Recommendations for the description of protein sequence variants (v2.0)

Last modified August 31, 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

- **Codons and encoded amino acids**
  - genetic code
  - amino acid descriptions *(1- and 3-letter code)*

Protein level

(suggestions extending the published recommendations NEW *in italics)*

***NOTE:** definitions of protein changes have been extensively reviewed (2013-Q2). This did not affect HGVS recommendations for variant descriptions but it did change under which category specific types are listed below. For example, where a nonsense change (p.Trp26Ter or p.W26*) was originally listed under Substitutions it is now listed under Deletions.*

The recommendations for the description of protein variants explain how changes in the sequence of a protein should be described. It should be noted that these changes are a consequence of a variant at DNA level that may or may not have influenced the processing of the RNA before it is translated into protein. Experimental evidence of protein level variants, e.g. from mass spectrometry amino acid sequencing, will rarely exist. In some cases indirect evidence might come from protein sizing (Western blot analysis) or localisation (immuno-histochemical staining). In most cases protein descriptions will however be ***deduced only***, predicted from the changes detected on DNA and/or RNA level.

Specific terms are used to describe the consequences of a change at protein level, like ***missense***, ***nonsense***, ***silent*** and ***frame shift***. These terms are not used in the descriptions given below. Missense is under substitution, nonsense under deletion, silent under no change and frame shift under deletion/insertion (indel).

General

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

- descriptions at protein level may only be given ***in addition to*** a description at DNA (and RNA) level
- descriptions at protein level should describe the changes observed on protein level and ***not try to incorporate any knowledge regarding the change at DNA-level*** (*see FAQ*)
- NEW to indicate that the description at protein level is without any experimental evidence it is recommended that, ***when RNA nor protein has been analysed***, the description is given ***between brackets***, like *p.(Arg22Ser)* (*see Discussion* ***2012-10-12***)
- a "***p.***" preceding the change is used to indicate a description at protein level
- amino acids are described as "***Trp26***" or "***W26***", i.e. with capital first letter (not as "*trp26*" or "Trp^{26}")
- the 3-letter amino acid code is preferred to describe the amino acid residues (*see Discussion*)
- for all descriptions the ***most C-terminal position possible*** is arbitrarily assigned to have been changed

- **alleles** are described using square brackets ("***p.[ ]***")
- ***Miscellaneous***
  - *unknown effect*
    - *p.?* - protein has not been analysed, an effect is expected but difficult to predict
    - *p.(=)* - protein has not been analysed, but no change is expected
    - *p.=* - protein has not been analysed, RNA was, but no change is expected (***silent change***)
  - *no protein*
    changes which affect the promoter of a gene, the transcription initiation site (cap site), the translation initiation site, etc. may affect the amount of protein produced;
    - *p.0* - no protein can be detected (*experimental data should be available*)
    - *p.0?* - probably no protein is produced
  - *amount of protein*
    changes which do not affect the protein sequence itself but only the amount of protein produced (other then ***no protein***) are described as p.= (*no change*). Remarks on the amount of protein should be made separately (e.g. under Remarks).
  - *protein modifications*
    currently no recommendations exist for the description of protein modifications. Remarks on protein modifications should be made separately (e.g. under Remarks).

Amino acid coding and numbering

- the Methionine encoded by the translation initiation site (*start codon*) is numbered as residue 1 ("***Met1***" or " ***M1*** ")
- NEW the protein coding sequence ends at a translation termination codon (*stop codon*), described at protein level as "***Ter***" or "*******" ("*******" in 1- and 3-letter amino acid code) (*see Important changes*)
- the protein reference sequence should represent the ***primary translation product***, not a processed mature protein, and thus include e.g. signal peptide sequences (*see FAQ*)
- amino acids originating from changes introducing ***upstream translation initiation*** are numbered like nucleotides; *..., Gln-2, Thr-1*
- amino acids originating from changes resulting in ***translation of intronic sequences*** are numbered like nucleotides; *Val4+1, Ser4+2, ..., Phe5-2, Gln5-1*
- amino acids originating from no-stop changes causing ***translation downstream of the translation termination codon*** are numbered like nucleotides; *Gln*1, Ser*2, ...*

Silent changes

Description of so called "silent" changes in the format p.(Leu54Leu) (or p.(L54L)) should not be used. When desired such changes can be described using ***p.(=)***. Descriptions should always be given at ***DNA level*** (*see Discussion*).

Substitutions

Substitutions (missense changes) ***replace one amino acid by one other amino acid*** and are descrbied using the format ***p.Trp26Cys***. The descrption does not use the "**>**"-character used on DNA- and RNA level (indicating "*changes to*").

- **missense changes**
  p.Trp26Cys denotes that amino acid Tryptophan-26 (Trp, W) is changed to a Cysteine (Cys)
- ***start codon*** *(initiating methionine change - Met1)* (*see Discussion*, *see Examples*)
  a change affecting the translation initiation codon (Met-1) is, depending on its consequence, either
  - a change which results in no protein being produced (*p.0*)
    Met1? - denotes that amino acid Methionine-1 (translation initiation site) is changed and that it is unclear what the consequences of the change are
  - an N-terminal deletion (*p.Phe2_Met46del*, i.e. activating downstream translation initiation)
    NOTE: *up to August 2015 the example given was p.Met1_Lys45del which is not correct, the 3' rule should be applied!*
  - an extension (*p.Met1ValextMet-12*, activating upstream translation initiation)
- ***nonsense change***
  a change introducing an immediate translation stop codon, is described as an amino acid deletion
- ***no-stop change (Ter)*** NEW *(change in stop codon, Ter/*)*
  a change affecting the translation termination codon (Ter, *) is described as an extension (*p.Ter110GlnextTer17* or *p.*110Glnext*17*).

Deletions

Deletions remove one or more amino acid residues from the protein and are described using "***del***" after an indication of the first and last amino acid(s) deleted separated by a "_" (*underscore*). Deletions remove either a small internal segment of the protein (*in-frame deletion*), part of the N-terminus of the protein (*initiation codon change*) or the entire C-terminal part of the protein (*nonsense change*). A ***nonsense*** change is a special type of deletion removing the entire C-terminal part of a protein starting at the site of the variant (specified ***2013-03-16***).

- **in-frame deletions** - are described using "***del***" after an indication of the first and last amino acid(s) deleted separated, by a "_" (*underscore*).
  - p.Gln8del in the sequence MKMGHQQ**Q**CC denotes a Glutamine-8 (Gln, Q) deletion to MKMGHQQCC
  - p.(Cys28_Met30del) denotes RNA nor protein was analysed but the predicted change is a deletion of three amino acids, from Cysteine-28 to Methionine-30
- **initiating methionine change (Met1) causing a N-terminal deletion** (*see Discussion*, *see Examples*)
  NOTE: changes extending the N-terminal protein sequence are described as an extension
  - *p.0* - no protein is produced (experimental data should be available)
    NOTE: this change is not described as *p.Met1_Leu833del*, i.e. as a deletion removing the entire protein coding sequence
  - *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
    - *p.Met1_Lys45del* - a new translation initiation site is activated (at *Met46*)
- **nonsense variant** - are a ***special type of amino acid deletion*** removing the entire C-terminal part of a protein starting at the site of the variant. A nonsense change is described using the format *p.Trp26Ter* (alternatively *p.Trp26**). The description does not include the deletion at protein level from the site of the change to the C-terminal end of the protein (stop codon) like *p.Trp26_Leu833del* (the deletion of amino acid residue Trp26 to the last amino acid of the protein Leu833).
    - p.(Trp26Ter) indicates RNA nor protein was analysed but amino acid Tryptophan26 (Trp, W) is predicted to change to a stop codon (Ter) (*alternatively p.(W26*) or p.(Trp26*)*)

NOTE: for all descriptions the **most C-terminal position** possible is arbitrarily assigned to have been changed

---

Duplications

Duplications are described using "***dup***" after an indication of the first and last amino acid(s) duplicated separated by a "_" (*underscore*).

- p.Gly4_Gln6dup in the sequence MKMGHQQQCC denotes a duplication of amino acids Glycine-4 (Gly, G) to Glutamine-6 (Gln, Q) (i.e. MKMGHQ**GHQ**QQCC)
- duplicating insertions in single amino acid stretches (or short tandem repeats) are described as a duplication, e.g. a duplicating HQ insertion in the HQ-tandem repeat sequence of MKMGHQHQCC to MKMGHQHQ**HQ**CC is described as p.His7_Gln8dup (not p.Gln8_Cys9insHisGln)

NOTE: for all descriptions the **most C-terminal position** possible is arbitrarily assigned to have been changed

---

Insertions

Insertions add one or more amino acid residues between two existing amino acids and this insertion is not a copy of a sequence immediately 5'-flanking (*see Duplication*). Insertions are described using "***ins***" after an indication of the amino acids flanking the insertion site, separated by a "_" (*underscore*) and followed by a description of the amino acid(s) inserted. Since for large insertions the amino acids can be derived from the DNA and/or RNA descriptions they need not to be described exactly but the total number may be given (*like "ins17"*).

- p.Lys2_Met3insGlnSerLys denotes that the sequence GlnSerLys (QSK) was inserted between amino acids Lysine-2 (Lys, K) and Methionine-3 (Met, M), changing MKMGHQQQCC to MK**QSK**MGHQQQCC
- p.Trp182_Gln183ins17 describes a variant that inserts 17 amino acids between amino acids Trp182 and Gln183
  NOTE: it must be possible to deduce the 17 inserted amino acids from the description given at DNA or RNA level

NOTE: duplicating insertions should be described as duplications (*see Discussion*), not as insertion.

Variability of short sequence repeats NEW

Variability of short sequence repeats are described as p.Gln6(3_6); the description indicates that a stretch of Glutamines (Gln, Q) is present, starting at amino acid position 6 (e.g. in MKMGHQQQCC), which is found with a variable length from 3 to 6 in the population

NOTE: the underscore is used to indicate the range (3 to 6 times).

Deletion/insertions (indels)

Deletion/insertions (indels) replace one or more amino acid residues with one or more other amino acid residues. Deletion/insertions are described using "***delins***" as a deletion followed by an insertion after an indication of the amino acid(s) flanking the site of the deletion/insertion separated by a "_" (*underscore*, *see Discussion*). ***Frame shifts*** are a ***special type of amino acid deletion/insertion*** affecting an amino acid ***between*** the first (*initiation, ATG*) and last codon (*termination, stop*), replacing the normal C-terminal sequence with one encoded by ***another reading frame*** (specified ***2013-10-11***). A frame shift is described using "***fs***" after the first amino acid affected by the change. Descriptions either use a ***short*** (*"fs"*) *or* ***long*** (*"fsTer#"*) ***description***. The description of frame shifts does not include the deletion at protein level from the site of the frame shift to the natural end of the protein (stop codon). The inserted amino acid residues are not described, only the total length of the new shifted frame is given (i.e. including the first amino acid changed).
NOTE: *typing error in den "Dunnen & Antonarakis (2000)"*. The suggestion to use "**>**" to indicate "***delins***" in frame shift descriptions has been retracted.
NOTE: when ***one*** nucleotide is replaced by ***one*** other nucleotide the change is called a ***substitution***

- **in-frame**
  - p.(Cys28_Lys29delinsTrp) indicates RNA nor protein was analysed but the predicted change is a 3 bp deletion that affects the codons for Cysteine-28 and Lysine-29, substituting them for a codon for Tryptophan
  - p.Cys28delinsTrpVal denotes a 3 bp insertion in the codon for Cysteine-28, generating codons for Tryptophan (Trp, W) and Valine (Val, V)
- **frame shifts**
  are described using the format *p.Arg97Glyfs*26* (alternatively *p.Arg97GlyfsTer26*, or short *p.Arg97fs*) where *Arg97Gly* describes the change of the first amino acid affected (Arg97 replaced by a Pro residue), "***fs***" indicating the ***frame shift*** and **16* giving the position of the translation termination codon (stop codon) in the new reading frame.
  NOTE: the description does not include a description of the deletion from the site of the change to the C-terminal end of the protein (stop codon) like *p.Arg97_Leu833delinsGlyfsTer26)* nor a specific description of the inserted amino acid residues.
  NOTE: the shifted reading frame includes the first new amino acid (*Gly*) and encounters a translation termination codon at position 26 (*Ter26* or **26*). The shifted reading frame is thus open for 'Ter26-1' amino acids.
  - short description - uses "fs" only, e.g. *p.Arg97fs*
  - long description - uses "fsTer#" (alternatively "fs*#") (*see Discussion*)

- includes the change occurring at the site of the frame shift, e.g. *p.Arg97Gly*
- "fsTer#" (or "fs*#") indicates at which position the new reading frame encounters a translation termination (stop) codon stop (Ter# / *#). The position of the stop in the new reading frame is calculated starting from the first amino acid changed by the frame shift, and ending at the first stop codon (fsTer# or fs*#)

- Examples
  - *p.Arg97ProfsTer23 (alternatively p.Arg97Profs*23; **short** p.Arg97fs)* denotes a frame shifting change with Arginine-97 as the first affected amino acid, replacing it for a Proline and creating a new reading frame ending at a stop at position 23 (counting starts with the Proline as amino acid 1)
  - *p.Glu5Valfs*5* describes a frame shifting insertion (do not use *p.Glu5Valins2fs*3*)
  - *p.(Tyr4*)* indicates RNA nor protein was analysed but the predicted consequence of the change c.12delC in the sequence ATG-GAT-GCA-TA***C***-GTG-ACG to ATG-GAT-GCA-TA.-G TG-A CG is a Tyr to translation termination codon.
  - *p.Asp2Metfs*4 (alternatively p.Asp2fs)* describes the consequence of the change c.4delG in the sequence ATG-***G***AT-GCA-TAC-GTG-ACG to ATG- .AT-G CA-T AC-G TG-A CG.
  - *p.Glu5Valfs*5 (alternatively p.Glu5fs)* describes the consequence of the change c.6_13dup in the sequence ATG-GA***T-GCA-TAC-G***AG-ATG-AGG to ATG-GAT-GCA-TAC-G***T-G CA-T AC-G*** AG-A TG-A GG.
  - ***date 2012-11-01*** *p.Ile327Argfs*? (alternatively p.Ile327fs)* describes the consequences of a frame shifting change (e.g. a 1-nucleotide insertion) with Isoleucine-327 as the first affected amino acid, replacing it for an Arginine and creating a new reading frame which does not encounter a new stop codon (*see FAQ*).

NOTE: the changes observed should be described on protein level and not try to incorporate any knowledge regarding the change at DNA-level (*see Recommendation*). Thus, *p.His150Hisfs*10* is not correct, but *p.Gln151Thrfs*9* is.

Extensions NEW

Extensions affect either the first (*start, translation initiation, N-terminus. ATG*) or last codon (*translation termination, stop*) and as a consequence extend the protein sequence N- or C-terminally with one or more amino acids. Extensions are described using "***ext***" after a description of the change at the first amino acid affected and followed by a description of the position of the new translation initiation or termination codon.

- **new translation initiation site** NEW (*see Discussion*) ***date 2012-08-31***
  a change affecting the translation initiation codon (Met-1) introducing a ***new upstream initiation codon*** extending the N-terminus of the encoded protein described using "***ext-#***" where "-#" is the position of the new initiation codon (*Met-#*)
  - *p.Met1ext-5* - a variant in the 5' UTR activates a new upstream translation initiation site starting with amino acid *Met-5* (Methionine -5)
  - *p.Met1Valext-12* - amino acid *Met1* is changed to *Val* activating an upstream translation initiation site at position -12 (Methionine -12)
    NOTE: recently modified from *p.Met1ValextMet-12* (*see Discussion*)

- **no-stop change** NEW *(substitution in stop codon)*
  a change affecting the translation termination codon (Ter/*) introducing a ***new downstream termination codon*** extending the C-terminus of the encoded protein described using "***extTer#***" (alternatively "*ext*#*") where "#" is the position of the new stop codon (*Ter# / *#*)
  - *p.*110Glnext*17* (alternatively *p.Ter110GlnextTer17* or *p.*110Qext*17*) describes a variant in the stop codon (Ter/*) at position 110, changing it to a codon for Glutamine (Gln, Q) and adding a tail of new amino acids to the protein's C-terminus ending at a new stop codon (Ter17/*17)
  - ***date 2012-11-01*** *p.*327Argext*?* (alternatively *p.Ter327ArgextTer?* or *p.*327Rext*?*) describes a variant in the stop codon (Ter/*) at position 327, changing it to a codon for Arginine (Arg, R) and adding a tail of new amino acids of unknown length since the shifted frame does not contain a new stop codon (*see FAQ*).

More changes in one individual

Two or more changes in one individual are described by combining the changes, per chromosome (maternal and paternal), between square brackets ("***[;];[;]***") and using a semicolon (";") as separator: *[first change maternal; second change maternal] ; [first change paternal; second change paternal]*". When changes are in different genes on different chromosomes a space (" ") is used to separate the different chromosomes ("***[;] [;]***").

- **Two changes in one gene on one chromosome**
  - deriving from *two independent changes at DNA level* are described as "[first change;second change]" (*see Discussion*).
    - p.[(Ala25Thr; Gly28Val)] indicates two predicted changes derived from one chromosome (RNA or protein not analysed); amino acid Alanine25 to Threonine and Glycine-28 to Valine
  - deriving from *one change at DNA level* that has more than one effect on RNA/protein level are described as "[first change, second change]" (*see Discussion*).
    - p.[Asn26His, Ala25_Gly29del] describes two protein changes deriving from one change on a chromosome (c.76A>C at DNA level) resulting in two transcripts (RNA level r.[76a>c, 73_88del]) yielding two predicted proteins, one where amino acid Asparagine25 changes to Histidine and one with a deletion of amino acids Asparagine25 to Glycine29
- **Two changes in one gene on different chromosomes** (e.g. in *recessive diseases*)
  p.[Ala25Thr];[Gly28Val] describes two changes derived from a gene on each chromosome (one paternal, one maternal); predicted change amino acid Alanine25 to Threonine on one chromosome and Glycine28 to Valine on the other chromosome (RNA or protein analysed)

Examples

- p.[(Ala25Thr)];[(Gly28Val)] describes two changes derived from one gene on each chromosome (one paternal, one maternal); predicted change amino acid Alanine25 to Threonine on one chromosome and Glycine28 to Valine on the other chromosome (RNA or protein not analysed)

NOTE: the description p.([Ala25Thr];[Gly28Val]) should not be used

- p.[Ala25Thr];[(Pro323Leu)] described a predicted change of amino acid Alanine25 to Threonine derived from one chromosome (RNA or protein analysed) and Proline323 to Leucine derived from the other chromosome (RNA or protein not analysed)
- p.[Ala25Thr];[?] describes a change of amino acid Alanine-25 to Threonine derived from one chromosome and an unknown change iderived from the other (RNA or protein not analysed)
  NOTE: "*unknown change in the other allele*" does not only mean that no DNA-change was detected in the other chromosome but includes cases where the consequence of a detected change is unclear or can not be predicted (e.g. the consequence of a change at the splice site)
- p.[Ala25Thr];[=] denotes a change of amino acid Alanine25 to Threonine derived from one chromosome and a normal sequence (indicated by "=") of the other chromosome (*see FAQ*)

-
- NEW ***Two sequence changes in one gene with chromosomes unknown*** *are described as "[change 1(;)change 2]" (see Disucssion).*
  - *p.[Ala25Thr(;)Pro323Leu] describes that two changes were identified in one individual (amino acid Alanine25 to Threonine and Proline323 to Leucine, RNA or protein analysed), but it is not known whether these changes are on the same chromosome (in cis) or on different chromosomes (in trans)*
  - *p.[(Ala25Thr(;)Pro323Leu)] describes that two changes were identified in one individual (amino acid Alanine25 to Threonine and Proline323 to Leucine, RNA nor protein analysed), but it is not known whether these changes are on the same chromosome (in cis) or on different chromosomes (in trans). Alternatively p.[(Ala25Thr)(;)(Pro323Leu)] can be used.*
- NEW **Mosaicism** is described using "/"
  - *p.[Arg83=/Arg83Ser] describes a somatic case where a chromosome in some cells contains a normal sequence (p.Arg83=), while other cells contain a Ser at this position (p.Arg83Ser)*
- NEW **Chimerism** is described using "//"
  - *p.[Arg83=//Arg83Ser] describes a chimeric organism where a chromosome in some cells contain a normal sequence (Arg83=), while other cells contain anotehr chromosome with Ser at this position (p.Arg83Ser)*
