HUMAN MUTATION 15:7–12 (2000)

MDI SPECIAL ARTICLE

Mutation Nomenclature Extensions and Suggestions to Describe Complex Mutations: A Discussion

Johan T. den Dunnen[1]* and Stylianos E. Antonarakis[2]*

[1]*MGC-Department of Human and Clinical Genetics, Leiden University Medical Center, Leiden, The Netherlands*
[2]*Division of Medical Genetics, University of Geneva Medical School, Geneva, Switzerland*

Consistent gene mutation nomenclature is essential for efficient and accurate reporting, testing, and curation of the growing number of disease mutations and useful polymorphisms being discovered in the human genome. While a codified mutation nomenclature system for simple DNA lesions has now been adopted broadly by the medical genetics community, it is inherently difficult to represent complex mutations in a unified manner. In this article, suggestions are presented for reporting just such complex mutations. Hum Mutat 15:7–12, 2000.

KEY WORDS: **complex mutation; mutation detection; mutation database; nomenclature; MDI**

INTRODUCTION

Recently, a nomenclature system has been suggested for the description of changes (mutations and polymorphisms) in DNA and protein sequences [Antonarakis et al., 1998]. 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. The goal of this article is to make suggestions for additional mutation nomenclature recommendations and to stimulate discussions as to how far the rules should go regarding the description of complex mutations. 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 on the World Wide Web (http://www.dmd.nl/mutnomen.html). We also invite investigators to communicate with us complicated cases with a suggestion of how to describe these (send e-mail to: ddunnen@ruly46.MedFac.LeidenUniv.nl and Stylianos.Antonarakis@medecine.unige.ch); an overview will be listed at the same WWW-address. Ultimately, this list of examples for the description of complicated cases may evolve into a uniformly accepted reference for mutation nomenclature.

The nomenclature needs to be accurate and unambiguous, but flexible, and the nucleotide change must always be included in the original report. The items missing in the current nomenclature recommendations which will be covered in this article include simple omissions (changes in mitochondrial DNA), simple mutations (duplications, inversions), and more complex mutations (mutations in recessive disease, insertion/deletions, changes inside codons, frame shifts). It should be noted that when describing a mutation as a "deletion," "insertion," or "other," one should always indicate which level is described; a substitution at the DNA level may cause an insertion at the RNA level and a nonsense mutation at the protein level. Furthermore, for descriptions at protein and RNA levels, which are mostly *derived from* the DNA sequence with only rarely experimental proof, it should be clear that recommendations cover a description of the *consequence* rather than the *nature* of the sequence change. It can be debated whether this should be done, but we believe it should.

A weak element in many reports is often the description at the level of mature or processed messenger RNA. Many reports fail to mention clearly, and discriminate in tabular listings, at which level the sequence variations reported were

Received 7 September 1999; accepted revised manuscript 4 October 1999.

*Correspondence to: Johan T. den Dunnen, Dept. of Human and Clinical Genetics, LUMC, Wassenaarseweg 72, 2333 AL Leiden, The Netherlands. E-mail: ddunnen@ruly46.MedFac.LeidenUniv.nl; or Stylianos E. Antonarakis, Division of Medical Genetics, University of Geneva Medical School, 9 Avenue de Champel, 1211 Geneva, Switzerland. E-mail: Stylianos.Antonarakis@medecine.unige.ch

analyzed. Consequently, it often remains unclear whether a change might have an effect on the RNA level or whether the suggested effect on RNA has been proven or not. That this is not trivial is exemplified by several cases where changes were originally reported as silent or missense (based on DNA data), although they actually represented changes affecting mRNA processing [see, e.g., Richard and Beckmann, 1995]. Several of the recommendations below are specifically made to clarify this issue.

Nomenclature recommendations by themselves do not safeguard against mistakes. A clear example is the recommendation to use the one-letter amino acid code for descriptions at the protein level. Several examples from recent publications show that it is not rare that mistakes are made. Codes are mixed up easily for amino acids which have the same initial letter (e.g., Alanine, Arginine, Asparagine, and Aspartic acid or Glycine, Glutamine, and Glutamic acid). Hopefully, electronic tools for submission of sequence changes, with automatic error checks, will help to avoid these problems (see http://ariel.ucs.unimelb.edu.au:80/~cotton/entry.htm).

SUMMARY OF PUBLISHED RECOMMENDATIONS

General [from Antonarakis et al., 1998]

- Sequence variations are best described at the DNA level.
- The accession number in primary sequence databases (Genbank, EMBL, DDJB, SWISS-PROT) should be mentioned in the publication/database submission. When available, the genomic reference sequence is preferred. For each gene, a reference sequence needs to be established.
- To avoid confusion, the nucleotide number is preceded by "g." when a genomic or by "c." when a cDNA reference sequence is used.
- For genomic DNA and cDNA sequences, the A of the ATG of the initiator Methionine codon is denoted nucleotide +1 (there is no nucleotide zero). The nucleotide 5′ to +1 is numbered –1.
- For variations in single nt (or amino acid) stretches or tandem repeats, the most 3′ copy is arbitrarily assigned to have been changed (e.g., ATGTGCA to ATGCA is described as 4-5delTG).
- Two mutations in the same allele are listed within brackets, separated by a semicolon; e.g. [1997G>T; 2001A>C] (see below).
- A unique identifier should be obtained for each mutation. Preferably, locus-specific database curators should assign unique identifiers. Alternatively, the OMIM unique identifier or the HGMD entry can be used as a reference source for previously cataloged mutations.

Description at the DNA Level

- Nucleotide changes start with the nucleotide number and the change follows this number; {nucleotide interval}{sequence changed nucleotide}{type of change}{sequence new nucleotide}.
- Substitutions are designated by ">"; 1997G>T denotes that at nt 1997 of the reference sequence a G is changed to a T.
- Deletions are designated by "del" after the deleted interval (followed by the deleted nts). 1997-1999del (alternatively 1997-1999delTTC) denotes a TTC deletion from nts 1997 to 1999. A TG deletion in the sequence ACTGTGTGCC (A is nt 1991) is designated as 1997-1998del (or 1997-1998delTG).
- Insertions are designated by "ins," followed by the inserted nts. 1997-1998insT denotes that a T was inserted between nts 1997 and 1998. A TG insertion in the TG-tandem repeat sequence of ACTGTGTGCC (A is nt 1991) is described as 1998-1999insTG (where 1998 is the last G of the TG-repeat).
- Variability of short sequence repeats, e.g., in ACTGTGTGCC (A is nt 1991), is designated as 1993(TG)3-22 with nt 1993 containing the first TG-dinucleotide which is found repeated 3 to 22 times in the population.
- Intron mutations are designated by the intron number (preceded by "IVS") or cDNA position; positive numbers starting from the G of the GT splice donor site, negative numbers starting from the G of the AG splice acceptor site. IVS4-2A>C (1998-2A>C) denotes the A to C substitution at nt –2 of intron 4, at the cDNA level positioned between nucleotides 1997 and 1998. IVS4+1G>T (1997+1G>T) denotes the G to T substitution at nt +1 of intron 4. When the full-length genomic sequence is known, the mutation is best designated by the nt number of the genomic reference sequence.

Description at the Protein Level

Note that recommendations for sequence variations at the protein level describe the deduced consequence and not the nature of the mutation.

TABLE 1. Sequence Variation Description at the DNA Level

Type of variation	Description	Alternative	Remark
	g.12T>A		change described in relation to a genomic sequence
	c.12T>A		change described in relation to a cDNA sequence
	m.12T>A		change described in relation to a mitochondrial sequence
	[6T>C + 13_14del]		one allele containing two changes
	[13_14del] + [?]		mutations in the 2 alleles of a recessive disease, one being unknown (=?)
Substitution	12T>A		
	15+1G>C	IVS1+1G>C	change in intronic sequence (splice donor site)
	93-2A>G	IVS1-2A>G	change in intronic sequence (splice acceptor site)
Deletion	13_14delTT	13_14del	
	IVS1_IVS5	*15+?_645+?del or EX2_EX5del*	genomic deletion of exons 2 to 5
Duplication	*10_11dupTG*	*10_11dup*	
Insertion	14_15insT		
Inversion	*4_15inv*		
Insertion/deletion	*112_117delAGGTCAinsTG*	*112_117delinsTG or 112_117>TG*	also known as ''indel''
Complex	*depends on change*		*examples at http://www.dmd.nl/mutnomen.html*

Suggestions from this article in italics. Sample sequence 5′-CATGC ATG CAT GCA TGT TTC GTGAGTATATC-3′ with nucleotides -CATGC- being the 5′ untranslated region in exon 1, -ATG-being the translation site and -GTGAGT...- being the first nucleotides in intron 1.

Furthermore, it should be avoided to report changes at the amino acid level without including the description at the nucleotide level.

- The codon for the initiator Methionine is codon 1.
- Stop codons are designated by X. R97X denotes a change of Arginine 96 to a termination codon.
- The single letter amino acid code is recommended, three letter code is acceptable.
- Amino acid changes are described in the format {code first amino acid changed}{amino acid interval}{code new amino acid or type of change}. Y97S denotes Tyrosine97 is substituted by a Serine.
- Deletions are designated by "del" after the amino acid interval. T97-C102del (or T97-C102del6) denotes that amino acid Threonine97 to Cysteine102 are deleted.
- Insertions are designated by "ins" after the amino acid interval, followed by the inserted amino acids or the number of amino acids inserted. T97-W98insLQS (or T97-W98ins3) denotes that three amino acids are inserted between amino acids 97 and 98 (i.e., after Threonine97).

EXTENSIONS AND ADDITIONAL SUGGESTIONS

Sequence Variations in Mitochondrial DNA Suggestion (from David Fung, Camperdown,

TABLE 2. Sequence Variation Description at the Level of Mature or Processed Messenger RNA

Type of variation	Description	Alternative	Remark
	r.15c>u		*change described in relation to a RNA sequence*
	as for DNA but lower case letters		
Splice variants	*[16g>t + 16_112del]*		*mutation affecting splicing, producing two mRNAs, one normal and one with a deletion of nucleotides 16 to 112 (exon 2)*
Complex	*depends on change*		*examples at http://www.dmd.nl/mutnomen.html*

Suggestions from this article in italics. Sample sequence 5′-CAUGC AUG CAU GCA UGU UUC GUC-3′ with nucleotides -CAUGC- being the 5′ untranslated region in exon 1 and -AUG- being the translation initiation codon.

TABLE 3. Sequence Variation Description at the Protein Level

Type of variation	Description	Alternative	Remark
	p.R5S		change described in relation to a *protein* sequence
	others		see Table I
Substitution	R5S		
	W4X		change to stop codon
Deletion	L3_W4del		
Duplication	*L3_W4dup*		
Insertion	W4_R5insK		
Inversion	–		not relevant
Frame shift	*W4fsX8*	*W4fs*	*frame shift causing a translational stop 5 codons downstream*
Complex	*depends on change*		*examples at http://www.dmd.nl/mutnomen.html*

Suggestions from this article in italics. Sample sequence M-I-L-W-R-R-C, with amino acid M representing the translation initiating Methionine.

Australia): when a mitochondrial reference sequence is used, the nucleotide number is preceded by "m." (e.g., m.1203A>T).

Sequence Variations in Protein Sequences

Suggestion: when a protein reference sequence is used, the amino acid is preceded by "p." (e.g., p.K93W).

Descriptions of a Range

Due to the fact that the "–" symbol is used for two different purposes, i.e., to indicate a range (R12-W13) as well as to indicate a negative distance (77-2A>G), current recommendations to describe sequence changes starting and or ending in intronic sequences may easily cause confusion. For example, does 5-77-77del describe a deletion from nt 5-77 to 77 or from nt 5 to 77-77? It would be better not to use the "–" symbol for two purposes. Since for intronic positions both the "–" and "+" symbols are used, a change should involve the "–" symbol separating the first from the last affected nt.

Suggestion: the "_" symbol (underscore) should be used to separate the first from the last affected nt (or amino acid), e.g., 85_86delAG. On the protein level as K23_W29.

Duplications

No recommendations have been made to describe duplications. Although they can be seen as a specific type of insertion, and could be described as such, they often originate through other mutational mechanisms. We therefore prefer to provide a distinctive designation of this type of sequence change

Suggestion: duplications are described as 1992_1994dupCTG (or 1992_1994dup). On the protein level as L78dup.

As a consequence, duplicating insertions in short tandem repeats (or single nucleotide stretches) can also be described as a duplication, e.g., 1997_1998dupTG (now 1998_1999insTG).

Inversions

Recommendations for inversions are missing but can be rather straightforward using the existing rules.

Suggestion: inversions are described as 203_506inv (or 203_506inv304) indicating that the 304 nt's from position 203 to 506 have been inverted.

Mutations in Recessive Diseases

The current suggestion for the description of recessive mutations is [1997G>T + 2001A>G], indicating the substitution of nt 1997 on one allele and of nt 2001 on the other allele. The description should be given the status of a recommendation, which is important for two reasons, 1) it makes clear whether a mutation was found on both alleles, and 2) it ensures that researchers show which mutations were identified in which combinations. The latter is important since severity might depend on the combination of mutations present, while other combinations might not be deleterious at all. However, the current description is not completely straightforward and might cause confusion with that for the description of two mutations in one allele, currently like [1997G>T; 2001A>C].

Suggestion: two variations in one allele are described as [1997G>T + 2001A>C] while variations in different alleles, e.g., in recessive diseases, are designated as [1997G>T] + [2001A>G]. In homozygous cases the format is [1997G>T] + [1997G>T]. When only one mutated allele has been identified, the format is [1997G>T] + [?]. On the protein level, the

designation is likewise, e.g., two variations in one allele [R175X + C305S] versus [R175X] + [C305S] for variations in different alleles.

Mutations Analyzed at Which Level?

Many reports fail to mention clearly, and discriminate in tabular listings, at which level (DNA, RNA, and/or protein) the sequence variations reported were analyzed and whether experimental proof was obtained regarding the descriptions provided at the RNA and protein level.

Suggestion: In tabular listings of the sequence variations identified, it should be stated clearly which variation was analyzed at which level, i.e., DNA, RNA, and/or protein.

Description of Mutations at the RNA Level

When RNA has been analyzed, the effect is often not described properly and recommendations for its description are lacking. Based on the nomenclature rules to describe mutations on the DNA level, suggestions can be simply copied when three additions are made.

Suggestion 1: An "r." is used to indicate that a change is described at the RNA level.

Suggestion 2: To discriminate between descriptions at the DNA and protein levels, lower case letters and the "u" for Uracil are used to describe changes at the RNA level.

Suggestion 3: When one change affects RNA-processing, yielding two or more transcripts, these are described between brackets, separated by a "+" symbol, e.g., [r.76a>c + r.70_77del], i.e., the nt change c.76A>C causes the appearance of two RNA molecules, one carrying the variation only and one containing a deletion of nucleotides 70 to 77 (shift of splice site).

It should be noted here that the designations at the RNA level, similar to that on the protein level, describe the consequence and not the nature of the mutation.

Exon Deletions Detected at the Genomic Level

In diseases such as Duchenne Muscular Dystrophy (DMD), many mutations are found which delete (sets of) whole exons, detected on Southern blot using cDNA probes or using exon specific PCR-tests. Current rules for mutation description do not cover these, with the consequence that everybody uses their own system.

Suggestion: exonic deletions are described as IVS2_IVS5del (alternatives 77+?_923+? or EX3_5del), indicating a deletion starting at an unknown position in intron 2 (after base 77) and ending at an unknown position in intron 5 (after base 923).

Insertion-Deletions

The occurrence of a combination of a deletion and insertion, sometimes named "indel," is not rare. Recommendations for their description have not yet been made. Based on existing terminology, a suggestion can be rather straightforward.

Suggestion: a combination of a deletion and insertion at the same site is described as 112_117delAGGTCAinsTG (alternatively 112_117delinsTG or 112_117>TG). On the protein level, likewise as W33_K35delinsR.

Mutations in the Translation Initiation Site

Currently, mutations in the translation initiating Methionine (M1) are mostly described as a substitution, e.g., M1V. We would like to note that 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.

Sequence Variation in Codons

Current recommendations do not fully cover the description of amino acid substitutions and changes inside codons that do not alter the reading frame.

Suggestion 1: C28_V29delinsW denotes a 3 bp deletion affecting the codons for Cysteine28 and Valine29, changing them to a codon for Tryptophan.

Suggestion 2: C28delinsWV denotes a 3 bp insertion in the codon for Cysteine28, generating codons for Tryptophan and Valine.

Frame-Shifting Mutations

At the protein level, no recommendations have been made regarding the description of frame-shifting mutations. 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.

Suggestion 1: Frame-shifting mutations are designated by "fs." R97fs denotes a frame-shifting change with Arginine97 as the first affected amino acid. The length of the shifted open reading frame may be described using the format R97fsX121, in-

dicating a frame-shifting change with Arginine97 as the first affected amino acid and the new reading frame being open for 23 amino acids.

REFERENCES

Antonarakis SE, the Nomenclature Working Group. 1998. Recommendations for a nomenclature system for human gene mutations. Hum Mutat 11:1–3.

Beaudet AL, the Ad Hoc Committee on Mutation Nomenclature. 1996. Update on nomenclature for human gene mutations. Hum Mutat 8:197–202.

Beutler E, McKusick VA, Motulsky A, Scriver CR, Hutchinson F. 1996. Mutation nomenclature: nicknames, systematic names and unique identifiers. Hum Mutat 8:203–206.

Richard I, Beckmann JS. 1995. How neutral are synonymous codon mutations? Nat Genet 10:259.