HUMAN MUTATION 11:1–3 (1998)

SPECIAL ARTICLE

Recommendations for a Nomenclature System for Human Gene Mutations

Stylianos E. Antonarakis[1]* and the Nomenclature Working Group[†]

[1]*Division of Medical Genetics, Department of Genetics and Microbiology, University of Geneva Medical School, Geneva, Switzerland*

Communicated by R.G.H. Cotton

INTRODUCTION

This document has been written in response to meetings on "Locus-specific Mutation Databases" held on March 24, 1996 in Heidelberg, Germany and "Mutation Databases" on October 29, 1996 in San Francisco, California. As a chairperson of the Nomenclature committee, the first author had accepted the task of preparing and circulating a document with recommendations for debate, further discussions and most importantly for final approval/acceptance at the October 1996 meeting in San Francisco. Four drafts of this document (August 5, 1996, September 5, 1996, November 29, 1996, and May 29, 1997) were distributed to a number of investigators, the majority of whom were co-authors in the Beaudet et al, 1996 and Beutler et al, 1996 papers. This document contains modifications according to their suggestions, opinions and criticisms. Furthermore, many colleagues had offered suggestions, criticisms, and ideas through electronic communication. During the October 1996 meeting in San Francisco, there was a sufficient discussion of these recommendations and it was agreed that another, further modified draft of the document would be posted on the "World Wide Web" for final debate for a period of several weeks. This present document is also the result of all of these discussions and was approved during that October 27, 1997 meeting in Baltimore.

Two manuscripts were recently published in "Human Mutation" that contain mutation nomenclature recommendations (Beaudet et al., 1996 and Beutler et al., 1996). These documents present the views of the authors after discussions during the October 19, 1994, Montreal, and the October 24-25, 1995 Minneapolis meetings. Other previously published papers/letters on nomenclature issues include the following: Beaudet & Tsui, 1993; Beutler, 1993; Antonarakis & McKusick, 1994.

It is obvious that the most unambiguous nomenclature system is that based on genomic DNA. Even in that case, however, length polymorphisms can create a problem in the numbering of nucleotides and therefore a standard, reference sequence ought to be established, preferably by experts. Unfortunately, the entire genomic sequence is only known for a minority of human genes. For the vast majority of genes, the known sequence is that of cDNA. The existence of more than one transcription start site, alternative splicing and utilization of alternative exons and variable number of repeats complicate the nucleotide numbering. Thus here too a reference sequence needs to be established. The nomenclature, at least in the present state of the human genome development,

Received 5 September 1997; accepted 3 October 1997.

*Correpondence to: Dr. Stylianos E. Antonarakis, University of Geneva Medical School, 9 avenue de Champel,1211 Geneva, Switzerland. Fax: 41-22-702.57.06

[†]Other members of the Nomenclature Working Group (in alphabetical order): Michael Ashburner, Cambridge University, U.K.; Arleen D. Auerbach, Rockefeller U, New York, NY, U.S.A.; Arthur L. Beaudet, Baylor College of Medicine, Houston, Texas, U.S.A.; Jacques S. Beckmann, Genethon, Paris, France; Ernest Beutler, Scripps Clinic & Res. Foundation, La Jolla, California, U.S.A.; David N. Cooper, U Wales Coll Medicine, Cardiff, U.K.; Richard G.H. Cotton, Mutation Research Ctr, Melbourne, Australia; Johan T. den Dunnen, Leiden U, Netherlands; Robert J. Desnick, Mt Sinai Sch Medicine, New York, NY, U.S.A.; Charis Eng, Dana-Farber Cancer Institute, Boston, MA, U.S.A.; Kenneth H. Fasman, MIT, Boston, MA, U.S.A.; David Goldman, NIAAA, Rockville, MD, U.S.A.; Kenshi Hayashi, Kyushu U, Fukuoka, Japan; Franklin Hutchinson, Yale U, New Haven, CT, U.S.A.; Haig H. Kazazian, Jr, U Pennsylvania, Philadelphia, PA, U.S.A.; Jeffrey Keen, University College London, U.K.; Mary-Claire King, U. Washington, Seattle, WA, U.S.A.; Heikki Lehvaslaiho, EMBL, EBI, Cambridge, U.K.; Phyllis J. McAlpine, U Manitoba, Winnipeg, Canada; Victor McKusick, Johns Hopkins U, Baltimore, MD, U.S.A.; Arno G. Motulski, U. Washington, Seattle, WA, U.S.A.; Sue Povey, Univ College London, U.K.; Daniel F. Schorderet, U Lausanne, Switzerland; Charles R. Scriver, McGill U, Montreal, Canada M.; Thomas B. Shows, Jr., Roswell Park Cancer Inst, Buffalo, NY, U.S.A.; Andrea Superti-Furga, U Zurich, Switzerland; Agnes H.N. Tay, U Singapore, Singapore; Lap-Chee Tsui, Hospital for Sick Children, Toronto, Ontario, Canada; David Valle, Johns Hopkins, Baltimore, MD, U.S.A.; Mauno Vihinen, U Helsinki, Finland.

needs to be accurate, unambiguous, but flexible. The nucleotide change must always be included in the original report; however, other terms, for example specifying the amino acid change, may be used. The genomic DNA-based nomenclature was termed "systematic" by Beutler at al, 1996 while all other mutation names were considered as "trivial" or "common" by these authors.

RECOMMENDATIONS

A list of recommendations follows:

- For genomic DNA and cDNA , the A of the ATG of the initiator Met codon is denoted nucleotide +1. There is no nucleotide zero (0). The nucleotide 5´ to +1 is numbered –1. If there is more than one potential ATG, a reference consensus may be used. The numbering of nucleotides in the reference sequence in the databases should not be changed and will always be associated with the same (original) accession number.
- The use of lower case g for genomic or c for cDNA in front of the nucleotide number is recommended. To avoid confusion, a dot should separate these from the nucleotide number (g. or c. for genomic or cDNA respectively). The accession number in primary sequence databases (Genbank, EMBL, DDJB) should also be included in the original publication/database submission.
- Nucleotide changes start with the nucleotide number and the change follows this number. 1997G>T denotes that at nucleotide 1997 of the reference sequence, G is replaced by a T.
- Deletions are designated by del after the nucleotide number. 1997delT denotes the deletion of T at nt 1997. 1997-1999del denotes the deletion of 3 nts. Alternatively, this mutation can be denoted as 1997-1999delTTC. For deletions in short tandem repeats, the most 3´ nt is arbitrarily assigned; e.g. a TG deletion in the sequence AATGTGTGCC is designated 1997-1998delTG or 1997-1998del (where 1997 is the first T before C).
- Insertions are designated by ins after the nucleotide interval number. 1997-1998insT denotes that T was inserted in the interval between nts 1997 and 1998. For insertions in short repeats the most 3´ nt interval is arbitrarily assigned; e.g. a TG insertion in the sequence AATGTGTGCC is designated 1997-1998insTG (where 1997 is the last G of the short TG repeat).
- Variability of short sequence repeats is designated as 1997(GT)6-22. In this case, 1997 is the first nucleotide of the dinucleotide GT, which is repeated 6 to 22 times in the population.
- A unique identifier for each mutation should be obtained. The OMIM (http://www3.ncbi.nlm.nih.gov/Omim/) unique identifier can be used, or database curators may assign such unique identifiers. Other existing databases such as the HGMD (http://www.cf.ac.uk/uwcm/mg/hgmd0.html) for example could also be used as a reference source for the already catalogued mutations.
- Intron mutations when the full genomic sequence is not known can be designated by the intron (IVS) number, positive numbers starting from the G of the donor site invariant GT, negative numbers starting from the G of the acceptor site invariant AG. IVS4+1G>T denotes the G to T substitution at nt +1 of intron 4. IVS4-2A>C denotes the A to C substitution at nt –2 of intron 4. Alternatively the cDNA nucleotide numbering may be used to designate the location of the mutation in the adjacent intron. For example, c.1997+1G>T denotes the G to T substitution at nt +1 after nucleotide 1997 of the cDNA. Similarly, c.1997-2A>C denotes the A to C substitution at nt -2 upstream of nucleotide 1997 of the cDNA. When the full length genomic sequence is known, the mutation can also be simply designated by the nt number of the reference sequence.
- Two mutations in the same allele can be listed within brackets as follows: [1997G>T; 2001A>C]. This will also allow the (i) designation of mutations that are only deleterious when they occur in the same allele with additional nucleotide substitutions; (ii) designation of haplotypes of different alleles.
- For amino acid-based systems, the codon for the initiator Methionine is codon 1.
- The single letter amino acid code is recommended. However the three letter code is also acceptable.
- For amino acid nomenclature, the format is Y97S (Tyrosine at codon 97 substituted by Serine). The "wild type" amino acid is given before and the mutant amino acid after the codon number. Therefore there is no confusion as to the significance of G,C,T and A in the nomenclature.
- Stop codons are designated by X. For example R97X (Arginine codon 96 substituted by a termination codon).
- Deletions of amino acids are designated as: T97del denotes that the codon 97 for Threonine is deleted.

- Insertions of amino acids are designated as: T97-98ins denotes that a codon for Threonine is inserted at the interval between codons 97 and 98 of the reference sequence of the protein.
- The first report of a mutation in the literature should contain both a nucleotide and amino acid based name when appropriate.

ADDITIONAL COMMENTS

No recommendations are made at this time for more complex mutations. Detailed description of such mutations and nomenclature proposals can be usually found in the original reference or by the unique identifier. A second phase of recommendations will deal with such issues in the future. In addition, the consequence of a mutation (frameshift, particular splicing abnormality, exon skipping etc) is not included in the mutation name. However, investigators that maintain mutation databases are encouraged to include a field of mutation consequence or mutation mechanism (if known) in their databases.

The recommendations listed above do not always represent a full consensus of the scientific community and the investigators involved in the discussions. Among the numerous other proposals/criticisms, it is worth mentioning the following:

- The "^" sign may be used to determine the interval of an insertion rather than the "–" sign. For example, 1997^1998insG instead of 1997-1998insG.
- The designation of both deleterious mutations in the two alleles of an homozygote for a recessive disorder may be designated as [1997G>T + 2001A>G] to indicate the substitution in nucleotide 1997 of one allele and in nucleotide 2001 of the other allele of the same gene.
- Analogous to g. or c. for the genomic or cDNA numbering system, the p. symbol may be used to clearly distinguish the protein-based nomenclature.
- X may not be the best symbol for a termination codon.

REFERENCES

Antonarakis SE, McKusick VA (1994) Discussion on mutation nomenclature. Hum Mut 4:166.

Beaudet AL and the Ad Hoc Committee on Mutation Nomenclature (1996) Update onnomenclature for human gene mutations. Hum Mut 8:197–202.

Beaudet AL, Tsui LC (1993) A suggested nomenclature for designating mutations. Hum Mut 2:245.

Beutler E (1993) The designation of mutations. Am J HumGenet 53:783–785.

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