General

Since references to web sites are not yet acknowledged as citations, please mention Den Dunnen et al. (2016) HGVS recommendations for the description of sequence variants: 2016 update. Hum.Mutat. 25: 37: 564-569 (http://onlinelibrary.wiley.com/doi/10.1002/humu.22981/pdf) when referring to these pages. Note that although the examples on these pages mainly give examples for human (*Homo sapiens*), the recommendations can be applied to **all species**.

Make sure you have also seen the (*Basics* (/bg-material/basics/), explaining the **history** of these recommendations, the process of making **changes**, the **versioning** of the recommendations and important remarks on **terminology**.

General recommendations

- 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
    - descriptions at RNA/protein level should describe the changes observed on that level (RNA/protein) and not try to incorporate any knowledge regarding the change at DNA-level (see Questions below)
- all variants should be described in relation to an accepted **reference sequence** (*see Reference Sequences* (/bg-material/refseq)).
    - the reference sequence file used should be **public and clearly described**, e.g. NC_000023.10, LRG_199, NG_012232.1, NM_004006.2, LRG_199t1, NR_002196.1, NP_003997.1, etc. (*see Reference Sequences* (/bg-material/refseq))
    - the reference sequence used must contain the residue(s) described to be changed.
    - the recommended reference is a genomic reference sequence based on a recent genome build
        - for human the recommended reference is based on genome build GRCh38/hg38, e.g. NC_000023.11 for the chromosome X
        - **when variants are reported in relation to a transcript, the preferred reference sequence is the reference suggested by the MANE project (see Ensembl (http://tark.ensembl.org/web/mane_project/) or NCBI (https://www.ncbi.nlm.nih.gov/refseq/MANE/))**
          NOTE: while Locus Reference Genomic (LRG) (http://www.lrg-sequence.org) reference sequences are still acceptable, new LRG's are no longer generated and RefSeq or Ensembl transcripts specified by the MANE project are preferred for all genes where available to help standardize clinical reporting.

    - the reference sequence used must contain the residue(s) described to be changed.

    - a **letter prefix** is mandatory to indicate the type of reference sequence used. Accepted prefixes are;
        - "**c.**" for a coding DNA reference sequence
        - "**g.**" for a linear genomic reference sequence
        - "**m.**" for a mitochondrial DNA reference sequence
        - "**n.**" for a non-coding DNA reference sequence
        - "**o.**" for a circular genomic reference sequence
        - "**p.**" for a protein reference sequence
        - "**r.**" for an RNA reference sequence (transcript)
    - numbering of the residues (nucleotide or amino acid) in relation to the reference sequence used should **follow the approved scheme** (*see Numbering* (/bg-material/numbering))
- two variants separated by one or more nucleotides should be described individually and **not** as a "delins"
    - **exception: two variants separated by one nucleotide, together affecting one amino acid, should be described as a "delins"**
      NOTE: the SVD-WG is preparing a proposal to modify this recommendation. To apply the current rule one needs to know whether the two variants are in a coding sequence and affecting one amino acid. Recommendations should be general. The new recommendation will be: **two variants separated by less then two nucleotides should be described as a "delins"**

- **3'rule**: for all descriptions the most 3' position possible of the reference sequence is arbitrarily assigned to have been changed
    - the 3'rule also applies for changes in single residue stretches and tandem repeats (nucleotide or amino acid)
    - the 3'rule applies to ALL descriptions (genome, gene, transcript and protein) of a given variant
    - **exception**: deletion/duplication around exon/exon junctions using **c.**, **r.** or **n.** reference sequences (*see Numbering* (/bg-material/numbering/#DNAc))
- descriptions at DNA, RNA and protein level are clearly different:
    - **DNA-level** 123456A>T (*see Details* (/recommendations/DNA)): number(s) referring to the nucleotide(s) affected, nucleotides in CAPITALS using *IUPAC-IUBMB assigned nucleotide symbols* (http://www.qmul.ac.uk/sbcs/iubmb/misc/naseq.html#500)
    - **RNA-level** 76a>u (*see Details* (/recommendations/RNA)): number(s) referring to the nucleotide(s) affected, nucleotides in lower case using IUPAC-IUBMB assigned nucleotide symbols (http://www.qmul.ac.uk/sbcs/iubmb/misc/naseq.html#500)
    - **protein level** Lys76Asn (*see Details* (/recommendations/protein)): the amino acid(s) affected in three- or one-letter code followed by a number IUPAC-IUBMB assigned amino acid symbols (http://www.iupac.org/publications/pac/1984/pdf/5605x0595.pdf)
        - **three-letter** amino acid code is preferred (*see Standards* (/bg-material/standards/#aacode))
        - the **"*"** can be used to indicate the translation stop codon in both one- and three-letter amino acid code descriptions
- **prioritisation**: when a description is possible according to several types, the preferred description is: (1) substitution, (2) deletion, (3) inversion, (4) duplication, (5) insertion
    - when a variant can be described as a duplication or an insertion, prioritisation determines it should be described as a duplication
    - descriptions removing part of a reference sequence replacing it with part of the same sequence are not allowed (e.g. NM_004006.2:c.[762_768del;767_774dup])
- **only approved HGNC gene symbols (http://www.genenames.org) should be used to describe genes**

NOTE: to avoid confusion, HGVS recommends to follow the HGNC guidelines (https://www.genenames.org/about/guidelines/) to use ***italics*** to denote genes and to describe products of gene translocations or fusions (format GENESYMBOL1::GENESYMBOL2) and readthrough transcripts (format GENESYMBOL1-GENESYMBOL2)
NOTE: for protein nomenclature see the *International Protein Nomenclature Guidelines* (https://www.ncbi.nlm.nih.gov/genome/doc/internatprot_nomenguide/), written with the involvement of the HGNC

Characters used

In HGVS nomenclature some **characters** have a **specific meaning**

- "**+**" (plus) is used in *nucleotide numbering* (/bg-material/numbering); c.123+45A>G
- "**-**" (minus) is used in *nucleotide numbering* (/bg-material/numbering); c.124-56C>T
- "*****" (asterisk) is used in *nucleotide numbering* (/bg-material/numbering) and to indicate a translation termination (stop) codon (*see Standards* (/bg-material/standards#RNAcode)); c.*32G>A and p.Trp41*
- "**_**" (underscore) is used to indicate a range; g.12345_12678del
- "**[ ]**" (square brackets) are used for alleles (see *DNA* (/recommendations/DNA/variant/alleles), *RNA* (/recommendations/RNA/variant/alleles), *protein* (/recommendations/protein/variant/alleles)), which includes multiple inserted sequences at one position and insertions from a second reference sequence
  - "**;**" (semi colon) is used to separate variants and alleles; g.[123456A>G;345678G>C] or g.[123456A>G];[345678G>C]
  - "**,**" (comma) is used to separate different transcripts/proteins derived from one allele; r.[123a>u,122_154del]
  - NC_000002.11:g.48031621_48031622ins[TAT;48026961_48027223;GGC]
  - NC_000002.11:g.47643464_47643465ins[NC_000022.10:35788169_35788352]
- "**:**" (colon) is used to separate the reference sequence file identifier (*accession.version_number*) from the actual description of a variant; NC_000011.9:g.12345611G>A
- "**::**" (double colon) is used to describe RNA fusion transcripts (*RNA Deletion-insertion* (/recommendations/RNA/variant/delins/)) and to designate break point junctions creating a ring chromosome (*DNA Complex (HGVS/ISCN)* (/recommendations/DNA/variant/complex/))
- "**( )**" (parentheses) are used to indicate uncertainties and predicted consequences; NC_000023.9:g.(123456_234567)_(345678_456789)del, p.(Ser123Arg)
  NOTE: the range of the uncertainty should be described as precisely as possible (*see below*)
- "**?**" (question mark) is used to indicate unknown positions (nucleotide or amino acid); g.(?_234567)_(345678_?)del
- "**^**" (caret) is used as "or"; c.(370A>C^372C>R) as back translation of p.Ser124Arg (i.e. changing the AGC codon to CGC, AGG or AGA)
- "**>**" (greater than) is used to describe substitution variants (DNA and RNA level); g.12345A>T, r.123a>u (see *DNA* (/recommendations/DNA/variant/substitution), *RNA* (/recommendations/RNA/variant/substitution))
- "**=**" (equals) is used to indicate a sequence was tested but found unchanged; p.(Arg234=)
- "**/**" (forward slash) is used to indicate mosaicism (*see Example DNA substitution* (/recommendations/DNA/variant/substitution/))
- "**//**" (double forward slash) is used to indicate chimerism (*see Example DNA substitution* (/recommendations/DNA/variant/substitution/))
- "**|**" (pipe) is used to indicate that not a direct change of the sequence is described but a modification (a change of state, e.g. methylation). (*see Example methylation* (/recommendations/DNA/variant/other/))

Abbreviations in variant descriptions

Specific abbreviations are used to describe different variant types.

- "**>**" (greater then) indicates a **substitution** (DNA and RNA level); g.123456G>A, r.123c>u (see *DNA* (/recommendations/DNA/variant/substitution), *RNA* (/recommendations/RNA/variant/substitution))
    - a substitution at the protein level is described as p.Ser321Arg (see *protein* (/recommendations/protein/variant/substitution))
- "**del**" indicates a **deletion**; c.76delA (see *DNA* (/recommendations/DNA/variant/deletion), *RNA* (/recommendations/RNA/variant/deletion), *protein* (/recommendations/protein/variant/deletion))
- "**dup**" indicates a **duplication**; c.76dupA (see *DNA* (/recommendations/DNA/variant/duplication), *RNA* (/recommendations/RNA/variant/duplication), *protein* (/recommendations/protein/variant/duplication))
- "**ins**" indicates an **insertion**; c.76_77insG (see *DNA* (/recommendations/DNA/variant/insertion), *RNA* (/recommendations/RNA/variant/insertion), *protein* (/recommendations/protein/variant/insertion))
    - duplicating insertions are described as duplications, not as insertions
- "**inv**" indicates an **inversion**; c.76_83inv (see *DNA* (/recommendations/DNA/variant/inversion), *RNA* (/recommendations/RNA/variant/inversion)). Not used at protein level, usually described as *"delins"* (/recommendations/protein/variant/delins/)
- "**fs**" indicates a **frame shift**; p.Arg456GlyfsTer17 (or p.Arg456Glyfs*17, *see Frame shifts* (/recommendations/protein/variant/frameshift))
- "**ext**" indicates an **extension**; p.Met1**ext**-5 (*see Extension* (/recommendations/protein/variant/extension))

- HGVS/ISCN (*see Community Consultation* (http://www.hgvs.org/mutnomen/comments004.html))
    - "**cen**" indicates the **centromere** of a chromosome
    - "**chr**" indicates a **chromosome**; chr11:g.12345611G>A (NC_000011.9)
    - "**pter** indicates the **first nucleotide** of a chromosome
    - "**qter**" indicates the **last nucleotide** of a chromosome
    - "**sup**" indicates an **supernumary** chromosome (marker chromosome)
- changes of state (modifications)
    - "**gom**" indicates a **gain of methylation**; g.12345678_12345901**|gom**
    - "**lom**" indicates a **loss of methylation**; g.12345678_12345901**|lom**
    - "**met**" indicates a **methylation**; g.12345678_12345901**|met=**

Questions

- **Some papers and web sites use a "-" (minus) to indicate a range, is this correct?**
  The sign used to indicate a range is "_" (underscore) and not a "-" (minus). The minus sign should only be used as a minus in the description of variants based on a coding DNA reference sequence. c.12-14del describes a deletion of nucleotide -14 in the intron directly preceding coding DNA nucleotide 12, **not** a deletion of nucleotides c.12 to c.14.

- **Why is it recommended to use three-letter amino acid code to describe protein variants?**
  Several amino acids start with the same initial letter (**A**la, **A**rg, **A**sn, **A**sp start with **A**, **G**ln, **G**lu, **G**ly with **G**, **L**eu, **L**ys with **L**, **P**he, **P**ro with **P** and **T**hr, **T**yr with **T**) but in one-letter amino acid code this letter is used

as abbreviation for only one. In practice this leads to many mistakes. It is therefore recommended to use three-letter amino acid code abbreviations.

- **When I want to report a variant on DNA, RNA and protein level do I need to use a specific separator?**
  No, best is to report the variant using the format NC_000023.11:g.32849790T>A NM_004006.3:c.124A>T r.(124a>u) p.(Ser42Cys)
  NOTE: when several NP_'s are annotated in the NM_ reference sequence, it is mandatory to add the NP_ reference sequence used to describe the variant at protein level.

- **I found a substitution variant (DNA) which alters splicing, inserting a short sequence in the transcript (RNA), giving a frame shift at #rotein level. How should I list this variant, as substitution, as splice variant, as insertion or as frame shift?**
  When listing variant types, HGVS recommends listing them separately for each level, i.e. DNA, RNA and protein. On DNA level you identified a substitution, on RNA level an insertion and on protein level a frameshift.

- **What do you mean with "variants should be described on the protein level and not incorporate knowledge regarding the change at the DNA-level"?**
  It means that protein variant descriptions should be derived from comparing the variant protein sequence with the reference protein sequence. Knowledge on the underlying change at the DNA level should not be used. E.g. when MetTrpSerSerSerHisAsp.. changes to MetTrpSerSer_HisAsp.. this is described as p.Ser5del. The information that at the DNA level the change is ..ATGTGGTCCAGTTCCCACGAT.. to ..ATGTGGTCC_TCCCACGAT.., so the codon for Ser4 is deleted, is not used; the description p.Ser4del is not correct.

- **Is it correct that when I apply the 3'rule for genes that are on the minus strand of a chromosome, the "g." and "c." variant descriptions differ regarding the nucleotide that I describe as deleted?**
  Yes, when a gene is on the minus strand of a chromosome (opposite transcriptional orientation) and the change is located in a repeated sequence (mono-, di-, tri-, etc. nucleotide stretches) the 3'rule has this as a consequence. When the chromosome sequence is -TGGGGCAT- and one of the G's is deleted (change to -TGGG_CAT-) the description based on chromosome coordinates is g.5delG. When the annotated coding DNA reference sequence is on the minus strand (ATGCCCCA) the description is c.7delC. Not only is the deleted nucleotide different (delG vs. delC), in fact the descriptions also point to another nucleotide, g.5 vs. g.2 (equivalent to c.7delC).

- **Can I describe a deletion when I have not yet sequenced the break point?**
  Yes, using the characters to indicate uncertainties, i.e. the question mark ("**?**") and brackets ("**( )**"), such cases can be described. Describe the range of uncertainty as precise as possible. For details *see Uncertain* (/recommendations/uncertain/).

Website created by William Hong
Nomenclature by J.T. Den Dunnen