 doi:10.1006/geno.2002.6748, available online at http://www.idealibrary.com on IDEAL

Guidelines for Human Gene Nomenclature

Hester M. Wain, Elspeth A. Bruford, Ruth C. Lovering, Michael J. Lush, Mathew W. Wright, and Sue Povey

HUGO Gene Nomenclature Committee, The Galton Laboratory, Department of Biology, University College London, Wolfson House, 4, Stephenson Way, London, NW1 2HE, UK. E-mail: nome@galton.ucl.ac.uk.

INTRODUCTION

Guidelines for human gene nomenclature were first published in 1979 [1], when the Human Gene Nomenclature Committee was first given the authority to approve and implement human gene names and symbols. Updates of these guidelines were published in 1987 [2], 1995 [3], and 1997 [4]. With the recent publications of the complete human genome sequence there is an estimated total of 26,000–40,000 genes, as suggested by the International Human Genome Sequencing Consortium [5] and Venter *et al*. [6]. Thus, the guidelines (http://www.gene.ucl.ac.uk/nomenclature/guidelines.html) have been updated to accommodate their application to this wealth of information, although gene symbols are still only assigned when required for communication. These updates were derived with input from the HUGO Gene Nomenclature Committee (HGNC) International Advisory Committee and attendees of the ASHG01NW Gene Nomenclature Workshop. All approved human gene symbols can be found in the Genew database [7].

The philosophy of the HGNC remains "that gene nomenclature should evolve with new technology rather than be restrictive as sometimes occurs when historical and single gene nomenclature systems are applied" [2].

A summary of the guidelines is presented here:

1. Each approved gene symbol must be unique.
2. Symbols are short-form representations (or abbreviations) of the descriptive gene name.
3. Symbols should only contain Latin letters and Arabic numerals.
4. Symbols should not contain punctuation.
5. Symbols should not end in "G" for gene.
6. Symbols do not contain any reference to species, for example, "H/h" for human.

1. CRITERIA FOR SYMBOL ASSIGNMENT

1.1 Gene

A gene is defined as: "a DNA segment that contributes to phenotype/function. In the absence of demonstrated function a gene may be characterized by sequence, transcription or homology." The overwhelming majority of objects named by HGNC are in this category.

1.2 Locus

The word "locus" is not a synonym for gene but refers to a map position. A more precise definition is given in the *Rules and Guidelines from the International Committee on Standardized Genetic Nomenclature for Mice*, which states: "A locus is a point in the genome, identified by a marker, which can be mapped by some means. It does not necessarily correspond to a gene; it could, for example, be an anonymous non-coding DNA segment or a cytogenetic feature. A single gene may have several loci within it (each defined by different markers) and these markers may be separated in genetic or physical mapping experiments. In such cases, it is useful to define these different loci, but normally the gene name should be used to designate the gene itself, as this usually will convey the most information" (http://www.informatics.jax.org/mgihome/nomen/gene.shtml#trdef).

1.3 Chromosome Region

In the context of gene nomenclature, "chromosome region" is defined as a genomic region that has been associated with a particular syndrome or phenotype, particularly when there is a possibility that several genes within it may be involved in the phenotype. Designation of such regions may be requested by the scientific community and approved by HGNC, e.g., *ANCR* "Angelman syndrome chromosome region" and *CECR* "cat eye syndrome chromosome region."

Symbols may therefore be assigned to the following:

a) Clearly defined phenotypes shown to be inherited predominantly as monogenic mendelian traits, e.g., *BBS1* "Bardet-Biedl syndrome 1."
b) Unidentified genes contributing to a complex trait shown by linkage or association with a known marker, e.g., *IDDM6* "insulin-dependent diabetes mellitus 6."
c) Cloned segments of DNA with sufficient structural, functional, and expression data to identify them as transcribed entities, e.g., *COX8* "cytochrome *c* oxidase subunit VIII."
d) Nonfunctional copies of genes (pseudogenes), e.g., *IL9RP1* "interleukin 9 receptor pseudogene 1."
e) Genes encoded by the opposite (antisense) strand that overlap a known gene, e.g., *IGF2AS* "insulin-like growth factor 2, antisense."
f) Transcribed but untranslated functional DNA segments, e.g., *XIST* "X (inactive)-specific transcript."

0888-7543/02 $35.00

g) Cellular phenotypes from which the existence of a gene or genes can be inferred, e.g., *LOH18CR1* "loss of heterozygosity, 18, chromosomal region 1."
h) EST clusters which suggest a putative gene, e.g., *C1orf1* "chromosome 1 open reading frame 1."
i) Fragments of expressed sequence will be designated a D-number by GDB (The Genome Database), e.g., *DXYS155E* (Appendix 1).
j) Polycistronic genes generated from a single mRNA, but with independent coding sequence, physically separable and non-overlapping with other coding sequence giving independent gene products, e.g., *SNURF* "SNRPN upstream reading frame" and *SNRPN* "small nuclear ribonucleoprotein polypeptide N."
k) Genes of unknown function which share highly similar sequences, e.g., *FAM7A1* "family with sequence similarity 7, member A1," etc.
l) Predicted (*in silico*) genes which show a high degree of sequence homology to well-characterized genes will be assigned the same symbol with an "L" for like, e.g., *TCP10L* "t-complex 10 (mouse)-like."
m) Intronic transcripts (on the same DNA strand) will be assigned separate symbols, usually relating to the gene in which they reside, e.g., *COPG2IT1* "coatomer protein complex, subunit g 2, intronic transcript 1."

Gene symbols will not usually be assigned to alternative transcripts or to genes predicted solely from *in silico* data (with no other supporting evidence, e.g., significant homology to a characterized gene).

2. GENE SYMBOLS

Human gene symbols are designated by upper-case Latin letters or by a combination of upper-case letters and Arabic numerals, with the exception of the *C#orf#* symbols. Symbols should be short in order to be useful, and should not attempt to represent all known information about a gene [8]. Symbols should be inoffensive and should not spell words or match abbreviations that would cause problems with database searching, e.g., DNA. Ideally, symbols should be no longer than six characters in length. New symbols must not duplicate existing approved gene symbols in either the human (Genew: Human Gene Nomenclature Database) or the mouse (MGD) databases.

a) The initial character of the symbol should always be a letter. Subsequent characters may be other letters, or if necessary, Arabic numerals.
b) All characters of the symbol should be written on the same line; no superscripts or subscripts may be used.
c) No Roman numerals may be used. Roman numbers in previously used symbols should be changed to their Arabic equivalents.
d) Greek letters are not used in gene symbols. All Greek letters should be changed to letters in the Latin alphabet (Appendix 2, Table 1). Note that such gene symbols will then appear in lists in the order of the Latin alphabet.
e) A Greek letter prefixing a gene name must be changed to its Latin alphabet equivalent and placed at the end of the gene symbol. This permits alphabetical ordering of the gene in listings with similar properties, such as substrate specificities, e.g., *GLA* "galactosidase, α"; *GLB* "galactosidase, β."
f) No punctuation may be used, with the exception of the HLA, immunoglobulin, and T-cell receptor gene symbols (which may be hyphenated). The HLA symbols are assigned by the WHO Nomenclature Committee for Factors of the HLA System [9] via the IMGT/HLA database [10]. The immunoglobulin and T-cell receptor gene symbols are assigned by the IMGT Nomenclature Committee via the IMGT/LIGM database [11].
g) Gene symbols will not usually be assigned to alternative transcripts. However, if a community working on a group of genes has a need for nomenclature where there are multiple small coding sequences which can be combined to form a number of different larger products, then these coding sequences may be assigned symbols, e.g., *UGT1A1* "UDP glycosyltransferase 1 family, polypeptide A1" to *UGT1A13* representing 13 distinct gene symbols.
h) Tissue specificity or molecular weight should be avoided; where necessary this may be included in the gene name.
i) Some letters or combination of letters are used as prefixes or suffixes in a symbol to give a specific meaning and their use for other meanings should be avoided (Section 10).
j) Oncogenes are given symbols corresponding to the homologous retroviral oncogene, but without the "v-" or "c-" prefixes, e.g., *JUN* "v-jun sarcoma virus 17 oncogene homolog (avian)," and *SRC* "v-src sarcoma (Schmidt-Ruppin A-2) viral oncogene homolog (avian)."

3. GENE NAMES

Gene names should be brief and specific and should convey the character or function of the gene, but should not attempt to describe everything known about it. The first letter of the symbol should be the same as that of the name in order to facilitate alphabetical listing and grouping. Gene names are written using American spelling. Tissue specificity and molecular weight designations should be avoided as they have only limited use as a description and may in time and across species prove inaccurate; however, they may be incorporated into the gene name if absolutely necessary. Gene names should not include terms such as nephew, cousin, sister, etc. to describe familial relationships with other genes. The following gene name syntax

should be used:

a) Names start with a lowercase letter unless it is a person's name describing a disease/phenotype, e.g., *AHDS* "Allan-Herndon-Dudley syndrome."
b) Descriptive modifiers should follow the main part of the name, separated by commas, e.g., *ACO1* "aconitase 1, soluble."
c) Where a complete alternative name (or names) is being included as part of the name, this should be in parentheses, e.g., *IDS* "iduronate 2-sulfatase (Hunter syndrome)."
d) Names of other species must be placed in parentheses at the end, e.g., *LFNG* "lunatic fringe homolog (*Drosophila*)" and *ANLN* "anillin, actin binding protein (scraps homolog, *Drosophila*)."

4. Gene Families

Hierarchical symbols for both structural and functional gene families should be used where possible. A stem (or root) symbol as a basis for a symbol series allows easy identification of other family members in both database searches and the literature. Gene family members should be designated by Arabic numerals placed immediately after the gene stem symbol, without any space between the letters and numbers used, e.g., *GPR1*, *GPR2*, *GPR3* (three G protein-coupled receptor genes). However, very occasionally, only if they exist historically, single-letter suffixes may be used to designate these different genes, e.g., *LDHA*, *LDHB*, *LDHC* (three lactate dehydrogenase genes). Some gene families are very large and so may include a variety of number/letter combinations to indicate their phylogenetic relationships, e.g., *CYP1A1*, *CYP21A2*, *CYP51A1* (three members of the cytochrome P450 superfamily). When symbols are approved by the HGNC consecutively in a gene family the hierarchical order of symbols will not necessarily reflect the chronological order of peer-reviewed publications. Consecutive symbols approved by HGNC take precedence over those published, although this will be a matter for discussion with the relevant scientific community.

Many genes receive approved symbols and names that are non-ideal when considered in the light of subsequent information. In the case of individual genes a change to the name (and subsequently the symbol) is only made if the original name is seriously misleading. However, groups of scientists working on particular gene families often coordinate a revised nomenclature when more information becomes available; such initiatives are welcomed. Groups planning to do this are strongly advised to liaise with the HGNC in order to avoid possible problems. Previously approved symbols that have been withdrawn should not be re-used where possible as this causes great confusion in information retrieval.

4.1 Anonymous families

When a series of genes can be shown to be related by sequence similarity but otherwise cannot be described by homology or function they can be assigned anonymous and temporary FAM# symbols. Membership in each family, indicated by the FAM number, will be assigned using the general criteria established by Mackenzie *et al.* [12] and Nelson *et al.* [13] that the gene products show greater than 40% amino acid identity. However, these criteria are not exclusive but merely a convenient starting point on which to base an appropriate symbol. The symbol will always include a letter designation for subfamily, with the final character indicating gene number, e.g., *FAM7A1* "family with sequence similarity 7, member A1," etc. Once further information is available to assign a more descriptive name the whole gene family should be updated with new symbols, with the FAM designations being withdrawn.

5. Homologies with Other Species

Homologous genes in different vertebrate species (orthologs) should, where possible, have the same gene nomenclature.

a) Human homologs of genes first identified in other species should not be designated by a symbol beginning with H or h for human.
b) When a gene or series of genes has been defined in one species, and it is reasonable to expect that in the future a homologous gene will be identified in human, we recommend that the designated symbol be reserved for the human locus. We recommend that this should be done in other species for genes first identified in human.
c) When necessary to distinguish the species of origin for homologous genes with the same gene symbol, the letter-based code for different species already established by SWISS-PROT should be used. This can be found at URL http://www.expasy.ch/cgi-bin/speclist and commonly used species are shown in Appendix 2, Table 2.
d) The agreement between human and mouse gene nomenclature for many homologous genes should be continued and extended to other vertebrate species where possible.
e) Human homologs of genes in invertebrate or prokaryote species may sometimes be represented by the symbol used in the other species, usually followed by an L to represent "like," and a number if there is (or is likely to be) more than one human homolog; the original organism is placed at the end of the gene name in parentheses, e.g., *BARHL1* "BarH-like 1 (*Drosophila*)." The use of H to represent homolog is no longer recommended, and will be discontinued.

6. Genes Identified from Sequence Information

6.1 Antisense

A gene of unknown function, encoded at the same genomic

locus (with overlapping exons) as another gene should have its own symbol. If the new gene regulates the first gene it may be assigned the symbol of the first gene with the suffix AS for antisense, e.g., *IGF2AS* "insulin-like growth factor 2, antisense." The gene symbol should not be written backwards.

6.2 Opposite strand
Genes of unknown function on the opposite strand that have no proven regulatory function should be assigned the suffix OS for "opposite strand."

6.3 Untranslated functional RNAs
These may be assigned symbols that are unique and relevant to the scientific community. However, the approved name should contain "untranslated RNA," e.g., *H19* "H19, imprinted maternally expressed untranslated RNA."

6.4 Related (-like) sequences
The designation of the suffix "L" for like has been used in the past for related sequences identified by cross-hybridization studies. Where genes are identified by database searching and where no other functional information is available, and there is some sequence similarity with a known gene, they are designated with the symbol of the known gene followed by an "L" for like, e.g., *ACY1L* "aminoacylase 1-like." Alternatively, if there are a number of similar genes they may be assigned numbers in a series, e.g., *BTNL1* "butyrophilin-like 1" to *BTNL3*.

6.5 Genes of unknown function
Genes predicted from EST clusters or from genomic sequence with EST evidence, but showing no structural or functional homology, are regarded as putative. These are designated by the chromosome of origin, the letters "orf" for open reading frame and a number in a series, e.g., *C2orf1* "chromosome 2 open reading frame 1." The use of the lowercase letters "orf" is to prevent confusion between the first letter "o" and the numeral "0" (zero), which may be part of the chromosome number.

6.6 Pseudogenes
Pseudogenes are sequences that are generally untranscribed and untranslated and which have high homology to identified genes. However, it has recently been shown that in different organisms or tissues functional activation may occur. Therefore, the previous policy of assigning the gene symbol of the structural gene followed by "P" and a number will only be approved on a case-by-case basis. In the future, pseudogenes will usually be assigned the next number in the relevant symbol series, suffixed by a "P" for pseudogene (or "PS" in specific cases) if requested, e.g., *OR5B12P* "olfactory receptor, family 5, subfamily B, member 12 pseudogene." However, the designation "pseudogene" will remain in the gene name.

7. Enzymes and proteins

The rules described in the sections on gene names and symbols apply, but in addition the names of genes coding for enzymes should be based on those recommended by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology, e.g., *FPGS* "folylpolyglutamate synthase." These can be found at URL http://ca.expasy.org/enzyme/. Names of genes encoding plasma proteins, hemoglobins, and specialized proteins are based on standard names and those recommended by their respective committees, e.g., *HBA1* "hemoglobin, α 1."

8. Clinical disorders

The first gene symbol allocated to an inherited clinical phenotype (monogenic mendelian inheritance) may be based on an acronym which has been established as a name for the disorder, while following the rules described previously, e.g., *ACH* for "achondroplasia." It is usual for this symbol to change when the gene product or function is identified; however, if there is no additional information derived from the cloned gene, the disease symbol, e.g., *ACH*, will be maintained. If an approved gene symbol for the cloned gene based on product or function already exists, this will take precedence over the symbol derived from the clinical disorder when the gene descriptions are merged. For example, in the case of achondroplasia the symbol *ACH* has been withdrawn and *FGFR3* "fibroblast growth factor receptor 3 (achondroplasia, thanatophoric dwarfism)" approved in its place. If the original symbol is to be maintained we do not recommend the sole use of invented names based upon the phenotype, e.g., tuberin for *TSC2* "tuberous sclerosis 2."

8.1 Complex/polygenic traits
Genome searches may suggest a contributing locus in a complex trait, which may for convenience be given a gene symbol, e.g., *IDDM3* "insulin-dependent diabetes mellitus 3," although a proportion of these will disappear in time.

8.2 Contiguous gene syndromes
Syndromes clearly associated with multiple widely dispersed loci should not be given gene symbols. Syndromes associated with a regional deletion or duplication may be assigned the letters CR (for chromosome region), in place of S for syndrome. Examples include *ANCR* "Angelman syndrome chromosome region" and *CECR* "cat eye syndrome chromosome region." However, as advances in database design have increased the possible ways of representing this type of information, we recommend that such symbols are classified as syndromic region symbols and not gene symbols. Candidate genes found within a chromosomal region may be assigned a symbol based on the region symbol with sequential numbering, e.g., *CECR1* "cat eye syndrome chromosome region, candidate 1."

doi:10.1006/geno.2002.6748, available online at http://www.idealibrary.com on IDEAL

8.3 Loss of heterozygosity
A chromosomal region in which the existence of genes may be inferred by loss of heterozygosity can be designated by a symbol consisting of the letters LOH for loss of heterozygosity, the chromosome number, CR (for chromosomal region), and then a sequential number, e.g., *LOH1CR1* "loss of heterozygosity, 1, chromosomal region 1."

9. Genomic rearrangements and features

9.1 Genes only found within subsets of the population
Symbols will be assigned to genes generated by recombination events. The symbols will be designated a number or letter in the series determined from the original gene symbol, e.g., the *KIR2DL5* gene is duplicated in some haplotypes, the original gene will be referred to as *KIR2DL5A* "killer cell immunoglobulin-like receptor, two domains, long cytoplasmic tail, 5A" and the duplication as *KIR2DL5B* "killer cell immunoglobulin-like receptor, two domains, long cytoplasmic tail, 5B." Information detailing position on the chromosome, e.g., telomeric. may be added to the gene name to aid understanding. These types of symbols will be agreed on an individual basis.

9.2 Gene recombination
When there has been recombination between genes, generating a new locus, the gene symbol should remain short but reflect the two (or more) initial gene symbols. The name may be as descriptive as necessary, e.g., *POMZP3* "POM (POM121 homolog, rat) and ZP3 fusion."

9.3 Functional repeat sequences
Sequence data for these records must be obtained by HGNC before symbol approval. Endogenous retroviruses have the stem symbol ERV, followed by a letter representing type and the next consecutive number in the series, e.g., *ERVK2* "endogenous retroviral sequence K(C4), 2." Transposable elements have the stem symbol TE, followed by a relevant character set, e.g., MAR, and the next consecutive number in the series. Thus, a mariner-type transposable element could be *TEMAR1* "transposable element mariner-type, 1." LINEs (long interspersed nuclear elements) are currently assigned the next consecutive number in the LRE series, e.g., *LRE1* "LINE retrotransposable element 1."

10. Characters reserved for specific usage

Letters, or combinations of letters, are used as prefixes or suffixes in a symbol to represent a specific meaning and so their use for other meanings should be avoided where possible (Table 1). The following letter usage is not acceptable: H or h for human, G or g for gene. If the name of a gene contains a character or property for which there is a recognized abbreviation, the abbreviation should be used, e.g., the single-letter abbreviation for amino acids used in aminoacyl residues (Appendix 2, Table 3). Other characters are also used within the symbol to represent genomic or genetic information (Table 1).

TABLE 1: Characters reserved for specific usage

Character	Meaning
AS	antisense
AP	associated/accessory protein
BP	binding protein
C	catalytic
CR	chromosome/critical region
C#orf#	chromosome # open reading frame #
D	domain containing
FAM	family with sequence similarity
N or NH	inhibitor
IP	interacting protein
IT	intronic transcript
LG	ligand
L	like
LOH	loss of heterozygosity
MT or M	mitochondrial
OS	opposite strand
OT	overlapping transcript
P or PS	pseudogene
Q	quantitative trait
R	receptor
RG	regulator
#	gene family
@	gene cluster

11. Symbol status

Each gene record in the Genew database has one of five states:

a) Pending: gene symbols in progress.
b) Approved: official gene symbols that are publicly available.
c) Reserved: official gene symbols that are not publicly available; these will be released after a maximum period of six months.
d) Symbol Withdrawn: previous symbols for genes that now have different approved symbols.
e) Entry Withdrawn: symbols that refer to a gene that has since been shown not to exist.

APPENDIX 1: GENE SYMBOL USE IN PUBLICATIONS

It is recommended that gene and allele symbols are underlined in manuscript and italicized in print; protein symbols should be represented in standard fonts. Italics need not be used in gene catalogs. To distinguish between mRNA, genomic DNA, and cDNA the relevant prefix should be written in parentheses: (mRNA)*RBP1*, (gDNA)*RBP1*, (cDNA)*RBP1*.

Appendix 1.1 DNA segments

These symbols can be obtained from The Genome Database (GDB) and are assigned automatically to arbitrary DNA fragments and loci. Please email requests to help@gdb.org. These symbols comprise five parts, e.g., *DXS9879E*, described by the following guidelines:

a) D for DNA.
b) 0, 1, 2,...22, X, Y, XY for the chromosomal assignment, where XY is for segments homologous on the X and Y chromosomes, and 0 is for unknown chromosomal assignment.
c) S, Z or F indicating the complexity of the DNA segment detected by the probe; with S for a unique DNA segment, Z for repetitive DNA segments found at a single chromosome site, and F for small, undefined families of homologous sequences found on multiple chromosomes.
d) 1, 2, 3,... a sequential number to give uniqueness to the above-concatenated characters.
e) When the DNA segment is known to be an expressed sequence the suffix E can be added to indicate this.

Appendix 1.2 Alleles

Allele terminology is the responsibility of the HUGO Mutation Database Initiative (now the Human Genome Variation Society; http://ariel.ucs.unimelb.edu.au/~cotton/mdi.htm). However, alleles should still be limited to an optimum of three characters using only capital letters or Arabic numerals. The allele designation should be written on the same line as gene symbol separated by an asterisk, e.g., *PGM1*1*, the allele is printed as **1*. When only allele symbols are displayed an asterisk should precede them. Sequence variation nomenclature, in the form of mutations and polymorphisms, is described by Antonarakis *et al.* [14] and den Dunnen and Antonarakis [15]. HLA alleles are assigned by the WHO Nomenclature Committee for Factors of the HLA System [9] via the IMGT/HLA database [10]. Immunoglobulin and T cell receptor alleles are assigned by the IMGT Nomenclature Committee via the IMGT/LIGM database [11].

Appendix 1.3 Splice variants

The HGNC has no authority over protein nomenclature; however, we are frequently asked how to designate splice variants so we suggest the following:

Proteins should be designated using the same symbol as the gene, printed in nonitalicized letters. When referring to splice variants, the symbol can be followed by an underscore and the lowercase letter "v," then a consecutive number to denote which variant is which, e.g., G6PD_v1.

Appendix 1.4 Species designations

When necessary to distinguish the species of origin for homologous genes with the same gene symbol, the letter-based code for different species already established by SWISS-PROT should be used. This can be found at URL http://www.expasy.ch/cgi-bin/speclist and commonly used species are shown in Appendix 2, Table 2. The code is for use in publications only and should not be incorporated as part of the gene symbol. The species designation is added as a prefix to the gene symbol in parentheses. For example HUMAN signifies *Homo sapiens* and MOUSE signifies *Mus musculus*, e.g., human genes: (HUMAN)*ABCA1*; (HUMAN)*SLC13A2*; homologous mouse genes: (MOUSE)*Abca1*; (MOUSE)*Slc13a2*.

APPENDIX 2: TABLES

TABLE 1: Greek-to-Latin alphabet conversion

Greek symbol	Greek	Latin upper case conversion
α	alpha	A
β	beta	B
γ	gamma	G
δ	delta	D
ϵ	epsilon	E
ζ	zeta	Z
η	eta	H
θ	theta	Q
ι	iota	I
κ	kappa	K
λ	lambda	L
μ	mu	M
ν	nu	N
ζ	xi	X
ο	omicron	O
π	pi	P
ρ	rho	R
σ	sigma	S
τ	tau	T
ϑ	upsilon	Y
φ	phi	F
χ	chi	C
φ	psi	U
ω	omega	W

doi:10.1006/geno.2002.6748, available online at http://www.idealibrary.com on IDEAL

TABLE 2: Common species abbreviations from URL http://www.expasy.ch/cgi-bin/speclist

Common name	Abbreviation	Species
cat	FELSI	*Felis silvestris catus*
chicken	CHICK	*Gallus gallus*
chimp	PANTR	*Pan troglodytes*
dog	CANFA	*Canis familiaris*
fruit fly	DROME	*Drosophila melanogaster*
human	HUMAN	*Homo sapiens*
mouse	MOUSE	*Mus musculus*
orangutan	PONPY	*Pongo pygmaeus*
pig	PIG	*Sus scrofa*
puffer fish	FUGRU	*Fugu rubripes*
puffer fish	TETFL	*Tetraodon fluviatilis*
rabbit	RABIT	*Oryctolagus cuniculus*
rat	RAT	*Rattus norvegicus*
worm	CAEEL	*Caenorhabditis elegans*
yeast	YEAST	*Saccharomyces cerevisiae*
yeast	SCHPO	*Schizosaccharomyces pombe*
zebrafish	BRARE	*Brachydanio rerio*

TABLE 3: Single-letter amino acid symbols

Amino acid	Three-letter symbol	One-letter symbol
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Asn +Asp	Asx	B
Cysteine	Cys	C
Glutamine	Gln	Q
Glutamic acid	Glu	E
Gln + Glu	Glx	Z
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

ACKNOWLEDGMENTS

Many thanks to Stylianos Antonarakis, Sergolene Ayme, Amos Bairoch, Richard Cammack, Panos Deloukas, Ian Dunham, Cara Hunsberger, Ian Jackson, Peter Little, Donna Maglott, Gerard Manning, Dan Nebert, Mark Paalman, Chris Porter, Kim Pruitt, Yoshiyuki Sakaki, Stephen Scherer, Alan Scott, Tom Shows, Cindy Smith, Nigel Spurr, Keith Tipton, and anyone else who contributed with feedback. The HGNC is supported by NIH contract N01-LM-9-3533 and by the UK Medical Research Council.

REFERENCES

1. Shows, T. B., *et al.* (1979). International System for Human Gene Nomenclature. *Cytogenet. Cell Genet.* **25:** 96–116.
2. Shows, T. B., *et al.* (1987). Guidelines for human gene nomenclature: An international system for human gene nomenclature. *Cytogenet. Cell Genet.* **46:** 11–28.
3. McAlpine, P. J. (1995). International system for human gene nomenclature. *Trends Genet.* **11:** 39–42.
4. White, J. A., *et al.* (1997). Guidelines for human gene nomenclature. HUGO Nomenclature Committee. *Genomics* **45:** 468–471.
5. Venter, C., *et al.* (2001). The sequence of the human genome. *Science* **291:** 1304–1351.
6. Lander, E. S., *et al.* (2001). The International Human Genome Sequencing Consortium: Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.
7. Wain, H. M., Lush, M., Ducluzeau, F., and Povey, S. (2002). Genew: The Human Nomenclature Database. *Nucleic Acids Res.* **30:** 169–171.
8. White, J., *et al.* (1998). Networking nomenclature. *Nat. Genet.* **18:** 209.
9. Marsh, S. G. E., *et al.* (2001). Nomenclature for factors of the HLA System, 2000. *Tissue Antigens* **57:** 236–283.
10. Robinson, J., Waller, M. J., Parham, P., Bodmer, J. G., and Marsh, S. G. E. (2001). IMGT/HLA Database—a sequence database for the human major histocompatibility complex. *Nucleic Acids Res.* **29:** 210–213.
11. Lefranc, M. P. (2001). IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res.* **29:** 207–209.
12. Mackenzie, P. I., *et al.* (1997). The UDP glycosyltransferase gene superfamily: recommended nomenclature update based on evolutionary divergence. *Pharmacogenetics* **7:** 255–269.
13. Nelson, D. R., *et al.* (1996). P450 superfamily: update on new sequences, gene mapping, accession numbers and nomenclature. *Pharmacogenetics* **6:** 1–42.
14. Antonarakis, S. E. (1998). Recommendations for a nomenclature system for human gene mutations. Nomenclature Working Group. *Hum. Mutat.* **11:** 1–3.
15. den Dunnen, J. T., and Antonarakis, S. E. (2000). Mutation nomenclature extensions and suggestions to describe complex mutations: A discussion. *Hum. Mutat.* **15:** 7–12.