HGVS
HUMAN GENOME
VARIATION SOCIETY

Standards - definitions, symbols, nucleotides, codons, amino acids *(v2.0)*

Last modified September 11, 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.

Content

Definitions

for the description of sequence variants the following definitions are used

DNA/RNA

- **conversion** = a sequence change where a range of nucleotides are replaced by a sequence from elsewhere in the genome
- **deletion** = a sequence change where one or more nucleotides are removed (*deleted*)
- **deletion/insertion (indel)** = a sequence change where one or more nucleotides are replaced by one or more other nucleotides
  NOTE: when ***one*** nucleotide is replaced by ***one*** other nucleotide the change is called a *substitution*
- **duplication** = a sequence change where a copy of one or more nucleotides are inserted directly 3'-flanking of the original copy
  NOTE: when the copied sequence is not inserted directly 3'-flanking of the original copy the change is called an *insertion*
- **insertion** = a sequence change where one or more nucleotides are inserted between two nucleotides but where the insertion is not a copy of a sequence immediately 5'-flanking (*see duplication*)

- **inversion** = a sequence change where more than one nucleotide replacing the original sequence are the reverse complement of the original sequence
- **substitution** = a sequence change where one nucleotide is replaced by one other nucleotide
  NOTE: a sequence change where ***one nucleotide*** is replaced by ***more*** than one other nucleotide is a deletion-insertion (indel)
  NOTE: a sequence change where ***more than one nucleotide*** is replaced by ***one or more*** other nucleotide is a deletion-insertion (indel)
- **translocation** = a sequence change where the sequence of one chromosome at the so called breakpoint (or junction) changes to that of another chromosome
  NOTE: a translocation occurs when 2 chromosomes break and the fragments rejoin to the another chromosome. A full description of a (reciprocal) translocation consists of 2 parts, one describing the first junction, the second describing the other junction (e.g. the chromosome 4;X junction as well as the chromosome X;4 junction)
- **transposition** = a sequence change where a range of nucleotides moves from one position to another position, i.e. a deletion at one positions combined with the insertion of the deleted sequence at another position

Protein

- **conversion** = a sequence change where a range of amino acids are replaced by a sequence from elsewhere in the genome
- **deletion** = a sequence change where one or more amino acids are removed (*deleted*)
- **deletion/insertion (indel)**
  - ***in frame*** = a sequence change where one or more amino acids are replaced by one or more other namino acids
    NOTE: when ***one*** amino acid is replaced by ***one*** other amino acid the change is called a *substitution*
  - ***frame shift*** = a sequence change that affects 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***)
- **duplication** = a sequence change where a copy of one or more amino acids are inserted directly 3'-flanking of the original copy
  NOTE: when the copied sequence is not inserted directly 3'-flanking of the original copy the change is called an *insertion*
- NEW **extension** = a sequence change that affects 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
- **insertion** = a sequence change where one or more amino acids are inserted between two amino acids but where the insertion is not a copy of a sequence immediately 5'-flanking (*see duplication*)
- **substitution** = a sequence change where one amino acid is replaced by one other amino acid
  NOTE: a sequence change where ***one amino acid*** is replaced by ***more*** than one other amino acid is a deletion-insertion (indel)
  NOTE: a sequence change where ***more than one amino acid*** is replaced by ***one or more*** other amino acids is a deletion-insertion (indel)

Characters used

Below an overview of all different characters and signs used in the description of sequence variants with their meaning.

- reference sequences
  - c. = coding DNA reference sequence
  - g. = genomic reference sequence

- m. = mitochondrial reference sequence
    - n. = non-coding RNA reference sequence (gene producing an RNA transcript but not a protein)
      NOTE: suggested addition, see SVD-WG002
    - r. = RNA reference sequence
    - p. = protein reference sequence
- **diferent transcripts / protein isoforms generated from one gene**
    - _v = specifies transcript variants in coding DNA variant descriptions (e.g. NM_000109.3(DMD_***v2***):c.4G>T)
    - _i = specifies protein isoforms in protein variant descriptions (e.g. NM_000109.3 (DMD_***i2***):p.Glu2*)
- **numbering**
    - *genomic, mitochondrial, non-coding RNA, RNA and protein reference sequence*
        - N = nucleotide N in reference sequence (e.g. 311A>G)
    - *coding DNA reference sequence*
        - N = nucleotide N in protein coding sequence (e.g. 11A>G)
        - -N = nucleotide N 5' of the ATG translation initiation codon (e.g. -4A>G)
          NOTE: so located in the 5'UTR or 5' of the transcription initiation site (upstream of the gene, incl. promoter)
        - *N = nucleotide N 3' of the translation stop codon (e.g. *6A>G)
          NOTE: so located in the 3'UTR or 3' of the polyA-addition site (incl.downstream of the gene)
        - N+M = nucleotide M in the intron after (3' of) position N in the coding DNA reference sequence (e.g. 30+4A>G)
        - N-M = nucleotide M in the intron before (5' of) position N in the coding DNA reference sequence (e.g. 301-2A>G)
        - -N+M / -N-M = nucleotide in an intron in the 5'UTR (e.g. -45+4A>G)
        - *N+M / *N-M = nucleotide in an intron in the 3'UTR (e.g. *212-2A>G)
        - ***NOTE:*** suggestions have been made to specifically number non-transcribed nucleotides (i.e. 5' of the transcription initiation site (cap-site) or 3' of the polyA-addition site), but these are currently not part of the HGVS recommendations (*see Discussion*).
- **specific characters**
    - + (plus) = *see Standards - numbering*
    - - (minus) = *see Standards - numbering*
    - * (asterisk) = translation termination (stop) codon (*see Standards - amino acids*)
    - _ (underscore) = nucleotide numbering, used to indicate a range (e.g. in combination with a deletion, duplication, insertion or variable sequence)
    - > (greater than) = changes to (*substitution*)
        - c.5T>G substitution
          NOTE: *used at DNA and RNA level, not at protein level*
    - : (colon) = separates the description of a reference sequence and the actual description of a variant
      e.g. M13855.3:c.1A>G
    - [] = encloses changes from one allele (chromosome)
        - c.[76A>C; 83G>C] two changes in one allele (chromosome)
        - c.[76A>C];[83G>C] changes in the two alleles (chromosomes)
        - c.123+74TG[4];[5] a TG di-nucleotide repeat of length 4 on one allele and of length 5 on the other allele
        - c.32-?_357+?[3] a **triplication** of an exon (coding DNA reference sequence running from nucleotide 32 to 357)
    - ; (semi-colon) = separator between different changes in one allele or between two alleles
        - c.[76A>C]; [83G>C] changes in the two alleles (chromosomes)

        - c.[76A>C; 83G>C] two changes in one allele (chromosome)
        - c.[76A>C (;) 83G>C] two changes where it is unknown whether they are in the same or different alleles (chromosomes)
    - ^ = or
        - c.[370A>C^372C>R] back translation of the variant p.Ser124Arg, where the Ser-124 codon at c.370_372 is AGC; assuming a substitution change Arg-124 can be encoded by six possible codons, CGN (CGA, CGG, CGC, CGT) and AGR (AGA, AGG)
    - , (comma) = separator between different transcripts or proteins generated from one allele (chromosome)
        - r.[76a>c, 73_87del] denotes the nucleotide change c.76A>C causing the appearance of two RNA molecules, one carrying variant 76a>c, and one containing a deletion of nucleotides 73 to 87
        - 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]); amino acid Asparagine-25 to Histidine and a deletion of amino acids Asparagine-25 to Glycine-29
        -
    - = (equals) = indicates 'identical to reference sequence' (no change, wild type sequence)
    - ? (question mark) = unknown

    - / = mosaic cases; separator between the different nucleotides, transcripts and proteins generated from one allele (chromosome, like used by the ISCN)
        - c.[=/85C>T] somatic or germline mosaicism, i.e. the sample is a mix of c.= and c.85C>T alleles
    - // = chimeric cases, separator between different nucleotides, transcripts and proteins generated from a mix of four alleles (chromosomes, like used by the ISCN)
        - c.[=//85C>T] chimerism, i.e. the sample is a mix of two different populations of genetically distinct cells
    - ( ) = indicates uncertainty in the description of a change
        - c.[76A>C (;) 83G>C] two changes where it is unknown whether they are in the same or different alleles (chromosomes)
        - c.123+74TG(3_6) a TG di-nucleotide repeat found repeated 3 to 6 times in the population (located at nucleotide 74 in the intron following coding DNA nucleotide c.123)
    - 0 (zero) = indicates no product / nothing
        - c.0 = no DNA from allele detected, e.g. c.[76A>C];[0] for a variant in a X-linked gene in a male
        - r.0 = no RNA from allele detected, e.g. from a promoter variant or deletion
        - p.0 = no protein from allele detected, e.g. from a variant in the translation initiation codon
    - NEW ***Under discussion***
      { } (curly braces) = enclose "sub-alleles", i.e. changes within the range of duplications, inversions, insertions and gene conversions using nested and composite change formats (*see Proposal for complex variants*)
        - c.24_65dup{46G>T} an duplication of nucleotides 24 to 65 with a the variant c.46G>T in the duplicated copy
- **nucleotides, codons & amino acids (V2.0)**
    - DNA
    - RNA
    - protein
        - one and three letter amino acid code
        - * = translation termination codon (*stop codon*)
- **others**
    - chr = chromosome (e.g. chr19 or chrX)

- del = deletion
- dup = duplication
- ext = extension (e.g. N- or C-terminus of protein)
- ins = insertion
- inv = inversion
- con = (gene) conversion
- fs = frame shift
- t = translocation; e.g. *t(X;4)(p21.2;q34)*

Nucleotides (DNA / RNA)

For the complete and official list with further details go to *IUPAC-IUBMB* or *NCBI* site.

DNA

Symbol	Meaning	Description
A	A	Adenine
C	C	Cytosine
G	G	Guanine
T	T	Thymine
B	C, G or T	not-A (B follows A in alphabet)
D	A, G or T	not-C (D follows C in alphabet)
H	A, C or T	not-G (H follows G in alphabet)
K	G or T	Keto
M	A or C	aMino
N	A, C, G or T	aNy
R	A or G	puRine
S	G or C	Strong interaction (3 H-bonds)
V	A, C or G	not-T / not-U (V follows U in alphabet)
W	A or T	Weak interaction (2 H-bonds)
Y	C or T	pYrimidine
Used in alignments only		

X	A, C, G or T	masked nucleotide
-	none	gap of indeterminate length

RNA

Symbol	Meaning	Description
a	A	Adenosine
c	C	Cytidine
g	G	Guanosine
u	U	Uridine
b	c, g or u	not-a (b follows a in alphabet)
d	a, g or u	not-c (d follows c in alphabet)
h	a, c or u	not-g (h follows g in alphabet)
k	g or u	keto
m	a or c	amino
n	a, c, g or u	any
r	a or g	purine
s	g or c	strong interaction (3 H-bonds)
v	a, c or g	not-u (v follows u in alphabet)
w	a or u	weak interaction (2 H-bonds)
y	c or u	pyrimidine

Genetic code

NOTE: *'*' (alternatively 'Ter') is used to indicate a translation stop codon (replacing the 'X' used previously). To support translation from a DNA sequence a "T" is used in the codons although in nature RNA is translated so the codons contain U's.*

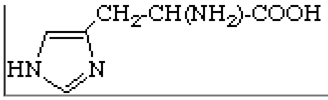
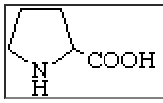
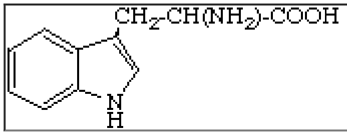
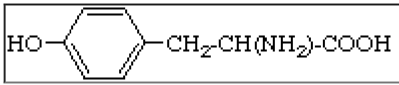
Nucleotide position in codon					
first	**second**				**third**
	T	**C**	**A**	**G**	

T	TTT - Phe	TCT - **Ser**	TAT - Tyr	TGT - Cys	**T**
	TTC - Phe	TCC - **Ser**	TAC - Tyr	TGC - Cys	**C**
	TTA - **Leu**	TCA - **Ser**	TAA - */Ter	TGA - */Ter	**A**
	TTG - **Leu**	TCG - **Ser**	TAG - */Ter	TGG - Trp	**G**
C	CTT - **Leu**	CCT - Pro	CAT - His	CGT - ***Arg***	**T**
	CTC - **Leu**	CCC - Pro	CAC - His	CGC - ***Arg***	**C**
	CTA - **Leu**	CCA - Pro	CAA - Gln	CGA - ***Arg***	**A**
	CTG - **Leu**	CCG - Pro	CAG - Gln	CGG - ***Arg***	**G**
A	ATT - Ile	ACT - Thr	AAT - Asn	AGT - **Ser**	**T**
	ATC - Ile	ACC - Thr	AAC - Asn	AGC - **Ser**	**C**
	ATA - Ile	ACA - Thr	AAA - Lys	AGA - ***Arg***	**A**
	ATG - Met	ACG - Thr	AAG - Lys	AGG - ***Arg***	**G**
G	GTT - Val	GCT - Ala	GAT - Asp	GGT - Gly	**T**
	GTC - Val	GCC - Ala	GAC - Asp	GGC - Gly	**C**
	GTA - Val	GCA - Ala	GAA - Glu	GGA - Gly	**A**
	GTG - Val	GCG - Ala	GAG - Glu	GGG - Gly	**G**

Amino acid descriptions

*For the complete and official list with further details go to IUPAC-IUBMB or NCBI site. (**NOTE:** formula-images were copied from the IUPAC-IUBMB site)*

One letter code	Three letter code	Amino acid	Possible codons	Systemic name	Formula
A	Ala	Alanine	GCA, GCC, GCG, GCT	2-Aminopropanoic acid	CH3-CH(NH2)-COOH
B	*Asx*	*Aspartic acid or Asparagine*	*AAC, AAT, GAC, GAT*		
C	Cys	Cysteine	TGC, TGT	2-Amino-3-mercaptopropanoic acid	HS-CH2-CH(NH2)-COOH
D	Asp	Aspartic acid	GAC, GAT	2-Aminobutanedioic acid	HOOC-CH2-CH(NH2)-COOH
E	Glu	Glutamic acid	GAA, GAG	2-Aminopentanedioic acid	HOOC-[CH2]2-CH(NH2)-COOH
F	Phe	Phenylalanine	TTC, TTT	2-Amino-3-phenylpropanoic acid	C6H5-CH2-CH(NH2)-COOH
G	Gly	Glycine	GGA, GGC, GGG, GGT	Aminoethanoic acid	CH2(NH2)-COOH

H	His	Histidine	CAC, CAT	2-Amino-3-(1H-imidazol-4-yl)-propanoic acid	
I	Ile	Isoleucine	ATA, ATC, ATT	2-Amino-3-methylpentanoic acid	C2H5-CH(CH3)-CH(NH2)-COOH
K	Lys	Lysine	AAA, AAG	2,6-Diaminohexanoic acid	H2N-[CH2]4-CH(NH2)-COOH
L	Leu	Leucine	CTA, CTC, CTG, CTT, TTA, TTG	2-Amino-4-methylpentanoic acid	(CH3)2CH-CH2-CH(NH2)-COOH
M	Met	Methionine	ATG *(translation initiation)*	2-Amino-4-(methylthio)butanoic acid	CH3-S-[CH2]2-CH(NH2)-COOH
N	Asn	Asparagine	AAC, AAT	2-Amino-3-carbamoylpropanoic acid	H2N-CO-CH2-CH(NH2)-COOH
P	Pro	Proline	CCA, CCC, CCG, CCT	Pyrrolidine-2-carboxylic acid	
Q	Gln	Glutamine	CAA, CAG	2-Amino-4-carbamoylbutanoic acid	H2N-CO-[CH2]2-CH(NH2)-COOH
R	Arg	Arginine	AGA, AGG, CGA, CGC, CGG, CGT	2-Amino-5-guanidinopentanoic acid	H2N-C(=NH)-NH-[CH2]3-CH(NH2)-COOH
S	Ser	Serine	AGC, AGT, TCA, TCC, TCG, TCT	2-Amino-3-hydroxypropanoic acid	HO-CH2-CH(NH2)-COOH
T	Thr	Threonine	ACA, ACC, ACG, ACT	2-Amino-3-hydroxybutanoic acid	CH_3-CH(OH)-CH(NH_2)-COOH
U	*Sec*	*Selenocysteine*	TGA, ...		*H2N-CH(COOH)--CH2-SeH*
V	Val	Valine	GTA, GTC, GTG, GTT	2-Amino-3-methylbutanoic acid	(CH3)2CH-CH(NH2)-COOH
W	Trp	Tryptophan	TGG	2-Amino-3-(lH-indol-3-yl)-propanoic acid	
X	Xaa	unknown or 'other'	NNN		
Y	Tyr	Tyrosine	TAC, TAT	2-Amino-3-(4-hydroxyphenyl)-propanoic acid	
Z	*Glx*	*Glutamic acid or Glutamine*			
*	* (Ter)	**Ter**mination	TAA, TAG, TGA *(translation*	HGVS addition (V2.0)	

			termination)		
Used in alignments only					
-	-	*gap of indeterminate length*			-

| Top of page | Homepage | Check-list | Standards |
| **Recommendations:** general, DNA, RNA, protein, uncertain |
| Discussions | FAQ's | History |
| **Example descriptions:** QuickRef, DNA, RNA, protein |

Website Created by Rania Horaitis, Nomenclature by J.T. Den Dunnen - Disclaimer