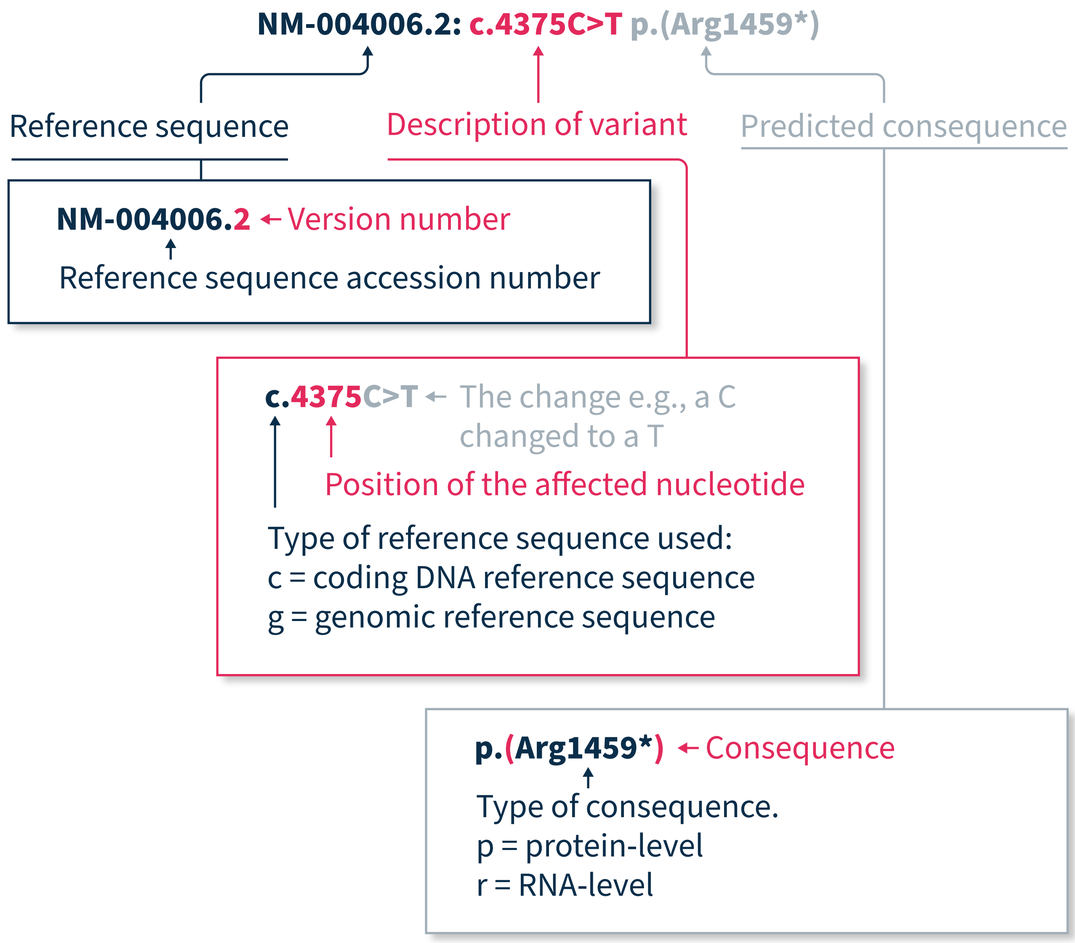


Figure 1 | Application of the HGVS recommendations for sequence variant nomenclature

Description of variant

The variant description begins by depicting the type of reference sequence used (c = coding DNA sequence, g = genomic reference sequence). When a protein-coding reference sequence is used (c), the nucleotide numbering begins with a 1 representing the first position in the protein-coding region (the A of the translation-initiating ATG) and ends at the last position of the stop codon. Thus, if you divide the position number by 3, you can identify the affected amino acid in the protein sequence e.g., using the same example, 4375/3=1459, indicating that the predicted consequence affects Arg1459. Different variants are indicated using different notations (Table 2).

Notation	Example	Explanation
>	c.4375C>T	Substitution of the C nucleotide at position c.4375 with a T
del	c.4375_4379del or c.4375_4379delCGATT	Nucleotides from position c.4375 to c.4379 deleted
dup	c.4375_4385dup or c.4375_4385dupCGATTATTCCA	Nucleotides from position c.4375 to c.4385 duplicated
ins	c.4375_4376insACCT	ACCT inserted between positions c.4375 and c.4376
delins	c.4375_4376delinsACTT or c.4375_4376delCGinsACTT	Nucleotides from position c.4375 to c.4376 (CG) are deleted and replaced by ACTT

Table 2 | HGVS notation for the most common types of variants[2]

Predicted consequence

When only DNA has been analyzed, the RNA- and protein-level consequences of the variant can only be predicted, and should thus be reported in parenthesis e.g., p.(Arg1459*) is the predicted effect at protein-level (p) for the example described above.

The 3' rule

For all variant descriptions using HGVS nomenclature, the nucleotide at the most 3' position of the variation in the reference sequence is arbitrarily assigned to have changed (see how to apply this rule in Figure 2).[4] Except for deletions/duplications around exon junctions where shifting the variant 3' would place it in the next exon.[5]

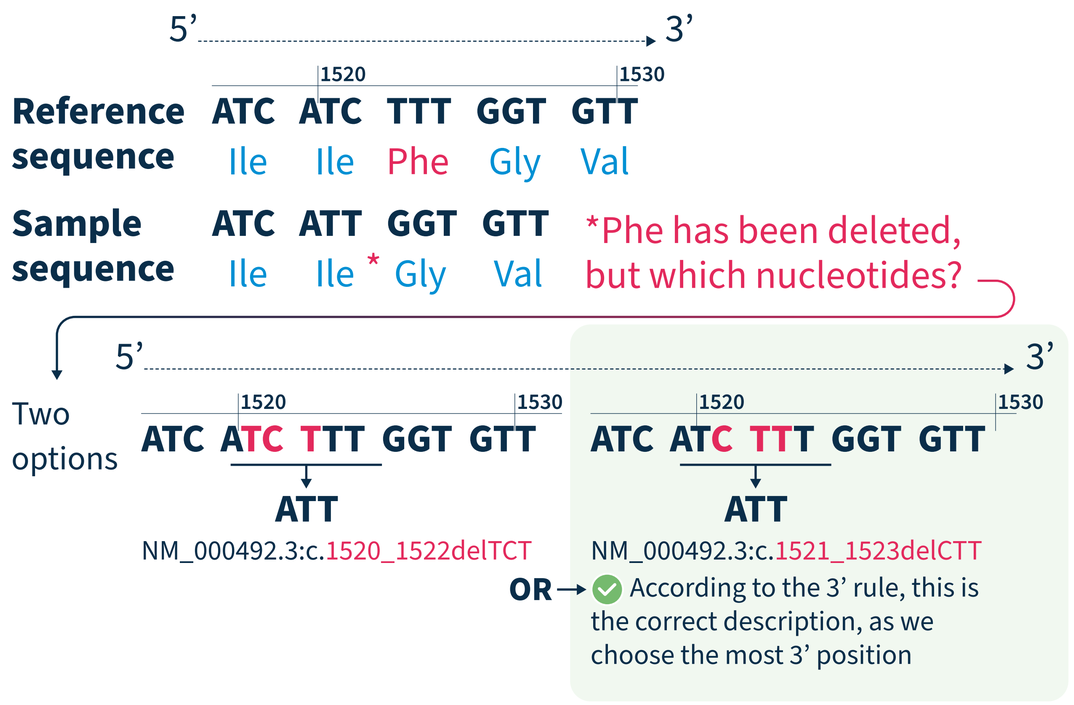


Figure 2 | Application of the 3' rule using the HGVS recommendations for variant nomenclature.

How does Alamut™ Visual Plus apply HGVS nomenclature?

Alamut™ Visual Plus currently applies the internationally recognized **version 20.05 of the HGVS nomenclature** (as of August 2022). If the variant is located at an **intron-exon junction** and at least 1 nucleotide is located in an exon, the variant is determined as being located in the exon. Alamut™ Visual Plus applies the **3' rule** at the genome level (choosing the most 3' position in the forward sequence) and at cDNA level (choosing the most 3' position in the transcript). This means the cDNA-level variant positions on the reverse strand do not always map to the same genome-level positions.

Genome Reference Assemblies

What are reference genome builds and who develops them?

In 2007, the **Genome Reference Consortium (GRC)** was founded to improve the human, mouse, and zebrafish reference genome assemblies for genetic research and analysis.[6] The GRC is a collaboration between The Wellcome Sanger Institute (represented by the Genome Reference Informatics Team), the McDonnell Genome Institute at Washington University (MGI), the European Bioinformatics Institute (EBI), and The National Center for Biotechnology Information (NCBI), with the NCBI hosting the assembly homepages. The GRC have a mission to continue to improve the human reference assembly by correcting errors and adding sequences to ensure that it provides the best representation of the human genome to meet basic and clinical research needs.[5] The most current reference assembly (as of August 2022) is Genome Reference Consortium Human Build 38 Referencepatch release 14 (**GRCh38.p14**), also known as human genome version 38 (**hg38**). Before GRCh38, the 2009 assembly was **GRCh37 (hg19)**.

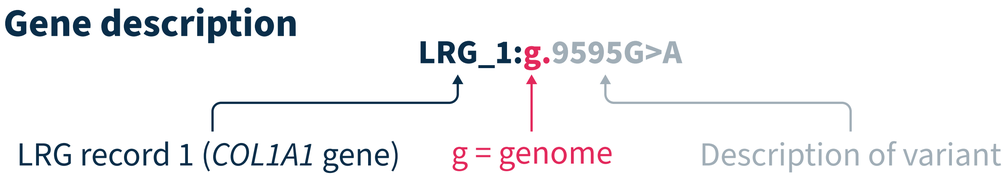
Reference genome builds		
GRCh38 (hg38)	GRCh37 (hg19)	LRG

In addition to GRC genome builds, **Locus Reference Genomic (LRG)** records are internationally recognized stable reference sequences specifically designed for the reporting of clinical and clinical research-relevant variants.[8] The aim of the LRG project was to provide stable, and thus un-versioned, reference sequences that were independent of changes to transcript models and reference genome assemblies and that did not change over time.[9]

The use of multiple sequences for a given locus and confusion over variants can result in inconsistent variant reporting. LRGs provide a stable framework for the reporting of clinical research-relevant variants in genomic DNA, transcript, or protein coordinates, using HGVS nomenclature with mapping to GRCh37 and GRCh38. LRG reference sequences are manually curated in collaboration with diagnostic and research communities, locus-specific database curators, and mutation consortia, compiled and maintained by the NCBI and EBI.

The LRG-specific exon numbering system is based on the transcript(s) in the LRG, with each exon numbered consecutively from 5' to 3'. Variant nomenclature adheres to the HGVS standard. For example, the *COL1A1* gene is represented by LRG record 1 (LRG_1) which has a single transcript (t1) and a single corresponding protein (p1). A variant in the *COL1A1* gene can be described in three different ways, as a single LRG record contains genomic DNA, mRNA, and protein sequences. The three corresponding descriptions for a variant in *COL1A1* are LRG_1:g.9595G>A, LRG_1t1:c.769G>A and LRG_1p1:p.Gly257Arg (illustrated in Figure 3).

Gene description



mRNA description

Protein description

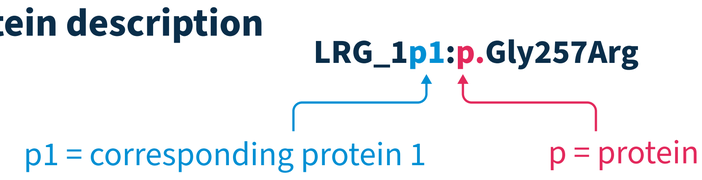


Figure 3 | LRG variant nomenclature with LRG_1 (*COL1A1* gene) as the reference genome build.

New LRGs ceased to be generated in March 2021, with Ensembl and RefSeq transcripts specified by the MANE collaboration now preferred for the standardization of reporting (read more in the Transcripts section of this app note).

Which reference assemblies are available in Alamut™ Visual Plus?

Alamut™ Visual Plus allows the visualization of transcripts and associated variants on the two most recent reference genome builds, **GRCh37 (hg19)** and **GRCh38 (hg38)**, as well as **LRG** records (Figure 4).

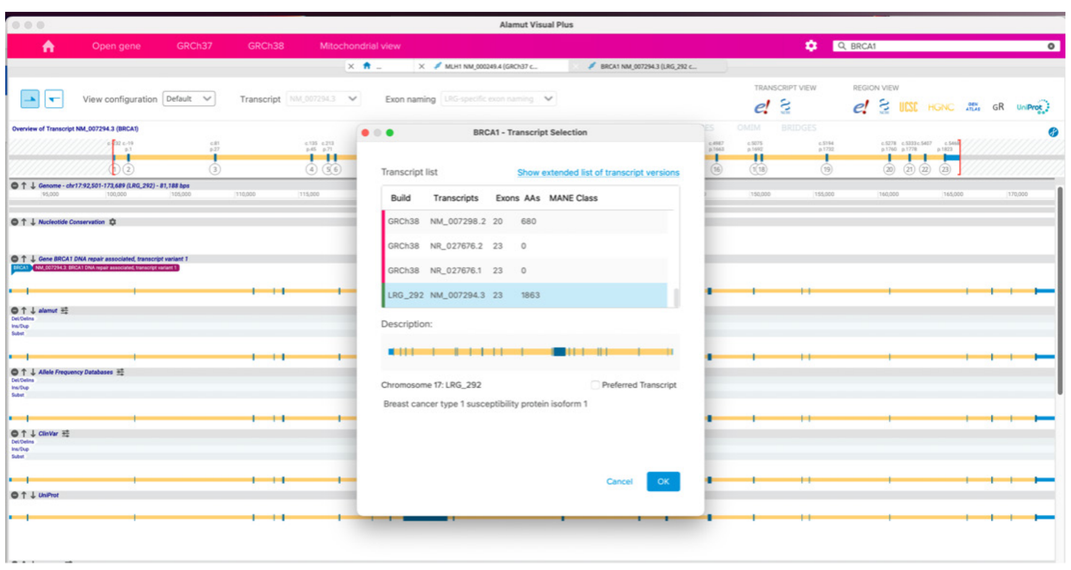


Figure 4 | Transcript selection for the *BRCA1* gene in Alamut™ Visual Plus showing reference genome build, transcript, number of exons, number of amino acids, and MANE class, with the LRG build highlighted at the bottom of the list.

Transcripts

What are MANE transcripts?

MANE (Matched Annotation from NCBI and EMBL-EBI) is a collaborative project between EMBL-EBI's Ensembl/GENCODE and NCBI's RefSeq with the aim of converging human gene and transcript annotation to ensure the consistent analysis and reporting of variants in clinical genomics and research.[10,11] The MANE initiative is developing a genome-wide representative transcript set that identifies **one well-supported transcript and corresponding protein for each protein-coding gene (MANE Select)** to be used as a universal standard for variant reporting and browser display. All MANE Select transcripts completely align to GRCh38 and represent an exact match between exonic sequences of RefSeq (NM) and Ensembl (ENST) transcripts, such that the identifiers can be used synonymously. MANE Select transcripts are chosen based on biologically relevant criteria such as transcript expression levels and conservation of coding regions, identified using computational methods followed by manual review and discussion. Additionally, the **MANE Plus Clinical** set provides additional transcripts for genes where Select transcripts are not sufficient to report all currently known disease-associated variants.

Which transcripts are recommended for variant interpretation and reporting?

Both the HGVS and Morales J, et al. (in Nature 2022) recommend the use of MANE transcripts.[10,12] HGVS recommend that **MANE transcripts** be used when describing sequence variants using HGVS nomenclature, and Morales J, et al. recommend the use of MANE transcript sets as a reference standard for reporting. In comparison to LRG transcripts, MANE transcripts cover all protein-coding genes rather than a limited subset, and provide transcript annotations that align with the GRCh38 reference assembly. Although new LRGs are no longer generated, existing LRG records now incorporate MANE transcript annotation and will continue to be supported.

Which transcripts are used in Alamut™ Visual Plus?

Both **RefSeq (NM)** and **Ensembl (ENST)** transcripts are available in Alamut™ Visual Plus. When selecting a transcript, there will be up to four transcript versions available for each gene. The transcripts are regularly updated, with the aim of keeping the four most recent versions and no more. When selecting a transcript in Alamut™ Visual Plus, you will see the build, transcript, exon, and amino acid information for each transcript, with the appropriate **MANE Select** or **MANE Plus Clinical** transcript tagged in the MANE class column (Figure 5).

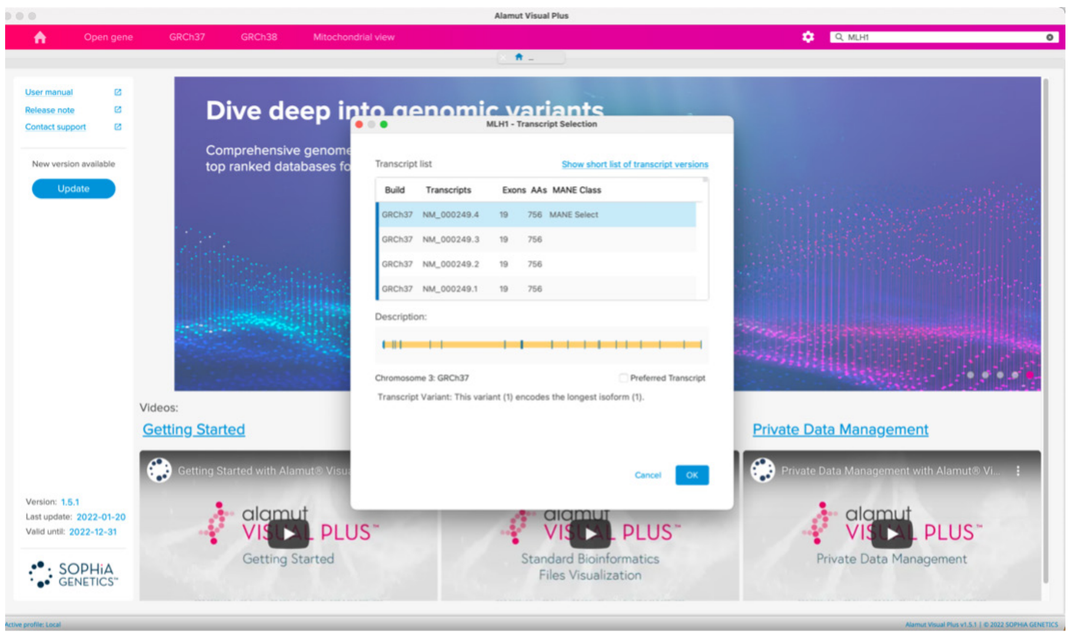


Figure 5 | Transcript selection for the *MLH1* gene in Alamut™ Visual Plus showing reference genome build, transcript, number of exons, number of amino acids, and MANE class, with the **MANE Select transcript** highlighted at the top of the list.

Why are there sometimes mismatches between genes and transcripts, or between RefSeq and Ensembl transcripts in Alamut™ Visual Plus?

Sometimes there are discrepancies between genes and transcripts **when the genome reference assembly includes polymorphism minor alleles but the corresponding transcript includes major alleles.** This results in some genomic variants being seen as "non-variants" in the transcript. In these instances, when the nucleotide in the transcript differs from the nucleotide in the genome build (GRCh37 or GRCh38), the nucleotide will be highlighted in **red** in Alamut™ Visual Plus. These discrepancies primarily occur in RefSeq transcripts (beginning with NM) as RefSeq do not make corrections to match the genome build, whereas Ensembl transcripts (beginning with ENST) are corrected to match the nucleotides present in the genome build.

Ensembl and RefSeq transcripts differ in that Ensembl transcripts are mapped onto the reference genome, whereas RefSeq transcripts are mapped onto mRNA sequences. Due to differences between reference genomes and individual mRNAs, some RefSeq mRNA's might not map perfectly to the reference genome, meaning that there may be small differences between Ensembl and RefSeq transcripts. Alamut™ Visual Plus uses Splign (a tool developed by RefSeq) to align all transcripts to the genome build.

Conclusion

In conclusion, Alamut™ Visual Plus follows a guideline-driven approach to variant nomenclature, applying universally accepted standards to ensure the consistent analysis and reporting of variants for clinical research.

References

1. https://varnomen.hgvs.org/bg-material/basics/
2. http://varnomen.hgvs.org/bg-material/simple/
3. http://varnomen.hgvs.org/bg-material/standards/
4. http://varnomen.hgvs.org/recommendations/general/
5. http://varnomen.hgvs.org/bg-material/numbering/#DNAc
6. https://www.sanger.ac.uk/data/genome-reference-consortium/
7. https://www.ncbi.nlm.nih.gov/grc
8. MacArthur JA, Morales J, Tully RE, et al. Locus Reference Genomic: reference sequences for the reporting of clinically relevant sequence variants. Nucleic Acids Res. 2014;42(Database issue):D873-D878.
9. https://www.lrg-sequence.org/
10. Morales J, Pujar S, Loveland JE, et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. Nature. 2022;604(7905):310-315.
11. https://www.ncbi.nlm.nih.gov/refseq/MANE/
12. den Dunnen JT. Describing Sequence Variants Using HGVS Nomenclature. Methods Mol Biol. 2017;1492:243-251.

About us

SOPHiA GENETICS (Nasdaq: SOPH) is a software company dedicated to establishing the practice of data-driven medicine as the standard of care and for life sciences research. We are the creator of the SOPHiA DDM™ Platform, a cloud-native platform capable of analyzing data and generating insights from complex multimodal data sets and different diagnostic modalities. The SOPHiA DDM™ Platform and related applications, modules, and services are currently used by a broad network of hospital, laboratory, and biopharma institutions globally.

Where others see data, we see answers.

Want to know more?

Contact us at: info@sophiagenetics.com