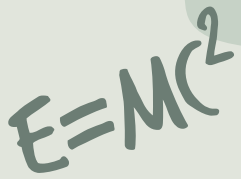
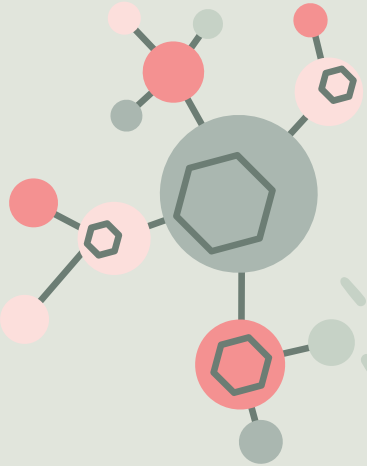




HGVS - BASICS

Presenter: Phan Thi Le Hang

Introduction

The concept of HGVS nomenclature is very familiar to those who work in the genomics field, but to those who do not, it may seem like an alien language. This is because it looks like a random mix of numbers and special characters. Though it may seem to be written in an indecipherable way, in fact, it accurately describes genetic variants according to strict rules and definitions.

Outline

01

What is HGVS nomenclature?

02

Reference Sequence

03

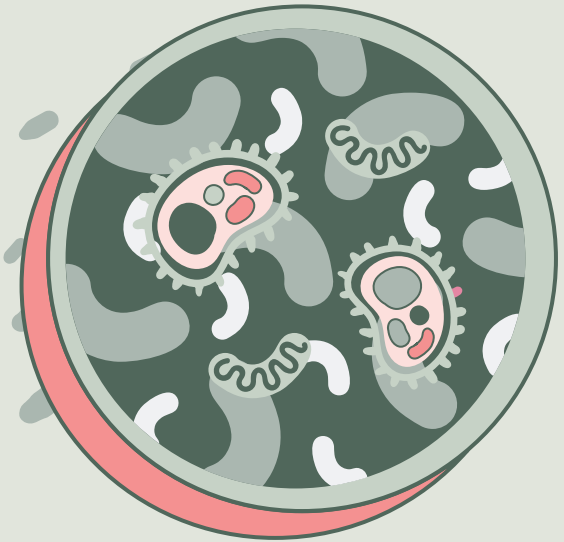
Standards

04

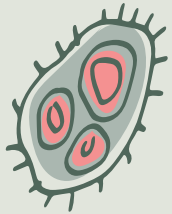
HGVS simple

01

What is HGVS nomenclature?

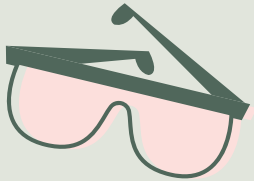


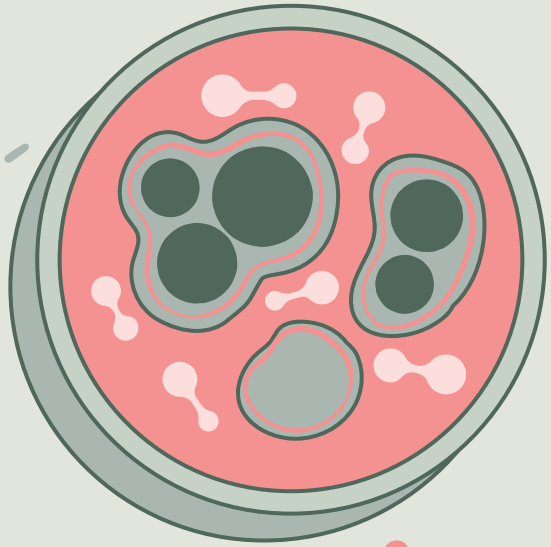
What is HGVS nomenclature?



HGVS stands for **Human Genome Variation Society** and is the name of the most prominent international academic organization that studies the human genome. HGVS nomenclature refers to the genetic variant nomenclature recommended by the Human Genome Variation Society.

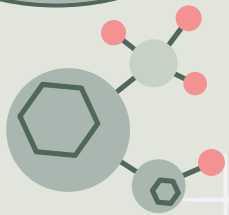
In 2000, **HGVS** first proposed rules and definitions for describing genetic variants, which have been gradually expanded and applied in various places, and today, they have become the global standard for describing genetic variants.





02

Reference Sequence



Reference Sequence

In order to have a unique description (that prevents confusion), you must include a reference sequence when using HGVS
BRCA1 c.4366A>G

	Option 1	Option 2
NM_007294.3	**c.4366A>G**	c.4358-2777A>G
NM_007300.3	c.4429A>G	**c.4366A>G**
Genomic (GRCh37)	chr17:g.41228623T>C	chr17:g.41231408T>C

Reference sequence used must contain the variant residue described - a coding DNA reference sequence does not contain intron and therefore cannot be used to describe intron variants

- **not correct:** NM_004006.2:c.357+1G>A
- **correct:** NG_012232.1(NM_004006.2):c.357+1G>A

Reference Sequence

Only public files from NCBI or EBI are accepted as reference sequence files

Approved reference sequence formats include:

- NC_# (e.g. NC_000023.10)
- LRG_# (e.g. LRG_199, LRG_199t1)
- NG_# (e.g. NG_012232.1)
- NM_# (e.g. NM_004006.2)
- NR_# (e.g. NR_002196.1)
- NP_# (e.g. NP_003997.1)

Type of reference sequence indicated by letter used:

c. (coding); g. (genomic); m. (mitochondrial); n. (non-coding); r. (RNA); p. (protein)

General Information

All variants should be described at the most basic level, the **DNA** level. Descriptions at the RNA and/or protein level may be given in addition.

Descriptions should make clear whether the change was **experimentally determined** or **theoretically** deduced by giving predicted consequences in parentheses

E.g.

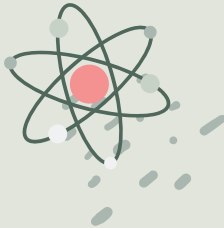
- NP_003997.1:p.(Trp24Cys) means amino acid Trp24is predicted to change to a Cys (no experimental proof, e.g. based on DNA level data)
- NP_003997.1:p.Trp24Cys means amino acid Trp24 is changed to a Cys (confirmed via RNA or protein sequence analyzed)

General Information

Prioritization: when a description is possible according to several types, the preferred description is: (1) deletion, (2) inversion, (3) duplication, (4) conversion, (5) insertion

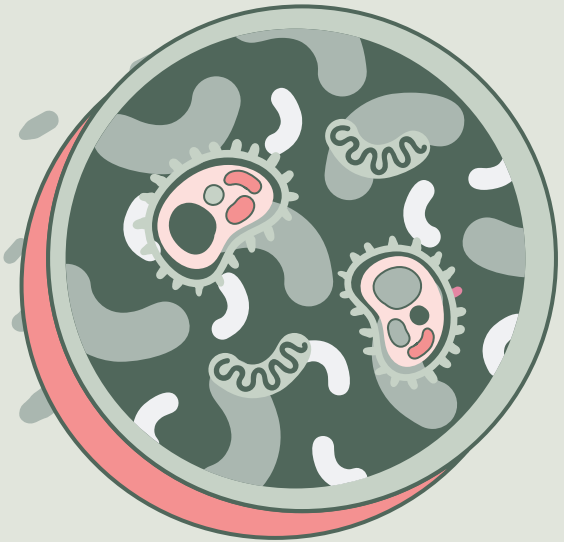
Descriptions at DNA, RNA and protein level differ:

- DNA-level 123456A>T: number(s) referring to the nucleotide(s) affected, nucleotides in CAPITALS
- **RNA-level** 76a>u: number(s) referring to the nucleotide(s) affected, nucleotides in lower case
- **protein level** Lys76Asn: the amino acid(s) affected in 3- or 1-letter followed by a number (* three-letter amino acid code is preferred)

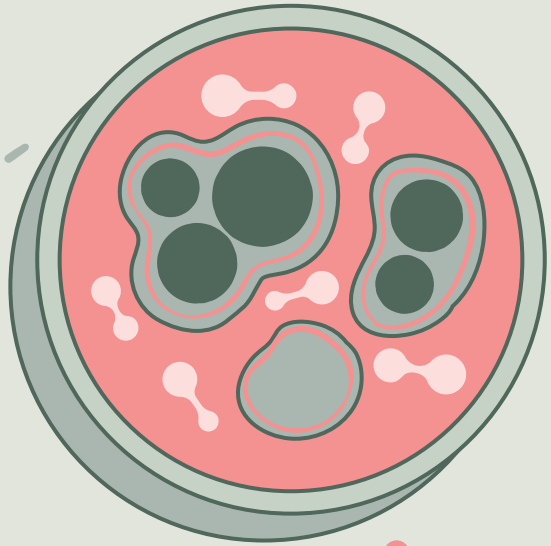


03

Standards

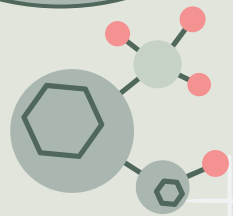


Standards: **https://varnomen.hgvs.org/bg-material/standards/**



04

HGVS simple



The Format

The format of a complete variant description is “**reference : description**” (spaces added for clarity only), e.g.;

```
*       NM_004006.2:c.4375C>T
*       NC_000023.11:g.32389644G>A
```

- **NM_004006.3, NC_000023.11** so called **reference sequence**
- After the reference a description of the variant is given, in the examples **c.4375C>T** and **g.32389644G>A.**

DNA > RNA > protein

In nature the DNA code is first transcribed in to a RNA molecule (see Wikipedia). Next, there are **two options:**

- the **RNA** molecule is translated in to a protein and the protein is the final product of a gene. Proteins perform a vast array of functions, including catalysing metabolic reactions, DNA replication, responding to stimuli, providing structure to cells, and organisms, transporting molecules from one location to another, etc.
- the **RNA** molecule is the final product of the gene (so the RNA is not translated in to a protein). RNA molecules perform a vast array of functions, including e.g. rRNAs (ribosomal RNA) and tRNAs (transfer RNAs) both active in protein translation.

Genomic reference sequences

- For human the reference sequence **accession number** directly in front of the version number gives the number of the chromosome: 1-22, 23 for the X-chromosome and 24 for the Y-chromosome. In NC_000023.10 this number is “23” so a reference sequence of human **chromosome X.**
- Genomic reference sequences can also be based on smaller sequences not covering an entire chromosome. They usually cover only a specific gene or specific genomic segment. The most frequently used are LRG’s (Locus Genomic Reference sequences, format LRG_199) or NG’s (RefSeq Gene reference sequences, format NG_012232.1).

Coding DNA reference sequences

- In a human diagnostic setting the most frequently used reference is a “**coding DNA reference sequence**” (description starting with “c.”, e.g. NM_004006.3:c.4375C>T).
- Variant descriptions based on this format are very popular because they directly link to the encoded protein.
- In protein coding DNA reference sequences numbering starts with 1 at the first position of the protein coding region, the A of the translation initiating ATG triplet.
- Numbering ends at the last position of the ending triplet, the last position of the translation stop codon (TAA, TAG or TGA).

Variants

prefix.position(s)_change

substitution

one letter (nucleotide) of the DNA code is replaced (substituted) by one other letter. On DNA and RNA level a substitution is indicated using ">".

deletion

one or more letters of the DNA code are missing (deleted). A deletion is indicated using "del".

duplication

one or more letters of the DNA code are present twice (doubled, duplicated). A duplication is indicated using "dup".

insertion

one or more letters in the DNA code are new (inserted). An insertion is indicated using "ins".

deletion/insertion (indel)

one or more letters in the DNA code are missing and replaced by several new letters. A deletion/insertion is indicated using "delins".

Eg Variants

substitution

c.4375C>T
the C nucleotide at position c.4375 changed to a T

deletion

c.4375_4379del
the nucleotides from position c.4375 to c.4379 (CGATT) are missing (deleted). Also reported as c.4375_4379delCGATT.

duplication

c.4375_4385dup
the nucleotides from position c.4375 to c.4385 (CGATTATTCCA) are present twice (duplicated). Often reported as c.4375_4385dupCGATTATTCCA or c.4385_4386insCGATTATTCCA (not a correct HGVS description).

insertion

c.4375_4376insACCT
the new sequence “ACCT” was found inserted between positions c.4375 and c.4376.

deletion/insertion (indel)

c.4375_4376delinsAGTT
the nucleotides from position c.4375 to c.4376 (CG) are missing (deleted) and replaced by the new sequence “AGTT”. Also reported as c.4375_4376delCGinsAGTT.

Aliases

It should be noted that one variant, based on different reference sequences used, can be described in many different ways.
Variant c.5234G>A in the DMD gene can be described based on different genomic reference sequences
(e.g. NC_000023.9:g.32290917C>T, NC_000023.10:g.32380996C>T, NC_000023.11:g.32362879C>T, NG_012232.1:g.981731G>A, LRG_199:g.981731G>A) as well as different coding DNA reference sequences (e.g. LRG_199t1:c.5234G>A, NM_004006.3:c.5234G>A, NM_004009.3:c.5222G>A, NM_000109.3:c.5210G>A, NM_004007.2:c.4865G>A, NM_004010.3:c.4865G>A, NM_004011.3:c.1211G>A, NM_004012.3:c.1202G>A).

Thank you!