Nomenclature for the description of sequence variations


Reproduced with kind permission from Prof. S. E. Antonarakis

(last modified March 7, 2001)

Questions and comments regarding nomenclature should be directed to Professor Stylianos Antonarakis (stylianos.antonarakis@medecine.unige.ch) or Dr. Johan T. den Dunnen (ddunnen@lumc.nl). This page can also be found at the HGVS site.

Contents

- **Introduction**
- **Recommendations**
  - General
  - DNA-level
  - RNA-level
  - protein-level
- **Codons and encoded amino acids**
  - genetic code
  - amino acid descriptions (one / three letter code)

Introduction

Recently, a nomenclature system has been suggested for the description of changes (mutations and polymorphisms) in DNA and protein sequences [Antonarakis, S.E. and the Nomenclature Working Group (1998) Recommendations for a nomenclature system for human gene mutations. Hum.Mut. 11: 1-3]. These nomenclature recommendations have now been largely accepted and stimulated the uniform and unequivocal description of sequence changes. However, current rules do not yet cover all types of mutations, nor do they cover more complex mutations. This document lists the existing recommendations and summarizes suggestions for the description of additional, more complex changes, (shown in italics) based on a manuscript published in Human Mutation [den Dunnen, JT and Antonarakis, SE (2000). Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion.
Discussions regarding the advantages and disadvantages of the suggestions are necessary in order to continuously improve the designation of sequence changes. The consensus of the discussions will be posted here and we invite investigators to communicate with us regarding these suggestions. Furthermore, we invite investigators to send us complicated cases not covered yet, with a suggestion of how to describe these (mail to ddunnen@lumc.nl and Stylianos.Antonarakis@medecine.unige.ch). We hope these pages will be used as a guide to describe any sequence change, ultimately evolving into a uniformly accepted reference for mutation nomenclature description.

General recommendations

(suggestions extending the current recommendations are in italics)

The term "sequence variation" is used to prevent confusion with the terms "mutation" and "polymorphism", mutation meaning "change" in some disciplines and "disease-causing change" in others and polymorphism meaning "non disease-causing change" or "change found at a frequency of 1% or higher in the population".

The basic recommendation is to use **systematic names** to describe each sequence variation. For this, variations are described at the most basic level, i.e. the DNA level, using either a genomic or a cDNA reference sequence. A genomic reference sequence is preferred because it overcomes difficult cases, including multiple transcription initiation sites (promoters), alternative splicing, the use of different poly-A addition signals, multiple translation initiation sites (ATG-codons) and the occurrence of length variations. When, like in most cases, the entire genomic sequence is not known, a cDNA reference sequence should be used instead.

- sequence variations are described in relation to a reference sequence for which the accession number from a primary sequence database (Genbank, EMBL, DDJB, SWISS-PROT) should be mentioned in the publication/database submission (e.g. M18533)
- **tabular listings of the sequence variations described should contain columns for DNA, RNA and protein and clearly indicate whether the changes were experimentally determined or only theoretically deduced**
- to avoid confusion in the description of a sequence change, preceed the description with a letter indicating the type of reference sequence used;
  - "g." for a **genomic** sequence (e.g. g.76A>T)
  - "c." for a **cDNA** sequence (e.g. c.76A>T)
  - "m." for a **mitochondrial** sequence (e.g. m.76A>T) (from David Fung, Camperdown, Australia)
  - "r." for an **RNA** sequence (e.g. r.76a>u)
  - "p." for a **protein** sequence (e.g. p.K76A)
- to discriminate between the different levels (DNA, RNA or protein), descriptions are unique;
  - at DNA-level, in capitals, starting with a number referring to the first nucleotide affected (e.g. c.76A>T)
  - at RNA-level, in lower-case, starting with a number referring to the first nucleotide affected (e.g. r.76a>u)
  - at protein level, in capitals, starting with a letter referring to first the amino acid (one-letter code) affected (e.g. p.T26P)
• a range of affected residues is indicated by a "_"-character (underscore) separating the first and last residue affected (e.g. 76_78delACT)
  
  NOTE: current recommendations use the "-"-character (i.e. 76-78delACT)

• for deletions, duplications or insertions in short tandem repeats, the most 3' nucleotide is arbitrarily assigned as the nucleotide changed

• two sequence variations in one allele are listed between brackets, separated by a "+"-character (e.g. [76A>C + 83G>C])
  
  NOTE: current recommendations use the ";"-character as a separator (i.e. [76A>C; 83G>C])

• sequence changes in different alleles (e.g. for recessive diseases) are listed between brackets, separated by a "+"-character (e.g. [76A>C + [87delG]])
  
  NOTE: the current recommendation is [76A>C + 87delG]

• a unique identifier should be assigned to each mutation. The unique OMIM-identifier can be used, otherwise database curators should assign unique identifiers

---

DNA level

• nucleotides are designated by the bases (in upper case); A (adenine), C (cytosine), G (guanine) and T (thymidine)

• nucleotide numbering:
  • nucleotide +1 is the A of the ATG-translation initiation codon, the nucleotide 5' to +1 is numbered -1; there is no base 0
  • non-coding regions;
   - the nucleotide 5' of the ATG-translation initiation codon is -1
   - the nucleotide 3' of the translation termination codon is *1
  • intronic nucleotides;
   - **beginning of the intron**: the number of the last nucleotide of the preceeding exon, a plus sign and the position in the intron, e.g. 77+1G, 77+2T (when the exon number is known, the notation can also be described as IVS1+1G, IVS1+2T)
   - **end of the intron**: the number of the first nucleotide of the following exon, a minus sign and the position upstream in the intron, e.g. 78-2A, 78-1G (when the exon number is known, the notation can also be described as IVS1-2A, IVS1-2G)
  • for deletions, duplications or insertions in single nucleotide (or amino acid) stretches or tandem repeats, the most 3' copy is arbitrarily assigned to have been changed (e.g. ACTTTGTGCC to ACTTTGCCC is described as 7_8delTG)

Description of nucleotide changes

• **substitutions** are designated by a “>”-character
  • 76A>C denotes that at nucleotide 76 a A is changed to a C
  • 88+1G>T (alternatively IVS2+1G>T) denotes the G to T substitution at nucleotide +1 of intron 2, relative to the cDNA positioned between nucleotides 88 and 89
  • 89-2A>C (alternatively IVS2-2A>C) denotes the A to C substitution at nucleotide -2 of intron 2, relative to the cDNA positioned between nucleotides 88 and 89

  NOTE: **polymorphic variants** are sometimes described as 76A/G, but this is not recommened!

• **deletions** are designated by "del" after the nucleotide(s) flanking the deletion site
  • 76_78del (alternatively 76_78delACT) denotes a ACT deletion from nucleotides 76 to 78
Nomenclature for the description of sequence variations (mutation nomenclature)

82_83del (alternatively 82_83delTG) denotes a TG deletion in the sequence ACTTTGTGCC (A is nucleotide 76) to ACTTTGCC

IVS2_IVS5del (alternatives 88+_923+? or EX3_5del) denotes an exonic deletion starting at an unknown position in intron 2 (after nucleotide 88) and ending at an unknown position in intron 5 (after nucleotide 923)

- **insertions** are designated by "ins" after the nucleotides flanking the insertion site, followed by the nucleotides inserted

  NOTE: as separator the "^"-character is sometimes used but this is not recommended (e.g. 83^84insTG)

  - 76_77insT denotes that a T was inserted between nucleotides 76 and 77
  - 83_84insTG denotes a TG insertion in the TG-tandem repeat sequence of ACTTTGTGCC (A is nucleotide 76) to ACTTTGTG_TGCC. Note that this sequence variation (a duplicating insertion) can also be described as a duplication, i.e. 82_83dupTG (see "duplications")

- **variability of short sequence repeats**, e.g. in ACTTGTTGCC (A is nt 1991), are designated as 1993(TG)3-6 with nucleotide 1993 containing the first TG-dinucleotide which is found repeated 3 to 6 times in the population.

- **insertion/deletions (indels)** are described as a deletion followed by an insertion after the nucleotides affected

  - 112_117delinsTG (alternatively 112_117delAGGTCAinsTG or 112_117>TG) denotes the replacement of nucleotides 112 to 117 (AGGTCA) by TG

- **duplications** are designated by "dup" after the nucleotides flanking the duplication site,

  - 77_79dupCTG denotes that the nucleotides 77 to 79 were duplicated
  - duplicating insertions in short tandem repeats (or single nucleotide stretches) can also be described as a duplication, e.g. a TG insertion in the TG-tandem repeat sequence of ACTTTGTGCC (A is nt 76) to ACTTTGTG_TGCC can be described as 82_83dupTG (now 83_84insTG)

- **inversions** are designated by "inv" after the nucleotides flanking the inversion site

  - 203_506inv (or 203_506inv304) denotes that the 304 nucleotides from position 203 to 506 have been inverted

- **translocations** (no suggestions yet)

- **changes in different alleles** (e.g. in recessive diseases) are described as "[change allele 1] + [change allele 2]"

  - [76A>C] + [76A>C] denotes a homozygous A to C change at nucleotide 76
  - [76A>C] + [?] denotes a A to C change at nucleotide 76 in one allele and an unknown change in the other allele

- **two variations in one allele** are described as "[first change + second change]"

  - [76A>C + 83G>C] denotes an A to C change at nucleotide 76 and a G to C change at nucleotide 83 in the same allele

  NOTE: current recommendations use the ",;"-character as a separator (i.e. [76A>C; 83G>C])

RNA level

Sequence changes at RNA level are basically described as those at the DNA level with the following modifications/additions;

- an “r.” is used to indicate that a change is described at RNA-level
- nucleotides are designated by the bases (in lower case); a (adenine), c (cytosine), g (guanine) and

http://www.hgmd.cf.ac.uk/docs/mut_nom.html
u (uracil)
- 78u>a denotes that at nucleotide 78 a U is changed to an A
- when one change affects RNA-processing, yielding two or more transcripts, these are described between square brackets, separated by a “;”-character
  - [r.76a>c; r.76a>c + r.73_88del] denotes the nucleotide change c.76A>C causing the appearance of two RNA molecules, one carrying this variation only and one containing in addition a deletion of nucleotides 73 to 88 (shift of the splice donor site to within the exon)
  - [r. =; r.88_89ins88+1_88+10 + r.88+2t>c] denotes the intronic mutation g.88+2T>C causing the appearance of two RNA molecules, one normal (r. =) and one containing an insertion of the intronic nucleotides 88+1 to 88+10 with the nucleotide change 88+2t>c
  - [r.88g>a + r.88_89ins88+1_88+10] denotes the nucleotide change c.88G>A causing an insertion of the intronic nucleotides 88+1 to 88+10 (shift of the splice donor site to an intronic position)

Protein level

Sequence changes at protein level are basically described as those at the DNA level with the following modifications/additions;

- the one letter amino acid code is used, with "X" designating a translation termination codon
- Amino acid numbering:
  - the translation initiator Methionine is numbered as +1

Description of amino acid changes

- substitutions:
  - missense changes
    W26C denotes that amino acid 26 (Tryptophan, W) is changed to a Cysteine (C)
  - nonsense changes
    W26X denotes that amino acid 26 (Tryptophan, W) is changed to a stop codon (X)
  - initiating methionine (M1)
    Currently, mutations in the translation initiating Methionine (M1) are mostly described as a substitution, e.g. M1V. This is not correct. Either no protein is produced or the translation initiation site moves up- or downstream. Unless experimental proof is available, it is probably best to report the effect on protein level as “unknown”. When experimental data show that no protein is made, the description "p.0" might be most appropriate

NOTE: polymorphic variants are sometimes described as 36L/I, but this is not recommened!

- deletions are designated by "del" after the nucleotide(s) flanking the deletion site
  - K29del in the sequence CKMGHQQQCC (C is amino acid 28) denotes a deletion of amino acid Lysine 29 (K) to CMGHQQQCC
  - C28_M30del denotes a deletion of three amino acids, from Cysteine 28 to Methionine 30
  - Q35del in the sequence CKMGHQQQCC (C is amino acid 28) denotes a Glutamine 35 (Q) deletion to CKMGHQQQCC
  - if a deletion creates a new amino acid at the deletion junction the change is described as an insertion/deletions, e.g. C28_M30delinsW (see below)
- insertions are designated by "ins" after the nucleotides flanking the insertion site, followed by the
nucleotides inserted

**NOTE:** as separator the "^"-character is sometimes used but this is not recommened (e.g. Q83^C84insQ)

- K29_M29insQSK denotes that the sequence QSK was inserted between amino acids Lysine 29 (K) and Methionine 30 (M), changing CKMGHQQQCC (C is amino acid 28) to CKQSKMGHQQQCC
- Q35_C36insQ in the sequence CKMGHQQQCC (C is amino acid 28) denotes a Glutamine (Q) insertion to CKMGHQQQQCC. Note that this sequence variation (a duplicating insertion) can also be described as a duplication, i.e. Q35dup (see "duplications")
- if an insertion creates a new amino acid at the insertion junction the change is described as an **insertion/deletions**, e.g. C28delinsWV (see below)

- **variability of short sequence repeats**, e.g. in CKMGHQQQCC (C is amino acid 28), are designated as 33(Q)3-6 with amino acid Glutamine 33 (Q, the first repeated amino acid) found repeated 3 to 6 times in the population.

- **insertion/deletions (indels)** are described as a deletion followed by an insertion after the nucleotides affected
  - C28_K29delinsW denotes a 3 bp deletion affecting the codons for Cysteine 28 and Lysine 29, substituting them for a codon for Tryptophan
  - C28delinsWV denotes a 3 bp insertion in the codon for Cysteine28, generating codons for Tryptophan (W) and Valine (V)

- **duplications** are designated by "dup" after the amino acids flanking the duplication site
  - G31_Q33dup in the sequence CKMGHQQQCC (C is amino acid 28) denotes a duplication of amino acids Glycine 31 (G) to Glutamine 33 (Q) CKMGHQQQGCCKMGHQQQCC
  - duplicating insertions in short tandem repeats (or single amino acid stretches) can also be described as a duplication, e.g. a HQ insertion in the HQ-tandem repeat sequence of CKMGHQQQCC (C is amino acid 28) to CKMGHQQQHQCC can be described as H34_Q35dup (now Q35_C36insHQ)

- **frame shifting mutations**: recommendations to describe these sequence changes have not yet been made. Although it is probably not useful to add much detail in this description, it might be sensible, e.g. in the case of C-terminal mutations, to include the length of the new, shifted reading frame.
  - R97fsX121 (alternative R97fs) denotes a frame shifting change with Arginine97 as the first affected amino acid and the new reading frame being open for 23 amino acids

---

| Top of page |