Description of sequence changes: examples protein-level

Last modified September 1, 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**
  - reference sequence
- **Examples - protein changes**
  - general
  - **basic changes**
    - silent
    - substitutions *(missense, nonsense, no-stop)*
    - deletion
    - duplication *(variability of short sequence repeats)*
    - insertion
    - inversion *(description not used at protein level)*
    - translocation *(fusion proteins)*
  - complex rearrangements
    - deletion/insertion *(indel, frame shift)*
  - miscellaneous
    - two changes in one allele
    - two change in one individual - allele unknown
    - changes in different alleles *(recessive diseases)*
  - uncertainties *(exact position not known; Southern blot, PCR, arrayCGH, SNP-array, ...)*

-
-

Introduction

Within this page examples will be given for the description of sequence variants on ***protein level***, examples to describe changes at [DNA](#) and [RNA level](#) are given at other pages. All examples are described ***relative to a reference sequence***, here the amino acid (protein) sequence.

Reference sequence

Part of gene		nucleotide numbering *protein* Reference Sequence	nucleotide numbering *coding DNA* Reference Sequence	nucleotide numbering *genomic* Reference Sequence
5' gene flanking region		-	(-300 to -31)	1 to 270
exon 1	5' UTR	-	-30 to -1	271 to 300
	coding region	1 to 4	1 to 12	301 to 312
intron 1		-	12+1 ... 12+50, 13-50 ... 13-1	313 to 412
exon 2		5 to 29 (30)	13 to 88	413 to 488
intron 2		-	88+1 ... 88+100, 89-100 ... 89-1	489 to 689
exon 3		30 to 41	89 to 123	689 to 723
intron 3	contains rare alternatively spliced exon from 800 to 859 (coding DNA 123+77 to 123+136)	-	123+1 ... 123+150, 124-150 ... 124-1	724 to 1023
exon 4		42 to 100	124 to 300	1024 to 1200
			300+1 ... 300+200,	

intron 4		-	*301-200 ... 301-1*	1201 to 1600
exon 5	coding region	101 to 109	301 to 330	1601 to 1630
	3' UTR, containing a (CA)$_7$-stretch from nts 1700 to 1713 (coding DNA *71 to *83); poly-A addition site at 1825 (coding DNA *195)	-	*1 to *220	1631 to 1850
3' gene flanking region		-	(*221 to *370)	1851 to 2000

Legend:

Reference sequence of imaginary gene used for the exaples given on this page. Nucleotide +1 in the coding DNA reference sequence is the A of the ATG translation initiation codon. Abbreviations used: nt = nucleotide, nts = nucleotides, UTR = untranslated region of the mRNA. For a picture of part of this hypothetical sequence *see Figure*.

General

It should be noted that the descriptions at protein level, even more than those at RNA level, are mostly **deduced** and not based on experimental evidence. Publications describing changes at protein level should make it clear whether experimental proof was available or not. In fact, when changes are reported for which experimental proof is not available one should consider to list them between brackets.

Sequence changes at protein level are basically described like those at the DNA level, with a few modifications;

- the three-letter amino acid code is prefered (see Discussion), with "***Ter***" or "*******" designating a translation termination codon; for clarity this page describes changes using the three-letter amino acid
- **amino acid numbering**
  - descriptions start with the amino acid, followed by its number (like p.*Cys24 or p.C24*)
  - the translation initiator Methionine is numbered as +1
  - amino acids after the stop codon (here after *110) are numbered relative to the stop codon as Gly*110+1, Trp*110+2, etc.
- **"silent" changes**
  description of so called "silent" changes in the format p.Leu54Leu (or p.L54L) is not allowed; descriptions should be given at **DNA level**, it is **non-informative** and **not unequivocal** (there are five possibilities at DNA level which may underlie p.Leu54Leu); correct description has the format c.162C>G.

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 should be described without using the specific **">"**-character which is used on DNA and RNA level (i.e. p.Trp26Cys, not p.Trp26>Cys).

- **translation initiation codon** *(Methionine codon, see Discussion)*
  description depends on the consequences of the change on the translation product (protein);
  - ***no protein produced*** - p.0
    e.g. as a consequence of a variant deleting the promoter/first exon or a change in the translation initiation codon (experimental data should be available)
  - ***effect unknown (most cases)*** - p.0? (alternatively p.Met1?)
    when experimental data show that no protein is made, the description *p.0* should be used.
  - ***new translation initiation site***
    - *upstream* - p.Met1ValextMet-12 (alternatively p.M1VextM-12)
      denotes an extension of 12 amino acids (Met-12 to Thr-1) of the protein combined with an amino acid change (Met1Val)
    - *downstream* - p.Phe2_Met46del (alternatively p.F2_M46del)
      denotes inactivation of the normal and activation of a downstream translation initiation site (MET) resulting in deletion of the first 45 amino acids (Met-1 to Lys-45) of the protein.
      NOTE: for the description of the change the 3' rule applies so deletion of Phe2 to Met46.
- **one amino acid to**
  - *another amino acid **(missense change)***
    Tryptophan 26 to a Cysteine; p.Trp26Cys (alternatively p.W26C)
  - *a stop codon **(nonsense change)***
    Tryptophan 26 to a stop codon (*); p.Trp26* (alternatively p.W26*)
    NOTE: this change is ***not*** described as a deletion of the C-terminal end of the protein (e.g. p.Trp26_Arg1623del or p.Trp26_Arg1623del)
- **translation termination codon** *(stop codon, **no-stop change**)*
  - p.*110Glnext*17 (alternatively p.*110Qext*17)
    the stop codon (*) at position 110 is changed to a codon for Glutamine (Gln, Q), adding a tail of 17 new amino acids (incl. Gln110) to the protein's C-terminus after which a new stop codon is reached ('*codon *127*')
  - p.*321Argext*? (alternatively p.*321Rext*?) describes a variant in the stop codon (*) at position 321, 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.

*NOTE: **polymorphic variants*** are sometimes described as p.36Leu/Ile (p.36L/I) or p.36Leu/Leu (p.36L/L) but this is not correct (*see Protein level recommendations*).

Deletions

Deletions are designated by "***del***" after a description of the deleted segment, i.e. the first (and last) amino acid(s) deleted.

- M**K**LGHQQQCC to M_LGHQQQCC is described as p.Lys2del (alternatively p.K2del)
- MKL**GHQ**QQCC to MKL___QQCC is described as p.Gly4_Gln6del (alternatively p.G4_Q6del)
- MKLGHQQQCC to MKLGHQQCC is described as p.Gln8del (p.Q8del)
  NOTE: for deletions in single amino acid stretches or tandem repeats, the most 3' residue is arbitrarily assigned to have been deleted
- if a deletion creates a new amino acid at the deletion junction the change is described as a **insertion/deletion** (*see indels*)
- **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_Lys45del* - a new translation initiation site is activated (at Met46)
  - *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
- **nonsense variant**
  a nonsense variants is 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 designated by "***dup***" after a description of the duplicated segment, i.e. the first (and last) amino acid(s) duplicated.

- MKLGHQQQCC to MKLGHQ**GHQ**QQCC is described as p.Gly4_Gln6dup (alternatively p.G4_Q6dup)
- MKLGHQQQCC to MKLGHQQQ**Q**CC is described as p.Gln8dup (alternatively p.Q8dup)
  NOTE: for duplications in single amino acid stretches or tandem repeats, the most 3' residue is arbitrarily assigned to have been duplicated
- MKLGHQQQCC to MKLGHQ**HQ**QQCC is described as p.His5_Gln6dup (alternatively p.H5_Q6dup)
  NOTE: duplicating insertions in single amino acid stretches (or short tandem repeats) should be described as a duplication and not as an insertion - so in the example shown p.Gln6_Gln7insHisGln (alternatively p.Q6_C7insHQ) is not correct
- **variability of short sequence repeats** are designated as p.Gln6(3_6) (alternatively p.Q6(3_6)) describing that the Glutamine (Gln, Q) stretch starting at position 6 in MKLGHQQQCC is found repeated 3 to 6 times in the population

Insertions

Insertions are designated by **"*ins*"** after a description of the amino acids flanking the insertion site, followed by a description of the inserted amino acids. When the insertion is large it may be described by its length (e.g. p.Lys2_Leu3ins34). However, it should be possible to derive the inserted sequence from the description at DNA level. Duplicating insertions should be described as duplications ([see Discussion](#)).

- p.Lys2_Leu3insGlnSer (alternatively p.K2_L3insQS) describes the cahnge from MKLGHQQQCC to MK**QS**LGHQQQCC
- p.Arg78_Gly79ins23 describes the in-frame insertion of a 23 amino acid sequence between residues Arg-78 and Gly79. Such an insertion can e.g. derive from the inclusion of intronic sequences resulting from a change affecting RNA splicing (see [Examples RNA-level](#)).
  NOTE: the inserted sequence at DNA/RNA level should be specified, where necessary using a sequence database submission (Genbank, EMBL, DDJB) and listing of the accession number.

When an insertion creates a new amino acid at the insertion junction the change is described as an **insertion/deletion** ([see indels](#))

Translocations

Translocations at protein level occur when a translocation at DNA level leads to the production of a fusion protein, joining the N-terminal end of the protein on one chromosome to the C-terminal end of the protein on the other chromosome (and vice versa). No recommendations have been made sofar to describe protein translocations.

- t(X;17)(DMD:p.Met1_Val1506; SGCA:p.Val250_*387) describes a fusion protein resulting from a translocation between the chromosomes X and 17; the fusion protein contains an N-terminal segment of DMD (dystrophin, amino acids Methionine-1 to Valine-1506), and a C-terminal segment of SGCA (alpha-sarcoglycan, amino acids Valine-250 to the stop codon at 387)

Complex rearrangements

Complex rearrangements are rearrangements which consist of several different types of the **six elementary content changes** *substitution, deletion, duplication, insertion, inversion and translocation*. Such rearrangements can be very complex and difficult to describe. Specific recommendations to describe such changes have not made. Complex rearrangements can be best described as a combination of the elementary changes.

Deletion/insertions (indels) are described as a *deletion* followed by an *insertion* (see Discussion)

- p.(Cys28_Lys29delinsTrp) (RNA not analysed, alternatively p.(C28_K29delinsW)) denotes a 3 bp deletion affecting the codons for Cysteine-28 and Lysine-29, substituting them for a codon for Tryptophan
- p.Cys28delinsTrpVal (alternatively p.C28delinsWV) denotes a 3 bp insertion in the codon for Cysteine-28, generating codons for Tryptophan (W) and Valine (V)
- **frame shift changes** (see Discussion) cause, from a specific point onwards, the replacement of the normal C-terminal end of a protein for a new segment, encoded by the shifted reading frame. Frame shift changes can thus be best considered as deletion/insertions (indels). Frame shifts are designated by **"*fs*"** after the amino acid(s) affected by the change. Descriptions either use a **short** ("fs" only) or **long** ("fs*#") notation; the long description should include the change occurring at the site of the frame shift (see Discussion). In "fs*#", "*#" indicates at which codon position the new reading frame ends in a stop codon (*). The position of the stop in the new reading frame is calculated starting at the first amino acid that is changed by the frame shift, and ending at the first stop codon (*#). Thus, effectively, on DNA/RNA level the change might be one or more coding triplets up- or downstream.
  NOTE: the shifted reading frame is thus open for '#-1' amino acids.
  - p.(Arg97Profs*23) (RNA not analysed, short p.(Arg97fs)) denotes a frame shifting change with Arginine-97 as the first affected amino acid, changing into a Proline, and the new reading frame ending in a stop at position 23
  - p.Arg97Hisfs*5 (short p.Arg97fs) denotes a frame shifting change with Arginine-97 as the first affected amino acid, changing into a Histidine, and the new reading frame ending in a stop at position 5
  - p.(Leu30Serfs*3) (RNA not analysed, short p.(Leu30fs)) denotes a frame shifting change that deletes amino acids Leucine-30 to Cysteine-42 (exon 3 of the gene), substituting these for a Serine at the deletion junction and ending in the new reading frame in a stop at position 3
  - p.(Ile327Argfs*?) (RNA not analysed, 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.

Miscellaneous

- **Two changes in one allele** are described as "[first change; second change]" (*see Discussion*)
  - p.[Trp13*; Pro43Ala] (alternatively p.[W13*; P43A]) describes a Tryptophan-13 to stop codon and Proline-43 to Alanine change in one allele (on one chromosome)
  - p.[(Ala25Thr; Gly28Val)] indicates two predicted changes in one allele (RNA nor protein was analysed); amino acid Alanine25 to Threonine and Glycine-28 to Valine
- **One change in one allele yielding two transcripts/two encoded proteins** (e.g. deriving from DNA change that generates 2 different transcripts) are described as "[first change, second change]" (*see Discussion*).
  - p.[Asn26His, Ala25_Gly29del] denotes two protein changes deriving from a change in one allele at DNA level (c.76A>C) resulting in two transcripts (r.[76a>c, 73_88del]) yielding two proteins, one where amino acid Asparagine25 changes to Histidine and one with a deletion of amino acids Asparagine25 to Glycine29
- **Two changes in one individual, unknown in which allele** are described as "[first change (;) second change]" (*see Discussion*)
  - p.[Trp13*(;)Glu61Gln] (alternatively p.[W13*(;)E61Q]) describes a Tryptophan-13 to stop codon and Glutamine-61 to Glutamic acid change in one individual while it is unknown whether these changes are on the same or different alleles
- **Recessive disease (one change in each allele)** are are described as "[change allele 1];[change allele 2]" (*see Discussion*)
  - p.[Trp13*];[Cys28Arg] (alternatively p.[W13*];[C28R]) describes a Tryptophan-13 to stop codon change in one allele (chromosome) and a Cysteine-28 to Arginine change in the other allele (chromosome)
  - p.[Trp13*];[?] (alternatively p.[W13*];[?]) describes a Tryptophan-13 to stop codon change in one allele (chromosome) and an unknown change in the other allele (chromosome)
  - p.[Trp13*];[=] (alternatively p.[W13*];[=]) describes a Tryptophan-13 to stop codon change in one allele (chromosome) and a normal sequence in the other allele (chromosome)
- **Two changes in different genes** are described as SGCA:p.[Arg175*]; SGCB:p.[Cys305Ser] - the SGCA gene contains a nonsense change, the SGCB gene a Cys-to-Ser substitution
- **Mosaicism** is described using "/"
  - p.[=/Arg83Ser] describes a mosaic organism or somatic tissue where the allele in some cells contain the normal sequence (Arg83 described as '='), while other cells contain a Ser at this position
- **Chimerism** is described using "//"
  - p.[=//Arg83Ser] describes a chimeric organism where the allele in some cells contain the normal sequence (Arg83 described as '='), while other cells contain a Ser at this position

---