Recommendations for the description of sequence variants

Last modified September 14, 2015

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.

Contents

Introduction

Discussions regarding the uniform and unequivocal description of sequence variants in DNA and protein sequences (mutations, polymorphisms) were initiated by two papers published in 1993; Beaudet AL & Tsui LC (DOI paper / abstract) and Beutler E (paper /

abstract). The original suggestions presented were widely discussed, modified, extended and ultimately resulted in nomenclature recommendations that have been largely accepted and are applied world-wide (*see History*).

Current rules (den Dunnen, JT and Antonarakis, SE (2000), paper / abstract) however do not extensively cover all types of variants and the more complex changes. These pages will list, based on the last publication, the existing nomenclature recommendations as well as the most recent suggestions (*in italics and marked* NEW). More details regarding the latest additions can be found at the *Discussion page*. These pages can be used as a guide to describe any sequence variant identified and should help to get a uniformly accepted standard.

Discussions regarding the advantages and disadvantages of the recommendations made are necessary in order to continuously improve the system. What is listed on these pages represents the current consensus of the discussions. We invite investigators to communicate to us regarding the recommendations as well as to send us complicated cases not yet covered, with a suggestion of how to describe these (E-mail to: VarNomen @ HGVS.org).

Mutation and polymorphism

In some disciplines the term "**mutation**" is used to indicate "*a change*" while in other disciplines it is used to indicate "*a disease-causing change*". Similarly, the term "**polymorphism**" is used both to indicate "*a non disease-causing change*" or "*a change found at a frequency of 1% or higher in the population*". To prevent this confusion we do not use the terms mutation and polymorphism (including SNP or Single Nucleotide Polymorphism) but use neutral terms like "**sequence variant**", "**alteration**" and "**allelic variant**". The Vol.19(1) issue of Human Mutation (2002) contains several contributions discussing these issues as well as the fact that the term "**mutation**" has developed a negative connotation (see Cotton RGH - p.2, Condit CM et al. - p.69 and Marshall JH - p.76). The recently published "*Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology*" (Richards 2012, Genet.Med. 17:405-424) also suggests the use of the neutral term "**variant**".

NEW Pathogenic

Another confusing term used frequently is "a **pathogenic** variant". While a non-expert concludes the variant described "*causes disease*", the expert probably means "*causes disease when in a specific context*"

- causes disease when found in a male (X-linked recessive disorder)
- causes disease when combined with a similar chance in the other allele (autosomal recessive)
- causes disease when inherited from the father (imprinted)

To prevent confusion it therefore seems best not to use the term "*pathogenic*". A good alternative seems a neutral term like "**affects function**". In fact this properly describes what one actually means, the variant affects the normal function of the gene/protein (in whatever way). This also solves the issue of what term to use for non-disease phenotypes like skin/hair/eye colour or blood group. In such cases it is problematic to choose the phenotype to call "*normal*" or "*pathogenic*". Using "**affects function**" is clear and effective. To classify variants

people use most frequently 5 categories. Based on affects function these could be; *affects funtcion, probably affects function, unknown, probably does not affect function (or probably no functional effect), does not affect function (no functional effect)*. Variants for which a functional effect is unknown can together be called "**variants of unknown significance**" (**VUS**).

General recommendations

(suggestions extending the published recommendations in italics)

The most important rule is that all variants should be described at the most basic level, i.e. the DNA level. Descriptions should always be in relation to a **reference sequence**, either a **genomic** or a **coding DNA** reference sequence. Discussions on which type of reference sequence to prefer, genomic or coding DNA, have been lively. Although theoretically a genomic reference sequence seems best, in practice a coding DNA reference sequence is preferred (see *Reference Sequence discussions*).

- ***describing genes / proteins, only official HGNC gene symbols should be used***
  all changes reported must be **described at DNA-level**. When descriptions at RNA or protein level are given in the text (including TITLE and ABSTRACT), on first mention, a format like "c.78G>C (p.Trp26Cys)" should be used.
- when several changes are described in one manuscript, a tabular listings should be provided summarizing the findings, using separate columns for DNA, RNA and protein and clearly indicating whether the changes were **experimentally determined** or only **theoretically deduced**. When changes in patients with a recessive disease are described, the listing should make clear in which combination the changes were found. An additional column can be used to mention additional findings and to make remarks.
- to avoid confusion in the description of a variant it should be preceded by a letter indicating the type of reference sequence used. Several different reference sequences can be used (*see Figure*);
  - "c." for a **coding DNA** sequence (like c.76A>T)
  - "g." for a **genomic** sequence (like g.476A>T)
  - "m." for a **mitochondrial** sequence (like m.8993T>C, *see Reference Sequence*)
  - NEW *"n" for a **non-coding RNA** reference sequence (gene producing an RNA transcript but not a protein)*
  - "r." for an **RNA** sequence (like r.76a>u)
  - "p." for a **protein** sequence (like p.Lys76Asn)
- NEW *the DNA reference sequence used should preferably be* a LRG (Locus Reference Genomic sequence, *see Reference Sequence*). *from the RefSeq database, listing both database accession and version number (like NM_004006.2)*
  - within one document only one DNA reference sequence should be used. When variants in more than one sequence (gene) are described, any confusion should be prevented by including a unique indicator in the description. Indicator and sequence description should be separated by a colon (":") like in NM_004006.1:c.3G>T or GJB2:c.76A>C.
    NOTE: this format is especially important for unequivocal descriptions of SNP's (*see Discussion*). NEW *When both HGNC-approved gene symbol and database accession.version number are indicated this should be done using the format NM_004006.1(DMD):c.3G>T (see Discussion*).

- NEW *the coding DNA reference sequence used should represent the major and largest transcript of the gene. Alternatively spliced exons (5'-first, internal or 3'-terminal) derived from **within the gene** can be numbered as for intronic sequences. Variants in transcripts initiating or terminating **outside this region** can be described as upstream / downstream sequences (see Reference Sequence discussions).*
- protein reference sequences should represent the primary translation product, not a processed mature protein (*see FAQ*).
- NEW *when changes start or end in another sequence (gene), e.g. for large deletions, the nucleotide numbering for that end is based on the nucleotide numbering of that sequence (like c.827_NM_004004.3:c.235del). When the endpoint occurs on the opposite, non-transcribed strand (anti-sense strand), an "o" precedes the reference identifier (like c.827_oNM_004004.3:c.233del, see Discussion).*
- NEW *when a variant affects more then one gene, to prevent confusion, the variant should be described in relation to all genes affected.*

- for a clear distinction, descriptions at DNA, RNA and protein level are unique;
  - DNA-level
    in capitals, starting with a number referring to the first nucleotide affected (like c.76A>T or g.476A>T)
  - RNA-level
    in lower-case, starting with a number referring to the first nucleotide affected (like r.76a>u)
  - protein level
    in capitals, starting with a letter referring to first the amino acid affected (like p.Lys76Asn)
- **nucleotide numbering** *(for details and examples see Reference Sequence discussions)*
  - coding DNA Reference Sequence (*see Figure and Numbering*)
    - there is no nucleotide 0
    - nucleotide 1 is the A of the ATG-translation initiation codon
    - the nucleotide **5' of the ATG-translation initiation** codon is -1, the previous -2, etc.
      NEW ***NOTE:** den Dunnen & Antonarakis (Hum.Mut. 15: 7-12) write "For **genomic DNA** and cDNA sequences, the A of the ATG of the initiator Methionine codon is denoted nucleotide +1". This is an error, correct is; "In coding DNA reference sequences, the A of the ATG of the initiator Methionine codon is denoted nucleotide +1".*
    - NEW *the nucleotide **3' of the translation stop codon** is *1, the next *2, etc.*
    - intronic nucleotides
      - beginning of the intron; the number of the last nucleotide of the preceding exon, a plus sign and the position in the intron, like c.77+1G, c.77+2T, etc.
      - end of the intron; the number of the first nucleotide of the following exon, a minus sign and the position upstream in the intron, like c.78-1G.
      - in the middle of the intron, numbering changes from "c.77+.." to "c.78-.."; for introns with an uneven number of nucleotides the central nucleotide is the last described with a "+" *(see Reference Sequence discussions)*
  - genomic Reference Sequence (*see Figure*)
    - nucleotide numbering is purely arbitrary and starts with 1 at the first nucleotide of the database reference file
      NEW ***NOTE:** in den Dunnen&Antonarakis (Hum.Mut. 15: 7-12) write "For **genomic DNA** and cDNA sequences, the A of the ATG of the initiator Methionine codon is denoted nucleotide +1". This is an error, correct is; "In genomic reference*

sequences, the first nucleotide is nucleotide 1".
    - no +, - or other signs are used
    - the sequence should include all nucleotides covering the sequence (gene) of interest and should start well 5' of the promoter of a gene
    - when the complete genomic sequence is not known, a coding DNA reference sequence should be used
- **specific changes**
    - ">" indicates a ***substitution*** at DNA level (like c.76A>T)
    - "_" (underscore) indicates a ***range*** of affected residues, separating the first and last residue affected (like c.76_78delACT, *see Discussion*)
    - "del" indicates a ***deletion*** (like c.76delA)
    - "dup" indicates a ***duplication*** (like c.76dupA); NEW *duplicating insertions are described as* ***duplications****, not as insertions; ACTTTGTGCC to ACTTTGTG**G**CC is described as c.8dupG (not as c.8_9insG, see Discussion)*
    - "ins" indicates a ***insertion*** (like c.76_77insG)
    - NEW "inv" indicates an ***inversion*** (like c.76_83inv)
    - NEW *"con" indicates a* ***conversion*** *(like c.123_678conNM_004006.1:c.123_678, see Recommendations)*
    - "[]" indicates an allele (like c.[76A>T], *see Recommendations*)
    - NEW *"()" is used when the exact position of a change is not known, the range of the uncertainty is described as precisely as possible and listed between brackets (like c.(67_70)insG, see Uncertainties)*
- **miscellaneous**
    - for all descriptions the **most 3' position** possible is arbitrarily assigned to have been changed, this is important especially in single residue (nucleotide or amino acid) stretches or tandem repeats (*see Recommendations*, *see Discussion*)
    - *variability in the number of repeated sequences (e.g. ATGCGA**TGTGTG**CC) are described as c.123+74TG(3_6) (see Recommendations)*
    - NEW *triplications, quadruplications, etc. are described as alleles of variable short sequence repeats; c.87_93[3] describes a triplication of the 7 nucleotides from coding DNA position 87 to 93 (not as c.87_93tri, see Discussion)*
    - ***two sequence variants in one individual***
        - two sequence changes in different alleles (e.g. for recessive diseases) are listed between square brackets, separated by a ";"-character; c.[76A>C];[87delG] (*see Discussion*)
        - two sequence variants in one allele are listed between square brackets, separated by a ";"-character; c.[76A>C; 83G>C] (*see Discussion*)
        - NEW *two sequence changes with alleles unknown are listed between square brackets, separated by "(;)"; c.[76A>C(;)83G>C] (see FAQ)*
        - NEW *descriptions of sequence changes in different genes (e.g. for recessive diseases) are listed between square brackets, separated by a ";"-character and include a reference to the sequence (gene) changed; DMD:c.[76A>C];GJB:c.[87delG] (see Discussion)*
    - NEW ***mosaic cases:*** *two different nucleotides at one position in one allele are listed between square brackets, separated by a "/"-character; c.[=/83G>C] (see Recent changes)*
    - NEW ***chimeric cases:*** *two different nucleotides at one position in one allele are listed between square brackets, separated by a*

"//"-*character; c.[=//83G>C] (see Recent changes)*

- a unique identifier should be assigned to each variant; when available, the OMIM-identifier can be used, otherwise database curators should assign a unique identifier.

Detailed recommendations

- DNA level
- RNA level
- protein level
