Checklist for the description of sequence variants

Last modified April 28, 2014

Since references to WWW-sites are not yet acknowledged as citations, please mention den Dunnen JT and Antonarakis SE (2000). Hum.Mutat. 15:7-12 when referring to these pages.

Purpose

Going through publications one can easily see where people tend to offend the ***"Current recommendations for the description of sequence variants"***. The checklist below covers the most problematic issues and should assist those preparing a publication to describe sequence variants following the current recommendations.

Checklist

1. **Reference Sequence** - do you clearly mention the reference sequence used for numbering (nucleotides/amino acids)?
   A publication should mention, preferably in the Materials & Methods section and/or Table legend, which sequence file was used as reference sequence for numbering of the residues (DNA, RNA and protein) and describing the variants; see *Recommendations*, *Discussion* and *mtDNA variants*.
   - do you mention a *GenBank (not GeneBank) RefSeq-file accession number* ***with version number***?; do not forget the underscore in the accession number (correct is NM_004006**.2**, not NM004006.2).
   - a genomic reference sequence starts with nucleotide 1; a genomic reference sequence can not have negative numbers
   - for a coding DNA reference sequence, do you clearly state that *nucleotide numbering uses the A of the ATG translation initiation start site as nucleotide 1*?
   - when you ***do not*** use a coding DNA reference sequence and ***do not*** start nucleotide numbering ***with 1 at the A of the ATG***

translation initiation start you should not use descriptions preceded by "***c.***".
- if legacy numbering is used, this can only be done ***in addition to*** the approved nomenclature.
- does your reference sequence contain introns?; not that ***NM_*** reference sequences cover mature transcripts and do not contain intronic sequences.

2. **Intronic variants** - do you indicate where the *reference intron sequence* can be found ?
   The recommendation is to describe ***intronic variants*** in the format "***c.89-2A>G***" and not like "*IVS4-2A>G*" (see *Discussion*). When the format "*IVS4-2A>G*" is used, it is essential to give a clear reference for ***intron / exon numbering*** and to mention the *reference sequence used for the intron*.
3. **Tabular overview** - do you provide a clear, unequivocal overview of all changes reported?
   Preferably, a publication contains a *tabular overview* of all variants reported. This overview contains columns describing the change at *DNA-level (absolutely essential)* and, optional, at *RNA* and *protein level*. When data on RNA and/or protein level are provided, it should be made clear whether the data were ***deduced or experimentally verified*** (i.e. state explicitly when RNA was analysed, e.g. to study the consequences of a variant affecting splicing).
4. **Insertions**
   - are insertions reported in the format c.51_52insT?
     Since it is not clear whether one means ***insertion at*** or ***insertion after*** position 52, insertions should not be reported as c.52insT but in the format c.51_52insT (see *Discussion*).
   - do you give the inserted sequence?
     Describing a variant like c.5439_430ins6 is ***not sufficient***, the inserted sequence (ins6, e.g. TGCCAT) ***should always be mentioned***.
   - are the insertions reported really insertions or are they in fact duplications?
     Duplicating insertions should be described as duplications, not as insertions; for the change CCAGTAAC to CCAGT***GT***AAC the description is c.4_5dup (or c.4_5dupGT) is correct, c.5_6insGT is not correct (see *Discussion*).
5. **Most 3' position** - do you correctly assign the change to the most 3' (or C-terminal for protein variants) position possible?
   For deletions, duplications and insertions the ***most 3' position possible is arbitrarily assigned*** to have been changed (see *Recommendations*); important especially in single residue (nucleotide or amino acid) stretches or tandem repeats. Example CCAGT***GT***AAC to CCAGTAAC is described as c.6_7del (or c.6_7delGT, not as c.3_4del or c.4_5del.
6. **Recessive diseases** - do you clearly describe which changes are found in which combination?
   a publication describing sequence changes found in patients suffering from a recessive disease should for each patient explicitly mention which ***combination of changes*** was identified (see *Recommendations*). Example c.[76C>T];[87G>A] or c.[76C>T];[?].
   NOTE: this description differs from that describing ***several changes in one allele,*** which has the format c.[76A>C;113G>C].
7. **Range** - the sign used to indicate a range is "_" (underscore) and not a "-" (minus)?
   To prevent confusion, the *underscore* should be used to indicate a *range* and not the minus sign. The *minus sign* should only be used to indicate *negative numbers*. The correct description to indicate a deletion of the coding DNA nucleotides 12 to 14 is c.12_14del. Not correct is c.12-14del, this describes a deletion of nucleotide -14 in the intron directly preceding cDNA nucleotide 12 (see *Discussion*).
8. **Deletion** - do you indicate the first and last residue involved in a deletion?
   A deletion of more than one residue should mention the first and last residue deleted, separated using a "_" (underscore), e.g. c.21_24del or p.Ala13_Gln16del. Descriptions like c.21del3 should not be used.

9. **Describe at DNA-level** - do you describe all changes reported at DNA-level?
   All changes reported must be described at DNA-level
   - when descriptions at protein level are given in the text, upon first appearance, use a format like "***c.78G>C (p.(Trp26Cys), RNA not analysed)***" or "***c.78G>C (p.Trp26Cys, RNA analysed)***"
   - description of "*silent variants*" in the format "p.(Leu54Leu) (or ***p.(L54L))*** " ***should not be used*** (*see Discussion*). Descriptions should be given at DNA level. Descriptions like p.(Leu54Leu) are non-informative and not unequivocal (there are five possibilities at DNA level); a correct description is ***c.162C>G***
10. **RNA protein level descriptions**
    Recommendations exist to describe alternative transcripts deriving from one allele (see *Recommendations*). Since these descriptions may be rather complex to explain, it is wise to include a link to the HGVS recommendations in the publication.
11. **Protein level descriptions**
    - ***protein reference sequence*** - the protein reference sequence should represent the primary translation product, not a processed mature protein, and thus include any signal peptide sequences (see *Recommendations*).
    - NEW ***Ter / **** - do you use Ter or * to indicate a translation stop codon; the X is not allowed anymore (see *Important changes*)
    - ***one/three letter amino acid code*** - are the correct amino acid codes used at protein level?; several amino acids start with the same initial letter (Ala, Arg, Asn, Asp start with ***A***, Gln, Glu, Gly with ***G***, Leu, Lys with ***L***, Phe, Pro with ***P*** and Thr, Tyr with ***T***) but that initial letter is used as one-letter-amino-acid-code for only one of these (see *Discussion* and *Codons and amino acids*)
    - ***initiating methionine (Met1)*** - *p.Met1?* denotes that amino acid Methionine-1 (translation initiation site) is changed and that it is unclear what the consequence of this change is. When experimental data show that no protein is made, the description *p.0* should be used. The description p.Met1Val is not allowed (see *Discussion*)
    - ***no-stop change*** - recommendations have recently been made to describe substitutions in the stop codon, so called ***no-stop changes*** like p.*110Tyrext*16 (see *Recommendations*)
12. **Polymorphisms**
    Do not describe polymorphic variants as c.127A/G (or p.43I/V). A description of a variant should be ***neutral*** and polymorphisms and pathogenic changes should not be described differently (see *Discussion*). Correct descriptions are c.127A>G and p.(Ile43Val).

| Top of page | MutNomen homepage |
| **Recommendations:** DNA, RNA, protein, uncertain |
| Discussions | FAQ's | Symbols, codons, etc. | History |
| **Example descriptions:** QuickRef / symbols, DNA, RNA, protein |