MEETING REPORT

A Report on the International Nomenclature Workshop Held May 1997 at The Jackson Laboratory, Bar Harbor, Maine, U.S.A.

Judith A. Blake,*[,1] Muriel T. Davisson,*
Janan T. Eppig,* Lois J. Maltais,*
Sue Povey,† Julia A. White,†
and James E. Womack‡

*The Jackson Laboratory, Bar Harbor, Maine 04609; †University College London, MRC Human Biochemical Genetics Unit, London NW12HE, United Kingdom; and ‡Texas A&M University, College of Veterinary Medicine, College Station, Texas 77843

DEDICATED TO THE TIRELESS EFFORTS OF PHYLLIS J. MCALPINE FOR HUMAN GENE NOMENCLATURE

> The beginning of wisdom is calling things by their right name.
> old Chinese proverb

Estimates for the number of human genes range from 60,000 to 100,000 (Antequera and Bird, 1993; Fields *et al.*, 1994). Many of the other mammalian genomes under intense scrutiny appear to be roughly the same size as human (whether they have the same number of protein-coding sequences, or genes, is open to debate). Nevertheless, the expected numbers are well beyond the current 8010 genes named in the mouse or the 6651 genes named in humans as of July, 1997. How to distinguish each, in human terms, is no trivial task because many of these genes share structural elements or functional attributes with each other. Added to this problem is the historical pattern of embedding in the names of the genes something about their function, relationship to other genes, expression patterns, chromosome location . . . and more. In addition, the same name for the same gene in different species has been intensely desired by many researchers to facilitate their work. So the effort to coordinate gene names for homologs between species continues endlessly as the gene names constantly change when further research reveals new details about them. Finally, scientists often publish using their own designation of the appropriate name for the genetic unit under discussion; thus the very task of gene identification is compounded by the understandable desire of researchers to name the gene they have discovered. The result can be a confusion of names for the same object, confusing efforts to make associations between the biology of similar objects.

In recognition of the importance of the gene nomenclature process, 53 people gathered at The Jackson Laboratory in Bar Harbor, Maine, from April 30 to May 3, 1997, for the International Nomenclature Workshop. Attendees included participants from many mammalian genome mapping efforts as well as representatives from model organism databases and from the major sequence databases (Table 1; also see for acronym legends). This eclectic mix brought together nomenclature coordinators for many of the organism databases, large database managers, computer programmers, and gene mappers from large human sequencing centers to small research efforts centered on marsupials. The result was 3 days of dynamic discussion and presentations on the nomenclature process and problems in symbolizing genes in organisms as divergent as humans and yeast.

Participants did not expect to solve the challenges of gene nomenclature in one gathering; the more modest goals of the meeting were to meet other members of nomenclature committees for a variety of organisms, to promote more coordination between databases, and to further the development and use in different species of similar concepts in the nomenclature process. The meeting itself generated several new collaborations and the creation of working groups across species. In addition, a registry of gene names and symbols has been established (see below).

Background

During the 1970s, there were regular mammalian comparative mapping committee meetings held annually as part of the Human Genome Mapping Workshops. Starting with the Human Mapping meeting in 1973 (New Haven Conference, 1973), these chromosome-based mapping meetings provided the venue for sorting out nomenclature issues among the genome mapping communities before the advent of sequence technologies and databases. The last comparative mapping meeting of this series, which included species representation beyond mammals, was held in Australia in 1995. Since then, the discovery of new genes and the interest in comparative mapping of genomes have only intensified.

Following the hiatus in comparative mapping meetings, two groups of users of nomenclature procedures and connections emerged. First, there are the researchers of mammalian genomes such as cattle, sheep, pig, and chicken who are largely mapping known genes and naming them on the basis of names and symbols already assigned for mouse or human genes. On the surface, this is a straightforward endeavor. However, the knowledge about human and mouse genes is expanding rapidly, and the names and symbols for the human and mouse genes change with increasing frequency to reflect the new knowledge. This has led to concern and confusion as the other mammalian genome efforts, with far less support, have struggled to accommodate the changes.

The second group of researchers affected by the nomenclature conundrum has been the annotators of the huge and burgeoning volumes of sequence data for mouse and human. These scientists are looking for assistance in the naming process . . . even as little is known about the function of most of the genes being discovered. So, for example, what symbols

[1] To whom correspondence should be addressed. E-mail: jblake@informatics.jax.org.

0888-7543/97 $25.00

TABLE 1

Primary Databases and Their URLs

Acronym or genome	Name of database	URL
AGIS	Agricultural Genome Information System	http://probe.nalusda.gov:8000/index.html
	Animal Comparative Mapping	http://www.ri.bbsrc.ac.uk/genome_mapping.html
Arabidopsis	*Arabidopsis thaliana* Database	http://genome-www.stanford.edu/
BovGBASE	Bovine Genome Database	http://bos.cvm.lamu.edu/bovgbase.html
ChickGBASE	Chick Database	http://www.ri.bbsrc.ac.uk/chickmap/chickgbase.html
FlyBase	*Drosophila* Genome Database	http://flybase.blo.indiana.edu
GDB	The Genome Database	http://gdbwww.gdb.org
MGI	Mouse Genome Informatics (GXD and MGD)	http://informatics.jax.org
	Mouse Nomenclature Submission Form	http://www.informatics.jax.org/nomen/
PiGBASE	Swine database	http://www.ri.bbsrc.ac.uk/pigmap/pigbase/pigbase.html
RatMap	Rat Genome Database	http://ratmap.gen.gu.se/
SheepBase	New Zealand Sheep Genome Program	http://dirk.invermay.cri.nz/
Yeast	*Saccharomyces* Genome Database	http://genome-www.stanford.edu/
Zebrafish	The Fish Net	http://zfish.uoregon.edu
ATCC	American Type Culture Collection	http://www.atcc.org
GenBank	National Center for Biotechnology Information—GenBank	http://www.nobi.nlm.nih.gov/Web/Genbank/index.html
GSDB	Genome Sequence DataBase	http://www.nogr.org/gsdb/
Mendel	Sequenced Plant Genes	http://probe.nalusda.gov:8300/cgi-bin/browse/mendel
NCGR	National Center for Genome Resources	http://www.nogr.org/gsdb/
P450	P450 database	http://www.icgeb.triesle.it/p450
PIR	Protein Information Resource	http://www-NBRF.Georgetown.edu/pir/
PROW	Protein Reviews On the Web	http://www.nobl.nlm.nih.gov.PROW
Swiss-Prot	Annotated Protein Sequence Database	http://expasy.hcuge.ch
Library of Congress-Cataloging Services		http://lcweb.loc.gov/catdir/pcc/poc.html
MeSH	Medical Subject Headings, National Library of Medicine	http://www.nlm.nih.gov/mesh

should be applied to the 64 new genes identified during the sequencing of a human BAC clone for which the only knowledge is that they all contain zinc-finger domains? And what is the relationship between all the sequences that are named as a "myosin"?

Of course, all of this could be theoretically solved by assigning sequential numbers to each unique gene as we identify these unique entities within each genome. This is not a humanly useful approach since we human beings use words and symbols to reference information and to simplify communication. And even if numbers were uniquely assigned within a genome, they would need to be cross-referenced to their homologs in other genomes, the most interesting part of our endeavor to understand the evolution of life. The effort to name genes, and even the strategy for doing so, has been further challenged by the development of high-throughput sequencing techniques and electronic collection of sequence data concurrent with multiple representations of these data in biological databases worldwide. The task has become overwhelming as interest and recognition of the importance of nomenclature continue to grow.

Synopsis of Sessions

The meeting started with an overview of several organism-specific databases and nomenclature guidelines. Several species and sequence database representatives described their operations and mechanisms for handling nomenclature or nomenclature needs. Lois Maltais and Julia White, the nomenclature coordinators for the mouse (MGD) and human (GDB) databases, respectively, spoke in detail about their efforts to coordinate assignments of symbols for mouse and human genes. The nomenclature committees for the mouse and human databases have worked together for many years to share gene symbol stem use and to clarify nomenclature for related and homologous genes. In general, other mammalian catalogs have followed the lead of the human and mouse community in naming their genes. However, recently the volume has risen to over 2800 new or reassigned gene symbols per year, and that number will only increase over the next few years. In addition to the high number of new genes being described, many older gene symbols for mutant phenotypes are modified as the genes responsible for the phenotypes are identified and cloned. Frequently, multiple systems are in use simultaneously for naming related members of the same gene family. Presentations included the status and process for naming genes in nonhuman mammalian genomes, cattle, sheep, pig, chicken, cat and others, as well as in some model organisms such as *Drosophila*, zebrafish, plants, and yeast.

Several speakers discussed theoretical considerations and knowledge concerning nomenclature issues from other intellectual endeavors. Norbert Weber (National Library of Medicine, NLM), explained the process of building the MESH system, both as a classification hierarchy and as the process of choosing keywords for searching. John Mitchell (Library of Congress) spoke from the perspective of the Program for Cooperative Cataloging about the development of consistent authority files and multiple thesauri in the handling of information systems. Stan Blum (Bishop Museum, Hawaii) discussed the 300-year history of naming species using the binomial system and the struggle for consistency and accuracy as the representation of species changed. He argued for the importance of defining the underlying concept of the object (e.g., gene) to which a name is being attached and then tracking the change of that concept over time.

In one of the most thought-provoking talks for many participants, Carl Price (Rutgers) described the system of the Commission Plant Gene Nomenclature (CPGN) system. Angiosperms, with the ability and propensity for duplicating whole genomes, have widely differing numbers of members in different gene families between species. In the CPGN system, developed for well-known groups of genes, each gene group is given a stem symbol designation. Then within each species, different members of the gene family are numbered sequentially. This leads to great stability of nomenclature within each species and the ability to compare members of the same gene group or family between species without the constraint of identifying strict orthologs. This system is practical for "known" genes and works well for the plant molecular geneticists who work cooperatively across genomes. As a nomenclature system, it is not so robust for the volume of uncharacterized genes being isolated from the human genome. Still, for comparative mappers, it presented a good alternative approach to naming genes individually that many people were interested in pursuing.

In a related system for naming members of a gene family, Dan Nebert (University of Cincinnati) described the stem system used for symbolizing human genes in large gene families, giving the cytochrome P450 family (CYP) as an example. Somewhat similar to the plant system, a symbol stem followed by a number assignment provides the symbol and a means of expanding members of the groups as related genes are discovered. Scientists expert in a gene family group are called upon for review of that group on a regular basis. Several other community-based committees for specific gene families have been formed over the years and have been tremendously helpful to nomenclature coordinators faced with trying to distinguish and name members of large gene families. Most such user committees, however, function best after a family is fairly well established. This approach is harder to employ during times of rapid gene discovery.

The speakers representing biological databases reflected common interests in linking across databases and in reflecting a common nomenclature for genes. Ultimately, the database managers envision increased computer-generated annotation for relating, through shared accession numbers, such information as protein structure or related sequences or information from organism-specific databases. Improved automatic annotation is aided by standardized nomenclature, particularly in that no two different objects (genes) share the same name. So for the database managers, anything to help resolve the nomenclature conundrums that currently exist would be welcome. The ideal would be a database complete within its defined scope that held minimal redundancy, was current, was interconnected to value-added resources, and was accessible for searching through robust controlled vocabularies.

One consensus of the meeting was the need for participation of the user community in keeping databases current. Carol Harger (GSDB) described how the community annotation system is working at GSDB, and Robert Cottingham (GDB) reviewed the development and function of the open annotation system at GDB.

The final session included presentations by Elbert Branscom (LLNL) and Rolf Apweiler (SWISS-PROT) on future database development and then a lengthy discussion of issues raised during the Workshop and possible solutions. Among the comments were the observations that we are "blessed in explicit ignorance" as we name genes as "putative transcript factors"; meaningful naming is not possible in many instances, and so we are released from that standard of nomenclature and can institute more practical approaches for the current research situation. We need to build a "nonprejudicial" database with synthetic, dynamic representations for all the new genes. Text storage is not a useful medium, and there is a huge difference between data storage and useful storage; i.e., we need to be able to extract answers to biological questions from the data. The challenge here is to generate improved machine-searchable storage procedures, and this can be done through the construction of controlled vocabularies. However, Rolf Apweiler highlighted the danger in using systems of gene nomenclature as functional or other classification systems: if the classification changes, as it will, the nomenclature of genes must be revised immediately. Thus, there is a pressing requirement to separate nomenclature from the requirement to retrieve information from databases based on function, evolutionary relationship, or information on time and tissue of expression.

The final component of this discussion was the immediate need for representation of comparative genomic sequence data. The scientist of today wants and needs to be able to electronically see sequence and mapping information from different genomes and to move seamlessly between these information types. Even as we await the imminent sequencing of the human genome, we recognize that the conserved elements, the exons and regulatory elements that hold the greatest biological interest, constitute perhaps only 3% of the genome, and we wish to know about this small portion of the genome for many species. In our search for understanding of the evolution of gene sequences, we seek also to understand the evolution of protein function and thus the development and maintenance of life.

Discussion of Emerging Issues

General Nomenclature Concepts

Recommendation: Encourage stability in nomenclature.

During the final summary session, participants agreed in general that investigators should be able to name and symbolize genes as they want as long as they do not use an already assigned symbol and that nomenclature is more important within a species than across species. While scientists should be encouraged to name new genes as they wish, they should resist renaming them until enough is known about the gene and related genes for a review of the gene group by scientists as community experts. Overall, the effort should be to assign unique symbols to each "genetic unit" within a species and to support stability of nomenclature terms as much as possible. The concept used by the plant gene nomenclature community of assigning a gene family stem name followed by a numbering system for the genes within a species was viewed favorably by many. The stability of gene names and symbols is more important for the immediate unique identification than is trying to assign "meaningful" names and symbols based on insufficient knowledge about a gene. Participants agreed on four general concepts regarding coordinating across species for homologous genes.

(1) The naming effort for genes within each species should continue to be toward an internally consistent set of gene names. Thus, the nomenclature for genes for a given species would be a unique set of symbols and names for the genes of that organism. These should be developed by those working with the species. This is already a prevalent practice for most species, particularly for the naming of genes based on their phenotypic effect.

(2) It may be too much to expect identical symbols for homologs across species. Database structures can handle this for us by creating displays that indicate homology relationships. We recognize that although the display is easy to create, identifying the relationship requires substantial effort from scientists as database editors or community curators to identify which genes are true homologs (orthologs).

This said, within mammals many genes do have a 1:1 correspondence, and the coordination between mouse and human nomenclature committees has been very productive in coordinating symbols between these two species. This effort should and will continue.

(3) Short-term solutions are needed for assigning symbols quickly without using the same symbols for different genes in different species. The most confusion comes from false associations of information across species because of the incorrect assumption that genes with the same symbol/name in different species refer to the same entity. One concrete step toward promoting a quick symbol assignment without duplication is the implementation of an all-species gene registry.

(4) The long-term solution relies on the development of multiple classification systems and other components so that the power of electronic databases can be used to answer complex biological questions. Database managers need to rely on community experts for the classification and reorganization of such related groups of genes. The symbols can evolve slowly from a species-consistent name to a name common across species.

Registry of Gene Symbols and Names

Recommendation: Establish a single registry of gene symbols and names for as many species as possible.

By the end of the Workshop, there did seem to be, especially among the database people, a feeling that there needs to be an "authority" certifying gene names and symbols. Perhaps the mammalian groups could agree on a common nomenclature for human, mouse, rat, livestock species, and others. If such agreement was reached, the standard set of names and symbols would need to be based on that used for human genes, in large part as a result of the Human Genome Initiative that guarantees the generation (and naming) of tens of thousands of new gene sequences for the human genome within the next 5 years. To this standard would be added species-specific genes not present in humans.

Donna Maglott of ATCC discussed how ATCC might serve as a holder of a gene symbol and name registry beginning with human and mouse nomenclature information and extending to other species as databases develop to the point that electronic downloads can be easily accommodated. This registry would include synonyms as well. It would provide a single resource for researchers to check when they name and symbolize a new gene. The International Committee on Standardized Genetic Nomenclature for Mice and the HUGO Gene Nomenclature Committee, members of whom were present at this meeting, embraced the registry concept. The registry could expand to include other species . . . human, rat, mouse, cattle, and fly would be an auspicious start.

Subsequent to the Workshop ATCC agreed to establish a gene registry at their WWW site, and a prototype is being set up. The registry will be electronically populated from existing species databases and will include gene/marker symbol, gene/marker name, species, authority (database source), and identifier (database accession number). The prototype is expected to be publicly available by September 1, 1997. Additional species will be added to those in the prototype as they become electronically available from species databases.

Development and Interconnection of Databases

Recommendation: Work with biologists to develop multiple classifications and controlled vocabularies.

As noted by Elbert Branscom, the "real challenge of nomenclature is making it possible to access data in the databases." Good nomenclature doesn't solve all database problems, but can help databases function better and will increase database value. Standard nomenclature, following suggested guidelines, helps avoid problems that confound access to data or cause misrepresentations.

The interconnection of biological databases is well underway, and further links are being planned and implemented at an accelerated pace. The requirements of providing fully supported unique identifiers for the major objects in the database are generally understood, and most databases are actively incorporating unique accession numbers for each gene recorded.

Many speakers noted the need to implement controlled vocabularies rapidly in all the databases. Many of these thesauri will be used in common, for example gene product function. The data in the databases could in principle answer many more questions than current search capabilities permit, and search capabilities will be greatly enhanced by the utilization of controlled vocabularies for as many types of information as possible. Graham Cameron pointedly noted that the large sequence databases have failed to provide classification systems that would alleviate the pressure to encode functional information in gene names. Database managers are aware of and working on this challenge.

Impact of Comparative Genomics

Recommendation: Continue to hold workshops and meetings that address comparative mapping and genomics and that include scientists and database specialists representing a wide range of organisms. Many participants suggested that a regular series of International Nomenclature Workshops might be held, perhaps at 2- to 3-year intervals. Encourage the development of database structures and biological classification systems that will enhance the ability to use databases to help answer detailed biological questions.

Summary

Scientists, whether species gene mappers or biological database developers, are looking for consistency and stability in the naming of genomic segments. For mammalian genome groups, this means relaxing the insistence on constant name revisions, allowing the functional and location information to be encoded elsewhere in the databases through the development of improved classification systems and controlled vocabularies. For gene discoverers, it means cooperation to determine unique gene designations without imposing too much knowledge in the naming process. For both groups, the power of database connections and structures needs to be exploited further to facilitate use of the massive amount of biological data being encoded.

This workshop allowed the voices of many active participants in the naming and mapping of genes to be heard. The discussions reflected the intense interest in the stabilization of gene names despite the rapid advances in our knowledge of gene function and evolutionary relationships. Relaxation of the constraints and conventions on the gene naming effort that required the encapsulation of gene function or other attributes in the symbol will promote stability in the naming efforts. The development of multiple classifications and more robust database structures to support the information explosion releases the gene names from constantly changing to reflect our current knowledge. The gain in stability of gene nomenclature will be a welcome result.

ACKNOWLEDGMENTS

This workshop was supported in part by the National Science Foundation, the National Human Genome Research Institute at NIH, the U.S. Department of Agriculture, the U.S. Department of Energy, The Jackson Laboratory, and the Burroughs Wellcome Fund. Special thanks to Merlene McIntire, Sandra Wilcomb, and Robert Shea for excellent technical and logistical support of the Workshop.

REFERENCES

Antequera, F., and Bird, A. (1993). Number of CpG islands and genes in human and mouse. *Proc. Natl. Acad. Sci. USA* **90:** 11995–11999.

Fields, C., Adams, M. D., White, O., and Venter, J. C. (1994). How many genes in the human genome? *Nature Genet.* **7:** 345–346.

New Haven Conference (1973/1974). Refers to the “First International Workshop on Human Gene Mapping, Birth Defects: Original Article Series,” Vol. 10, No. 3, The National Foundation, New York; *Cytogenet. Cell Genet.* (1974). **13**(1–2): 1–216.

NOMENCLATURE

Guidelines for Human Gene Nomenclature (1997)

J. A. White,[1] P. J. McAlpine,[2] S. Antonarakis,[3] H. Cann,[4] J. T. Eppig,[5] K. Frazer,[5] J. Frezal,[6] D. Lancet,[7] J. Nahmias,[1] P. Pearson,[8] J. Peters,[9] A. Scott,[10] H. Scott,[11] N. Spurr,[12] C. Talbot Jr.,[13] and S. Povey[1,14]

[1]MRC Human Biochemical Genetics Unit, University College London, Wolfson House, 4 Stephenson Way, London, NW1 2HE, United Kingdom; [2]Department of Human Genetics, University of Manitoba, T250-770 Bannatyne Avenue, Winnipeg, Manitoba, Canada R3E OW3; [3]Division of Medical Genetics, University of Geneva School of Medicine, CMU, 9 Avenue de Champel, Geneva 1211-4, Switzerland; [4]CEPH, 27 rue Juliette Dodu, Paris 75010, France; [5]The Jackson Laboratory, 600 Main Street, Bar Harbor, Maine 04609; [6]GENATLAS, Service de Genetique Medicale Hopital des Enfants Malades, 149, rue de Sevres, 75743, Paris, Cedex 15, France; [7]Department of Membrane Research and Biophysics, The Weizmann Institute of Science, Rehovot, 76100, Israel; [8]Department of Human Genetics, Utrecht University, P.O. Box 80030, Utrecht 3508 TA, Netherlands; [9]Mammalian Genetics Unit, Medical Research Council, Harwell, Didcot, Oxon, OX11 0RD, United Kingdom; [10]OMIM, Center for Medical Genetics, Blalock 1007, The Johns Hopkins Hospital, 600 N. Wolfe Street, Baltimore, Maryland 21287-4922; [11]Division of Medical Genetics, University of Geneva Medical School, 1 rue Michel Servet, 1211 Geneva 4, Switzerland; [12]Smithkline Beecham Pharmaceuticals, New Frontiers Science Park (North), Third Avenue, Harlow, Essex CM19 5AW, United Kingdom; and [13]Genome Data Base (GDB), Johns Hopkins University, 2024 E. Monument Street, Baltimore, Maryland 21205

Received August 15, 1997; accepted August 18, 1997

Introduction

Thirty scientists, database experts, and representatives of scientific journals met on March 5, 1997 in Toronto, in association with HGM97, to discuss human gene nomenclature.The guidelines drawn up at this meeting were discussed further at the cross-species International Workshop on Gene nomen-

[14] To whom correspondence should be addressed.

0888-7543/97 $25.00

clature held in Bar Harbor in May 1997 (J. Blake, M. Davisson, J. Eppig, L. Maltais, S. Povey, J. White, and J. Womack, in preparation.) These now constitute the current guidelines approved by the HGMW Nomenclature sub-committee of HUGO and replace those previously published (Shows *et al.*, 1979; McAlpine, 1995). The use of these Guidelines should help to ensure that a gene symbol is unique and appropriate. However, for it to be listed as an approved gene symbol and accepted as such by journals requiring approved nomenclature it must be submitted to the HUGO Nomenclature Committee of the Genome Database, which will verify its compliance with the guidelines and suitability as a new symbol. Information submitted prior to publication is treated in confidence, and symbols may be reserved up to 6 months before manuscript submission. Information, guidance, and a submission form for approval of human gene symbols are to be found on the Nomenclature Committee WWW page at URL http://www.gene.ucl.ac.uk/nomenclature/.

Definition: A gene is a DNA segment that contributes to phenotype/function. In the absence of demonstrated function a gene may be characterized by sequence, transcription, or homology.

1. General Rules for Gene Nomenclature

1.1 Requirements for Designation by Gene Symbol

For a gene symbol to be allocated at least one of the following criteria must apply.

1.1.1. A gene symbol may be used to designate a clearly defined phenotype shown to be inherited as a monogenic Mendelian trait (Example: TSC1).

1.1.2. Gene symbols may be allocated to as yet unidentified genes contributing to a complex trait shown by linkage or association with a known marker (Example: IDDM6).

1.1.3. A gene symbol may be used to designate a cloned segment of DNA with sufficient structural, functional, and expression data to identify it as a transcribed entity. However, alternate transcripts from the same gene should not in general be given different gene symbols.

1.1.4. Gene symbols are also allocated to nonfunctional copies of genes (pseudogenes).

1.1.5. Genes encoded by the opposite (anti-sense) strand of a known gene will be given their own symbols.

1.1.6. A gene symbol may be given to a transcribed but untranslated DNA segment (Example: XIST).

1.1.7. A cellular phenotype from which the existence of a gene or genes can be inferred may have its own designation (Example: LOH#CR#).

1.1.8. If insufficient data are available to allocate a unique and meaningful gene symbol, a putative gene may be designated by the symbol C#orf#. This symbol will also be used for EST clusters. Other fragments of expressed sequence will be designated by a D-number.

1.2 Gene Symbols

1.2.1. Gene symbols are designated by uppercase Latin letters or by a combination of uppercase letters and Arabic numbers. Symbols should be short to be useful and should not attempt to represent all known information about a gene. Ideally symbols should be no longer than six characters in length. Based on classical genetic guidelines, it is recommended that gene symbols are either underlined or italicized when referring to genotypic information (phenotypic information is represented in standard fonts). Exceptions to this rule are in catalogs of known genes and when referring to fragments or synthesized segments of genes. New symbols must not duplicate existing gene symbols (check theGenome Database at http://gdbwww.gdb.org/ or the HUGO/GDB Nomenclature Committee list of approved gene symbols).

1.2.2. The initial character of the symbol should always be a letter. Subsequent characters may be other letters, or if necessary, Arabic numerals.

1.2.3. All characters of the symbol should be written on the same line; no superscripts or subscripts may be used.

1.2.4. No Roman numerals may be used. Roman numbers in previously used symbols should be changed to their Arabic equivalents.

1.2.5. Greek letters are not used in gene symbols. All Greek letters should be changed to letters in the Latin alphabet (see Web page for details).

1.2.6. 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 (Examples: GLA (galactosidase, α); GLB (galactosidase, β).

1.3 Gene Names

1.3.1. Gene names should be brief and specific and should convey the character or function of the gene.

1.3.2. The first letter of the symbol should be the same as that of the name to facilitate alphabetical listing and grouping, with the exception of the abbreviations noted in 2.6.2.

1.4 DNA Segments

In naming arbitrary DNA fragments and loci, the following guidelines determine each part of the symbol.

Part I. D for DNA.

Part II. 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.

Part III. A symbol indicating the complexity of the DNA segment detected by the probe, with S for a unique DNA segment and Z for repetitive DNA segments found at a single chromosome site or F for small undefined families of homologous sequences found on multiple chromosomes.

Part IV. 1, 2, 3, . . ., a sequential number to give uniqueness to the above concatenated characters.

Part V. When the DNA segment is known to be an expressed sequence the suffix E can be added to indicate this fact.

These numbers can now be generated automatically in the Genome Database, following entry of clone details.

2. Recommendations for Symbol Construction

2.1 Hierarchical Symbols, Gene Families, and Series

2.1.1. Every attempt should be made to represent information in a hierarchical form to facilitate retrieval of sets of related genes from computerized databases.

2.1.2. Where gene products of similar function are encoded by different genes, the corresponding loci are designated by Arabic numerals placed immediately after the gene symbol, without any space between the letters and numbers used (Examples: PGM1, PGM2, PGM3 (three loci for phosphoglucomutase activity); ADH1, ADH2, ADH3 (three alcohol dehydrogenase loci); HBA1, HBA2 (duplicated forms of the α-hemoglobin gene). However, if they exist historically, single-letter suffixes may be used to designate these different loci (Example: LDHA, LDHB, LDHC (three lactate dehydrogenase loci)).

2.1.3. A final character in the gene symbol may be used to specify a characteristic of the gene. While letters to specify tissue distribution have been used historically, Arabic numbers are now preferred as experience has shown that tissue specificity may not be as restricted as described initially.

2.2 Homologies with Other Species

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

2.2.2. Human homologs of genes first identified in other species should *not* be designated by a symbol beginning with H for human.

2.2.3. When a locus 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 humans, 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.

2.2.4. When necessary to distinguish the species of origin for homologous genes with the same gene symbol, the three-letter code for different species already established by the Committee on Standardization in Human Cytogenetics (see Table 1) is recommended. The code is for use in publications only and not incorporated as part of the gene symbol. The species designation is added as a prefix to the gene symbol. For example, HSA signifies *Homo sapiens* and MMU stands for *Mus musculus*. Examples of using the species designation with the gene symbol: human loci: (HSA)G6PD, (HSA)HBB, (HSA)ALB; homologous mouse loci: (MMU)G6pd, (MMU)Hbb, (MMU)Alb.

2.2.5. The agreement between human and mouse gene nomenclature for many homologous gene loci should be continued and extended to other vertebrate species where possible.

2.2.6. Human homologs of genes in invertebrate, or prokaryote, species may be represented by the symbol used in the other species, possibly followed by an L to represent like and a number, if there is more than one human homolog. The use of H to represent homolog is no longer recommended and will be discontinued.

TABLE 1

Species Abbreviations

Abbreviation	Species
HSA	*Homo sapiens*
PTR	*Pan troglodytes* (chimpanzee)
GGO	*Gorilla gorilla*
PPY	*Pongo pygmaeus* (orangutan)
MMU	*Mus musculus*
RNO	*Rattus norvegicus*
MML	*Macaca mulatta* (Rhesus monkey)
CAE	*Cercopithecus aethiops* (African green monkey)
PPA	*Papio papio* (baboon)
FCA	*Felis catus* (cat)
CGR	*Cricetulus griseus* (hamster)
OOV	*Ovies ovies* (sheep)
BBO	*Bos bovinus* (cattle)
SSC	*Sus scrofa* (pig)
OCU	*Oryctolagus cuniculus* (rabbit)
MRU	*Macropus rufus* (red kangaroo)

2.3 Genes Identified from Sequence Information

2.3.1 Predicted genes. Genes predicted from EST clusters or from genomic sequence alone are regarded as putative and are designated by the chromosome of origin and arbitrary number (Example: C2orf1). 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.

2.3.2 Pseudogenes. Molecular technology has identified sequences (generally not transcribed) that bear striking homologies to structural gene sequences. These sequences are termed pseudogenes. To show the relatedness of pseudogenes to functional genes, pseudogenes will be identified with the gene symbol of the structural gene followed by a P for pseudogene. To reserve P for pseudogenes, the use of P as the last character of a structural gene symbol should be avoided where possible (Examples: HBBP1 (hemoglobin, β pseudogene 1); ACTBP1 (actin, β pseudogene 1); ACTBP2 (actin, β pseudogene 2), etc). Pseudogenes may be on different chromosomes or closely linked to the functional gene and occur in varying numbers.

2.3.3 Related sequences. Related sequences identified by cross-hybridization and/or by computer searching of sequence databases (BLAST, FASTA), where no other functional information is available for the construction of a symbol, are designated with the symbol of the known gene followed by an L for like (see also homology, Section 2.2.6).

2.4 Enzymes and Proteins

The rules described in Section 1 apply, but in addition the following should be noted:

2.4.1. Names of genes coding for enzymes are based on those recommended by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (http://alpha.qmw.ac.uk/~ugca000/iupac/jcbn/). Names of plasma proteins, hemoglobins, and specialized proteins are based on standard names and those recommended by their respective committees.

2.5 Clinical Disorders

2.5.1. Inherited clinical disorders (monogenic Mendelian inheritance). The first gene symbol allocated to an inherited clinical phenotype may be based on an acronym that has been established as a name for the disorder, while following the rules described in Section 1 (Example: ACH for achondroplasia). However, it is usual for this symbol to change when the gene product or function is identified. In some cases a gene symbol based on product or function will already exist, and 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 changed to FGFR3 and the name to fibroblast growth factor receptor 3 (achondroplasia, thanatophoric dwarfism).

2.5.2 Complex/polygenic traits. Genome searches may suggest a contributing locus in a complex trait, which may for convenience be given a gene symbol, although a proportion of these will disappear in time. A symbol allocated to such a gene will not be reused.

2.5.3 Contiguous gene syndromes. Syndromes clearly associated with multiple 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: ANCR (Angelman syndrome chromosome region), DCR (Down syndrome chromosome region)). However, as advances in database design have now increased the possible ways of representing this type of information, we recommend that such symbols are now classified as syndromic region symbols and not gene symbols.

2.5.4 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, the chromosome number, CR (for chromosomal region), and then an arbitrary number.

2.6 Letters Reserved for Specific Usage

2.6.1. Certain letters or combinations of letters are used as the last letter in a symbol to represent a specific meaning, these are P for pseudogene (but note also BP for binding protein), L for like (see Section 2.1.), R for receptor or regulator, and N or NH for inhibitor. The use of these for other meanings should be avoided where possible.

2.6.2. If the name of a gene contains a character or property for which there is a recognized abbreviation, the abbreviation should be used (Example: the single-letter abbreviation for amino acids used in aminoacyl residues or approved biochemical abbreviations such as GLC for glucose and GSH for glutathione).

3. Allele Terminology

Allele terminology is now the responsibility of the Mutation Database (http://www2.ebi.ac.uk/mutations/cotton/).

4. Printing Gene and Allele Symbols

It is recommended that gene and allele symbols are underlined in the manuscript and italicized in print. Italics need not be used in catalogs. It may be convenient in manuscripts, computer printouts, and printed text to designate a gene symbol by following it with an asterisk (Example: PGM1*). When only allele symbols are displayed they can be preceded by an asterisk. For example, for PGM1*1, the allele is printed as *1.

ACKNOWLEDGMENTS

The Nomenclature meeting held in Toronto on March 5, 1997, was made possible by the support of the EU, through a contract to HUGO. We also thank all the participants at the meeting and everyone who has subsequently commented on the guidelines.

REFERENCES

McAlpine, P. J. (1995) *Trends Genet.* **Suppl.** 39–42.

Shows, T. B., McAlpine, P. J., Boucheix, C., Collins, F. S., Conneally, P. M., Frezal, J., Gershowitz, H., Goodfellow, P. N., Hall, J. G., Issitt, P., Jones, C. A., Knowles, B. B., Lewis, M., McKusick, V. A., Meisler, M., Morton, N. E., Rubinstein, P., Schanfield, M. S., Schmickel, R. D., Skolnick, M. H., Spence, A. M., and Sutherland, G. (1979). International system for human gene nomenclature. *Cytogenet. Cell. Genet.* **25:** 96–116.

NOMENCLATURE

Rules and Guidelines for Mouse Gene Nomenclature: A Condensed Version

Lois J. Maltais,[1] Judith A. Blake,
Janan T. Eppig, and Muriel T. Davisson[2]

The Jackson Laboratory, 600 Main Street, Bar Harbor, Maine 04609

Introduction

Gene nomenclature guidelines are based on the premise that the primary purpose of a gene or locus symbol is to provide a brief and universally acceptable symbol that uniquely identifies

[1] To whom correspondence should be addressed. Telephone: (207) 288-6429. Fax: (207) 288-6132. E-mail: ljm@informatics.jax.org.

[2] Chairperson, International Committee on Standardized Genetic Nomenclature for Mice.

0888-7543/97 $25.00

a specific gene or locus; all other purposes of a symbol are secondary and should not interfere with this primary purpose. Gene names, brief descriptive phrases that define the symbol, and gene descriptions in publications and electronic databases should be the primary means for conveying information about the gene. Complex information about a gene or locus, such as the properties of the assay used to identify it, should be conveyed in the description accompanying the gene and not be part of the unique identifying symbol.

The rules and guidelines reported here are a condensed version of the most recently published unabridged set (Committee, 1996; International Committee, 1994). All subsections are numbered as they appear in the complete guidelines for ease of cross-reference. An electronic version of the complete nomenclature rules and guidelines, to include chromosome anomalies and different types of strains, is available at the Mouse Genome Informatics (MGI) website URL: http://www.informatics.jax.org/. Information, guidance, and an electronic submission form for approval of mouse gene symbols are also at the MGI site (see Fig. 1).

1.1. Rules for Gene Nomenclature

1.1.1. Names of Genes or Loci

Names of genes and loci should be brief and should convey the character by which the gene is recognized. Genes are functional units, whereas a locus can be any distinct, recognizable DNA sequence (see 1.1.3). Hyphens are no longer used in gene names, except when they are part of a compound word.

1.1.2. Symbols for Genes

The initial letters of names and symbols should be the same for convenience in alphabetical listings. Arabic numbers may be included but the symbol should always begin with a letter. Roman numbers and Greek letters should not be used. Names of persons or places, in general, are discouraged in gene names or symbols. Symbols for related genes, such as genes in the same family, usually have a common stem symbol of 2 or 3 characters followed by 1 or 2 distinguishing characters to make each symbol unique. Some characters at the ends of symbols have special meaning; e.g., *r* is commonly used for receptor and regulatory loci, *bp* is commonly used for binding protein loci, and *e* is added to DNA locus symbols to indicate they are expressed. Gene symbols ending in *v* should be reserved for genes pertaining to viruses.

1. Hyphens are used only in gene symbols for clarity, primarily to separate characters that together might be confusing, e.g., *Lamb 1-2*. The hyphenated suffixes *-rs* and *-ps* are used to denote related sequences and pseudogenes, respectively.
2. Symbols for genes should not exceed 10 characters.
3. Except in the case of loci first discovered because of a recessive mutation, the initial letter of the locus symbol should be capital, and all others lowercase, e.g., *Ath1*.
4. In published articles, gene symbols should be set in italics.
5. Identification of new loci should not be assumed from the discovery of variation, whether morphological, biochemical, quantitative, or antigenic, without appropriate genetic tests.
6. A proposed new symbol must never duplicate one already used for another locus.
7. When a well-known locus has been recognized initially by a mutation and later identified, e.g., by cloning, the locus is identified by the symbol for the identified gene and the mutant allele symbol is designated as a superscript to the gene symbol (see 1.1.7 Alleles).
8. Symbols for quantitative trait loci genes follow the same rules as above; those affecting the same trait shall be given the same stem symbol and serially numbered. A "*q*" may be used as the final letter preceding the serial number but is not required.
9. Expressed sequence tagged (EST) loci, when mapped to chromosomes, may be given either D symbols with a final *e* for expressed or the symbol *ESTM###*, in which *M* identifies mouse as opposed to human EST loci and ### is a serial number assigned from the Mouse Genome Database.
10. Genes encoded by the opposite (antisense) strand of a known gene shall be given their own symbols.
11. Alternate transcripts from the same gene should not be given different "locus" symbols.

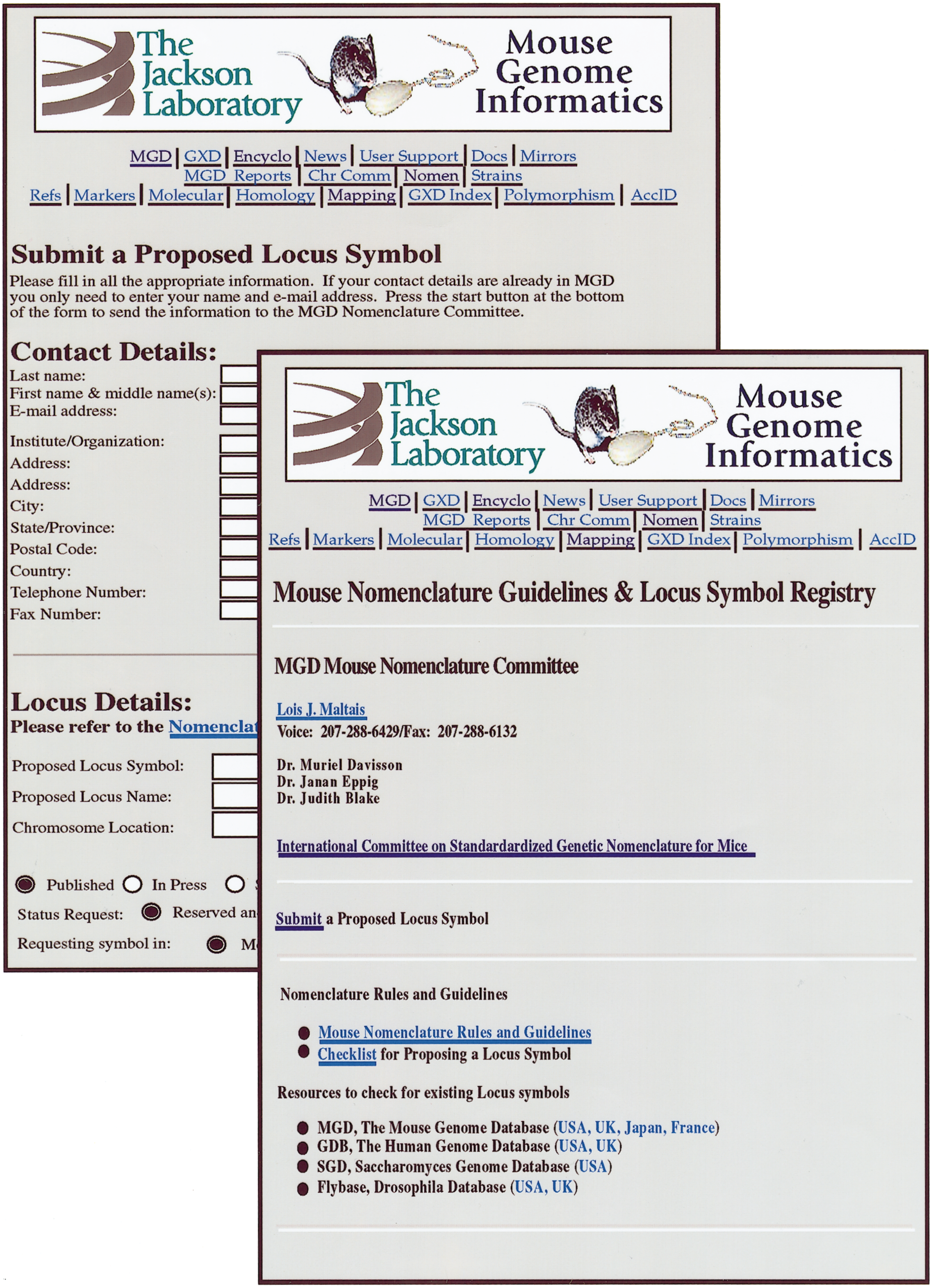
1.1.3. Genes and Loci Recognized by DNA Sequence

D symbols are used in two ways:

1. Loci recognized by anonymous DNA probes should be given D symbols.
2. Intragenic loci may be given D symbols to distinguish individual sites within a gene. Intragenic D-locus symbols should be used only (a) in describing intragenic mapping analysis or (b) in stating that a gene was typed using an intragenic D locus. The use of D symbols for intragenic DNA segments is discouraged. YAC ends may be given D symbols when they are used for genetic mapping.

Loci recognized by variation in copy number, such as mini- or microsatellites, should be given D symbols. If such microsatellite (D-symbol) loci are within or very near known genes and thus can be used to detect those genes, then the gene symbol should always be used to refer to the gene; that is, the gene's location on the chromosome, and the D symbols should be used only to refer to specific sites within the gene, e.g., to convey intragenic mapping information. When D-symbol loci fall within known genes, on general maps the gene will be identified by the gene symbol, and the locus (D) symbols will appear in locus lists and databases cross-referenced with the gene. Locus symbols would, of course, be used on fine-structure, high-resolution maps around and within genes.

FIG. 1. World Wide Web pages of the mouse nomenclature guidelines and locus symbol registry from the Mouse Genome Database, Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, Maine 04609. URL:http:www.//informatics.jax.org/.



MGD | GXD | Encyclo | News | User Support | Docs | Mirrors
MGD Reports | Chr Comm | Nomen | Strains
Refs | Markers | Molecular | Homology | Mapping | GXD Index | Polymorphism | AccID

Submit a Proposed Locus Symbol

Please fill in all the appropriate information. If your contact details are already in MGD you only need to enter your name and e-mail address. Press the start button at the bottom of the form to send the information to the MGD Nomenclature Committee.

Contact Details:

Last name:
First name & middle name(s):
E-mail address:
Institute/Organization:
Address:
Address:
City:
State/Province:
Postal Code:
Country:
Telephone Number:
Fax Number:

Locus Details:

Please refer to the Nomenclat

Proposed Locus Symbol:
Proposed Locus Name:
Chromosome Location:

● Published ○ In Press ○ S

Status Request: ● Reserved an

Requesting symbol in: ● M

MGD | GXD | Encyclo | News | User Support | Docs | Mirrors
MGD Reports | Chr Comm | Nomen | Strains
Refs | Markers | Molecular | Homology | Mapping | GXD Index | Polymorphism | AccID

Mouse Nomenclature Guidelines & Locus Symbol Registry

MGD Mouse Nomenclature Committee

Lois J. Maltais
Voice: 207-288-6429/Fax: 207-288-6132

Dr. Muriel Davisson
Dr. Janan Eppig
Dr. Judith Blake

International Committee on Standardardized Genetic Nomenclature for Mice

Submit a Proposed Locus Symbol

Nomenclature Rules and Guidelines

- Mouse Nomenclature Rules and Guidelines
- Checklist for Proposing a Locus Symbol

Resources to check for existing Locus symbols

- MGD, The Mouse Genome Database (USA, UK, Japan, France)
- GDB, The Human Genome Database (USA, UK)
- SGD, Saccharomyces Genome Database (USA)
- Flybase, Drosophila Database (USA, UK)

D symbols are composed of four parts:

1. D for DNA.
2. 1 . . . 19, X, and Y for the chromosomal assignment, and 0 for unmapped loci.
3. A 2- or 3-letter laboratory registration code indicating the laboratory or scientist describing the locus. Only the first letter of the laboratory code is capitalized. Laboratory registration codes are assigned from a central registry and can be obtained by contacting the Institute for Laboratory Animal Research in Washington, DC. For a current list of approved lab codes or to register a new lab code, go to URL address http://www2.nas.edu/labcode/.
4. A unique serial number.

If an anonymous DNA locus identified only by a D symbol is later identified as a known locus or the function of the gene is determined, the D symbol should either be replaced by the known gene's symbol or be changed to a new gene symbol that is an acronym for the new gene's name.

Anonymous DNA loci from the human genome that cross hybridize with mouse DNA and are mapped to a mouse chromosome retain their human symbol in all uppercase letters, and the mouse chromosome number and a capital *H*, for human, are inserted after the *D*, e.g., *D16H21S56*.

Xrf shall be used as the "lab code" in D symbols that designate cross-referenced genes in mice and yeast or other species. Symbols will be *DChr#Xrf###*, in which ### is a serial number assigned by the Genome Database at Johns Hopkins.

Some chromosomal regions can be detected by specific cytological staining methods that reveal the whole region, or loci within the region can be detected with DNA probes. Chromosomal nomenclature guidelines should be followed for cytologically detected chromosomal regions and gene nomenclature for genetic loci.

1.1.4. Pseudogenes and Related Sequences

Loci related by sequence (i.e., recognized by the same probe) but not proved to be pseudogenes are symbolized by adding *-rs* to the symbol of the primary locus that the probe identifies followed by serial numbers. If a locus related by sequence is proved to be a pseudogene, it is denoted by the suffix *-ps* and followed by an appropriate serial number.

1.1.5. Loci That Are Members of a Series or Family

Loci that are members of a series specifying similar proteins or other phenotypic characteristics should be designated by the same character symbol with the addition of distinguishing serial numbers, e.g., *Es1, Es2*.

Genes that are members of a gene family should have a common stem or root of 2 or 3 letters.

1.1.6. Homology with Other Organisms

It is highly desirable that terminology for homologous genes be standardized among species. In choosing a gene symbol, an attempt should first be made to discover and use any symbols already adopted for this gene in other species. Care should be taken that such symbols *do not duplicate any already in use in the mouse* for other loci. Do not insert the letter *m* or *M* (for mouse) as the first letter of the symbol for a locus with homologs in other species. *Note:* To describe conservation between species, "synteny" should not be used synonymously with "homology" because synteny means literally on the same thread or chromosome. The appropriate phrases are "conserved synteny" when a gene and its other species homolog are assigned to homologous chromosomes in the mouse and the other species and "conserved linkage" or "conserved segment" when the two genes are positionally mapped within the same region on the two species' chromosomes. "Conserved ordered segment" may be used when the order of genes within the segment is conserved.

1.1.7. Alleles

Alleles are usually designated by the locus symbol with an added superscript (in italics when printed).

1. In the case of mutant genes for which there is clearly a wildtype, the symbol for the first discovered mutant allele becomes both the gene symbol and the symbol for that allele. No superscript is then used, e.g., *Ca*. When further alleles are discovered, the first mutant allele may still be without a superscript and allele symbols for new alleles are added as superscripts, e.g., Ca^{J}.

When a spontaneous mutation is cloned or shown to occur in a previously named candidate gene, the mutation's symbol is changed to become an allele at the cloned locus by turning the mutation symbol into an allele symbol; e.g., the *shi* (shiverer) mutation in the *Mbp* (myelin basic protein) gene becomes Mbp^{shi}. If the original mutation symbol already has a superscript, the mutation and allele symbols are placed on one line in the new superscript and hyphenated, e.g., $Mbp^{shi\text{-}mld}$.

2. Recessive alleles should be indicated by the use of a lowercase initial letter for a mutant gene. Two exceptions to this rule are allowed for targeted and cloned mutant genes when the original cloned gene symbol starts with an uppercase letter:

   (1) If the phenotype of mutant alleles may be recessive or codominant depending on the method of determination (e.g., visual, protein assay, or DNA genotyping), the use of upper- or lowercase letters will depend on what the naming investigator considers the defining phenotype.

   (2) When a mutation is shown to occur in a cloned candidate gene and its symbol is changed to become an allele of the cloned gene, the first letter of the gene symbol may remain uppercase and the inheritance pattern may be conveyed in the allele symbol.

3. Refer to unabridged version.
4. Refer to unabridged version.
5. Wildtype should be designated by a + sign, with the locus symbol as superscript, e.g., $+^{pe}$. A + sign alone may be used when the context leaves no doubt as to the locus

represented. Reversions from mutant allele to wildtype should be distinguished from the original wildtype allele by designating them by the locus symbol, with a + sign as superscript, e.g., *pe*$^{+}$.

6. Indistinguishable alleles of independent origin are designated by the existing gene symbol with a series appended as a superscript in italics, e.g., *Kit*$^{W\text{-}81J}$.

When two named mutant genes are found to be alleles at the same locus, the symbol published or assigned first remains the locus symbol and the symbol of the second gene is superscripted as an allele symbol for that mutation.

7. Mutations or other variations occurring in known alleles may be denoted by a superscript *m* followed by an appropriate series symbol and separated from the original allele symbol, if one exists, by a hyphen, e.g., C57BL/6J-*Hprt*$^{b\text{-}m3}$. Mutant alleles created by targeted mutagenesis should have a *t* preceding the *m*, e.g., *Cftr*^{tm1Unc}. When an existing gene is replaced with a different, functional, gene, called a ''knock-in,'' the symbol shall be written as an allele of the original gene.

8. Mutant alleles that turn out subsequently to be deletions retain their allelic designation. If the deletion deletes more than one gene and is cytologically visible, the deletion should be given a chromosome anomaly designation containing the original allele designation, and the allele symbol is used as the abbreviation, e.g., *Del(10)Mgf*$^{Sl\text{-}12H}$*1H*, abbreviated *Mgf*$^{Sl\text{-}12H}$.

9. Refer to unabridged version.

1.1.8. Lethals

Appropriate locus symbols for recessive lethals with no known heterozygous effect and unidentified function consist of a lowercase letter *l* followed by the number of the chromosome on which the locus is located in parentheses and a series symbol indicating the serial number of the lethal in the laboratory of origin, e.g., *l(17)2Pas.*

1.1.9. Viruses

Nomenclature for genes related to the expression of viral antigens, or to sensitivity or resistance to viruses, should follow the standard rules for gene nomenclature.

1.1.10. Oncogenes

Nomenclature for mouse cellular oncogene sequences should follow the standard nomenclature for oncogenes. For the mouse cellular locus, however, in lists of symbols and maps, the prefix *c-* denoting cellular sequence should be omitted, and the initial letter of the symbol should be capitalized; e.g., *c-myc* becomes *Myc.*

1.1.11. Phenotype Symbols

Phenotype symbols, typically for antigen or biochemical genes, should be the same as genotype symbols except that symbols for phenotypes should be in capital letters, not italicized, and with superscripts lowered to the line; e.g., the *Gpi1*^{a} allele encodes the GPI1A protein.

1.1.12. Gene Complexes

Gene complexes are considered to exist when a number of apparently functionally or evolutionarily related loci are genetically closely linked. Alternative states of complexes are referred to as haplotypes rather than alleles. Detailed rules for complexes and the loci within them are given in the complete version of the guidelines.

1.1.13. Mitochondrial Genome

Loci in the mitochondrial genome should be denoted by the prefix *mt*- set off from the main symbol by a hyphen.

1.1.14. Antigenic Variants

Symbols adopted for loci concerned in cell-membrane alloantigens should be based on the method of demonstrating such loci. Details of antigen gene nomenclature are given in the complete guidelines.

1.2. Nomenclature for Special Classes of Genes and Gene Complexes

Biochemical nomenclature should be in accord with the rules of the International Union of Biochemistry, Commission on Biochemical Nomenclature. Detailed guidelines for symbolizing special classes of biochemical, immunological, and homeobox genes may be found in the complete version of the guidelines.

1.3. Rules for Naming Transgenes

All DNA sequences that are experimentally and stably introduced into the germ line of animals are technically transgenes. However, only genes or gene segments that are introduced to study their expression or phenotypic effect are considered transgenes and symbolized with the conventions given below. "Knockout" or directed mutation of a specific known gene should be designated using standard allele symbol conventions, e.g., *Cftr*^{tm1Unc}.

1. The transgene symbol consists of three parts, all in Roman typeface, as follows, TgX(YYYYYY)#####Zzz, where TgX is the mode, (YYYYYY) is the insert designation, ##### is the laboratory-assigned number, and Zzz is the laboratory registration code.

The mode shows the object is a transgene and always consists of the letters *Tg* followed by a letter designating the mode of insertion of the DNA: *H* for homologous recombination, *R* for insertion via infection with a retroviral vector, and *N* for nonhomologous insertion.

The insert designation is a symbol for the salient features of the transgene, as determined by the investigator. It is always contained within parentheses and consists of no more than six characters. It should identify the inserted sequence and indicate important features. When the insertion utilizes sequences from a named gene, it should contain the standard symbol for that gene, e.g., TgN(GPDHIm)1Bir. A transgene symbol may be abbreviated by omitting the parentheses and insert designation once the full symbol has been used in a paper: TgN1Bir.

The laboratory-assigned number is a number from 1 to 99,999 that is uniquely assigned by the laboratory to each stably transmitted insertion. A laboratory registration code is uniquely assigned to each laboratory originating transgenic animals, DNA loci, or inbred strains (see 1.1.3). When a mutation that produces an observable phenotype is caused by the insertion, the locus so identified must be named according to standard procedures for the species involved. The allele of the locus identified by the insertion can then be identified by the abbreviated transgene symbol, e.g., $ho^{TgN447Jwg}$.

REFERENCES

Committee on Standardized Genetic Nomenclature for Mice, Davisson, M. T., Chairperson. (1996). Rules and guidelines for gene nomenclature. *In* "Genetic Variants and Strains of the Laboratory Mouse" (M. F. Lyon, S. Rastan, and S. D. M. Brown, Eds.), 3rd ed., Vol. 1, pp. 1–16, Oxford Univ. Press, Oxford.

International Committee on Standardized Genetic Nomenclature for Mice, Davisson, M. T., Chairperson. (1994). Rules and guidelines for genetic nomenclature in mice. *Mouse Genome* **92:** vii–xxxii.