MEETING REPORT

# Report of the Second International Nomenclature Workshop

## Cambridge, United Kingdom, May 1–2, 1999

Julia A. White,*[,1] Rolf Apweiler,† Judith A. Blake,‡ Janan T. Eppig,‡ Lois J. Maltais,‡ and Sue Povey*

**MRC Human Biochemical Genetics Unit, University College London, 4 Stephenson Way, London NW1 2HE, United Kingdom; †EMBL Outstation–European Bioinformatics Institute (EBI), Cambridge CB10 1SD, United Kingdom; and ‡The Jackson Laboratory, Bar Harbor, Maine 04609*

### INTRODUCTION

A group of 47 participants met at the Second International Nomenclature Workshop (INW2) on May 1–2, 1999, to discuss issues of importance to gene nomenclature and changes that had occurred since the previous nomenclature workshop in 1997. The organizers set out the key objectives for the meeting as follows:

1. To work toward establishing guidelines for gene family nomenclature.
2. To consider strategies for dealing with increased volumes of data.
3. To redefine the point at which nomenclature committees become involved in the naming of genes.
4. To consider the future role(s) of nomenclature committees and other organizational structures that will be required to deal with gene nomenclature issues.

The participants, from nine different countries, included representatives from species gene nomenclature committees; scientists with specialized knowledge of particular gene families; genetic database curators, developers, and managers; and representatives from the pharmaceutical industry.

### GENE FAMILY ANALYSIS

#### Nomenclature Perspective

The current interest in gene families is of great relevance to the issue of gene nomenclature, as outlined by Hester Wain of the Human Gene Nomenclature Committee. The assignment of groups of genes, related by defined levels of sequence similarity, to superfamilies, families, and subfamilies is helpful to nomenclature editors trying to assign gene symbols in a logical, hierarchical, and meaningful system. Once a root (sometimes referred to as a stem) symbol is assigned and a numbering scheme devised for a particular gene family, the choice of gene symbol for new members of the family becomes considerably easier. The root symbol consists of a small number of letters (usually three or four), which represent some known structural or functional characteristic of the gene family (e.g., CYP for the cytochrome P450 family, SEMA for the semaphorins). The numbering scheme can be simple consecutive numbers or hierarchical combinations of letters and numbers representing subfamilies as well as individual genes (e.g., CYP2B7, SEMA3A). The system ensures an efficient use of letter combinations, allowing more genes to be assigned unique and meaningful symbols without them becoming too long and difficult to remember.

#### Defining Gene Families

The concept of the gene family is not clearly defined, and the term has been used to signify groups of genes related by function, by sequence, or by phenotype caused. Gene symbol series representing all these possible definitions already exist, and widespread changes, to reflect only families related by sequence, are not recommended. However, the term "superfamily" is considered to be more specifically defined and should be used only to refer to groups of genes related by evolutionarily defined sequence similarities.

#### Advantages and Drawbacks of Gene Family Hierarchy Systems

All the gene families considered at the INW2 had member genes across a range of species; in some cases, the gene family group contained both prokaryotic and mammalian representatives. This raises problems of communication between various species nomenclature committees when gene symbol schemes for the superfamilies are proposed and discussed. It was agreed that specialists in the field were the best people to decide on the criteria for membership of a gene superfamily, with advice from committees on choosing a root symbol and a suitable numbering scheme. However, how many committees they should consult, and how the results of their discussions should be communicated to other committees, was not clear at present. Often the mouse and HUGO nomenclature committees work collaboratively with experts for the particular gene family to structure the family designations. Examples of several such schemes (including the cytochrome P450 and UDP glycosyltransferase superfamilies) were described by Dan Nebert, who estimates that each of 8000–15,000 core ancestral genes (that is, those present several hundred million years ago), might now represent a unique superfamily. This estimate is arrived at by comparison of the complete genome sequences that are now available (mainly bacteria, plus archaea, and more recently a few eukaryotes). These are relatively manageable numbers for the assignment of unique short letter combinations as root

Details of this meeting, including the abstracts, can be seen at http://www.gene.ucl.ac.uk/nomenclature/INW2.html.

[1] To whom correspondence should be addressed: Telephone: +44 (0) 207 679 5027. Fax: +44 (0) 207 387 3496. E-mail: julia@galton.ucl.ac.uk.

0888-7543/99 $30.00




Genomics **62,** 320–323 (1999)
Article ID geno.1999.6013

symbols, compared to the alternative of naming, for example, approximately 10 times as many individual genes in the human genome.

As more superfamilies are analyzed, interesting new situations arise. In the case of the carbonic anhydrase gene family (described by David Hewett-Emmett), about one-third of the gene products are not functioning as carbonic anhydrases, yet their sequences are highly conserved in mammals. In terms of the sequence relationship, it makes sense to allocate symbols in the same series to all the genes of the family. The gene products may, however, be very different in function and may even belong to more than one gene family, by having more than one functional domain (whether active or inactive).

Some further drawbacks of the system of grouping genes into superfamilies based on their sequence relationships were emphasized by Kirill Degtyarenko, who described the importance of bioinorganic motifs in functional diversity with reference to metalloprotein families. However, it was clear from the example of the Mendel database used in plants that grouping newly discovered genes by sequence similarity can be very informative and potentially useful in contributing to the automation of sequence annotation. David Lonsdale described the Mendel-ESTs database and how new plant sequences were assigned to gene families. The separation of this activity from the longer term assignment of meaningful names by committees of experts is in contrast to the current activities of most animal species nomenclature committees.

## DATABASE INTEGRATION ISSUES

A major issue currently facing nomenclature committees is the huge volume of information accumulating in sequence databases. Rolf Apweiler, of SWISS-PROT, described current efforts to provide meaningful annotation for new sequences by automated means. This allows value to be added to the data without a huge increase in the "manual" curation effort, but must be carefully controlled to prevent recursive editing and accumulation of automated "nonsense" annotation. Standardized nomenclature clearly has a role to play in this process, and nomenclature editors can in turn benefit from the information provided. It must be recognized, however, that it is not possible to encapsulate all known information about a gene in its symbol and name.

In the past, approved gene nomenclature has been associated largely with model organism databases, and the link between approved symbol and sequence data has been surprisingly difficult to achieve with a reasonable level of accuracy. Donna Maglott (NCBI) described the current efforts in the LocusLink and RefSeq projects to improve this connection and thus add value to the sequence data. These projects involve collaboration between teams at NCBI (including the OMIM database) and the Human Gene Nomenclature Committee and are already raising the profile of approved gene nomenclature in the human genetics community. The key role of the nomenclature committees in data integration and exchange between databases was highlighted by Liora Yaar in describing the GeneCards database, which relies on the use of approved gene symbols to gather data. However, the volume of data is increasing the workload of the committees, and the GeneCards project is proposing the development of automated systems to replicate at least some of the extensive checks and searches performed by nomenclature editors.

## FUNCTION AND PHENOTYPE

It has long been a recommendation of gene nomenclature committees, at least for human and mouse, that gene symbols should be based on function whenever possible. This reflects to a large extent the types of data that were generally available at the time the various guidelines were first formalized. The use of phenotypic descriptions as a basis for gene nomenclature is common in some species, but in human and mouse is considered only when no other information is available. The recommendation has been that a mutant phenotype designation is changed as soon as normal gene function is defined. The situation is changing, however, and many genes are identified and described in publications with little or no accompanying phenotypic or functional data. It now often seems more appropriate to base symbols on sequence similarities and the presence of various motifs or domains that may indicate possible function, but are not conclusive. This can lead to problems of overlap, when the presence of multiple domains means that some genes fall into more than one grouping. Clearly the symbol and name alone cannot contain all the information that might be useful, and different interest groups require different types of information and different classification schemes. This was exemplified by the clinical perspectives described by Jean Frézal (GenAtlas). For many researchers in human genetics, classifications of human disease phenotypes are of great importance and must not be lost when the causative genes are themselves renamed. Likewise, other information, such as the presence of various domains not represented in the approved nomenclature, needs to be standardized, linked, and easily searchable in the databases.

Judy Blake and Selina Dwight described different aspects of a recent collaboration between the yeast, fly, and mouse model organism databases to provide controlled vocabularies for common use in the areas of gene function, biological processes, and cellular locations. This gene ontologies project aims to provide terms and definitions that anyone annotating gene characteristics can use, and by so doing, they will be using the terms of known definition to describe their genes. By annotating genes in this way, some of the pressure on nomenclature revisions will be relieved since the gene symbol/name need not be changed to reflect increased knowledge about the gene.

## HOMOLOGY, ORTHOLOGY, AND PARALOGY

One strong argument in favor of approved nomenclature, and collaboration between nomenclature committees for different species, has been the utility of related symbols in comparative genomics. Comparative maps can be easier to interpret with consistent nomenclature, and human and mouse committees in particular have collaborated closely for many years to achieve this. Anne Kwitek-Black described a new radiation hybrid map platform for constructing virtual comparative maps between rat, mouse, and human, which has already been useful in prediction of QTL (quantitative trait locus) gene regions. The use of human genetic markers in other species raises issues of common nomenclature, which were described by Nicole Créau, with

reference to nonhuman primates. The resulting maps and integrated data can be of great interest in highlighting complex evolutionary events.

Jenny Graves, however, described how recent advances were now making the situation more complex. The X and Y chromosomes in particular provide a number of examples where orthology between species was very difficult to establish and where the existence of variable numbers of closely related genes and pseudogenes could easily confound efforts to make the gene symbols the same across all mammals.

While human and mouse nomenclature committees work closely together to keep gene symbols of orthologs in synchrony where possible, other database mechanisms for relating genes as orthologs, paralogs, etc., need to be developed more fully. As in the ontologies mentioned above, the development of a controlled list of gene families would facilitate this process.

Janan Eppig also commented on the difficulties now being encountered in this area and even in the definition of the terms homology, orthology, and paralogy. Difficulties with assessing homology and orthology, even among mammalian species, include highly overlapping and similar sequence and function among gene families, low sequence similarity among functional orthologs, variable copy number among closely related species, and convergent functional evolution of genes with different origins. There is a need to be aware of differences, and not just similarities, between homologs and orthologs, on which much genomic work is currently focused. The inference of function and the identification of genes based on sequence similarity and syntenic mapping alone are useful working tools, but also can lead to erroneous assumptions and error propagation. Nomenclature committees must therefore be wary of oversimplification in deciding on naming schemes to avoid presenting a misleading interpretation of the data. In addition, horizontal gene transfer across phyla and even kingdoms now appears much more common than first believed and could present significant nomenclature problems. This again accentuates the need for good communication and coordination between those responsible for nomenclature in different species.

## NOMENCLATURE AUTHORITY—ORGANIZATIONAL STRUCTURES

The presentations of several expert groups clearly demonstrated the importance of involving those knowledgeable in the field in the naming of genes. The complexity of the immunoglobulin, T cell receptor, and MHC genes, for example, is best understood by those working in this field. The IMGT database (described by Marie-Paule Lefranc) serves well the needs of this particular community. However, it is important that other databases of wider scope can obtain data easily from these specialized resources to maintain consistency.

Jonathan Hodgkin, who provided the quote of the meeting (see below), emphasized the need for coordination between the various species nomenclature groups. His description of the progress in naming 20,000 genes in *Caenorhabditis elegans* again showed that it was possible, and indeed desirable, to separate the naming process from the initial identification of the gene. It was also clear that the lines of communication between the gene family expert groups and the smaller nomenclature committees of some of the model organisms needed considerable improvement, to prevent duplication of effort and subsequent confusion.

Some parallels could be identified with the situation in mutation nomenclature and databases. Richard Cotton described the locus-specific mutation databases, which were very varied in format, style, and content and which needed some form of quality control and coordination to maintain their usefulness to the wider community. One important outcome of this meeting was the introduction of gene family researchers to one another, resulting in some very useful discussions.

Other areas of biology also have organizational structures to deal with nomenclature issues, which may provide useful examples for the gene nomenclature committees. The International Union of Biochemistry and Molecular Biology (IUBMB) and the International Union of Pharmacology (IUPHAR), for example, have numerous subcommittees dealing with nomenclature of specific areas. Pat Humphrey (IUPHAR) illustrated this in describing the process of establishing the receptor coding system. In summing up, Julia White suggested that in the future, some form of overall gene nomenclature coordinating body might be required to ensure communication among researchers of the various species and special interest groups and to identify areas currently not being covered.

## CONCLUSIONS

The main conclusions reached at the meeting were as follows:

- It should be possible to standardize gene nomenclature across species, and this is a worthwhile goal even though some changes in principles and practice will be required. It is recognized that genes in different species may not have a one to one relationship.
- A gene needs a name when some human shows interest in it, usually wanting to publish something about it. Temporary identification for putative genes identified from sequence, in the form of clone name and number of gene, is acceptable until this point. Unique database identifiers (accession numbers) and the name of the database should always be used as stable identifiers for the sequence.
- The name of a gene cannot contain all known information and is essentially a unique human-friendly identifier, containing some clue about function if possible. The development of controlled vocabularies for function will be useful in this respect, allowing consistent additional annotation independent of the gene symbol. Interoperability of databases allowing approved symbols to propagate through to the sequence databases is essential.
- In many cases there are genes that are clearly related but nothing useful is yet known about the structure or function. When names are required for such genes, an arbitrary identifier and a number will be an acceptable approved symbol. The letters to be used for the arbitrary identifier (possibly FAM for "family") and the criteria by which genes are accepted for inclusion in such families or subfamilies are still under discussion.
- Specialist advisors are required to clarify the nomenclature of particular gene families. It is important that such advisors consult not only with the scientific communities

involved with those genes, but also with the species-specific nomenclature committees to avoid major inconsistencies. The publication of such nomenclature schemes is to be strongly encouraged, both as a point of reference and as subject for further discussion.

- The main nomenclature groups should maintain updated Web pages (or a link to such Web pages) that include comprehensive information on gene family information. Those gene family Web pages using approved nomenclature should be indicated, but others should not be excluded.
- Continued pressure should be exerted on journals to promote standardized nomenclature.
- Input from those involved in bacterial gene nomenclature, who were not represented at the meeting, would be welcomed.
- When a species identifier is needed in a name, the 5-letter designation adopted by SWISS-PROT is recommended (http://www.expasy.ch/cgi-bin/speclist).
- All groups experience funding difficulties, and nontraditional sources of funding should be sought. There is considerable enthusiasm from industry for approved nomenclature, and efforts should be made to obtain sponsorship (although the idea of naming genes after benefactors did not meet with wide approval).
- It was recognized that this particular grouping would rarely come together in other contexts and that it would be difficult to attach another workshop to any other meeting. Another meeting will probably be necessary within 2 years. To produce more concrete conclusions in the future, a more formal structure should be set up for decision making.
- In view of the huge size of the task facing gene nomenclature committees over the next few years (a potential 70,000 genes for the attention of the human gene nomenclature committee alone), it was generally agreed that coordination was vital to prevent duplication of effort. New organizational structures to promote communication, to increase efficiency, and to involve more people in the naming process will be considered.

## PARTICIPANTS

The following people attended this meeting:

Organizers included Dr. Sue Povey (HUGO Nomenclature Committee, UK); Dr. Julia White (HUGO Nomenclature Committee, UK); Dr. Janan Eppig (The Jackson Laboratory, USA); Dr. Judith Blake (The Jackson Laboratory, USA); Ms. Lois Maltais (The Jackson Laboratory, USA); Dr. Rolf Apweiler (The European Bioinformatics Institute, UK).

Invited Participants included Dr. Benedict Arnold (John Innes Centre, UK); Dr. Michael Ashburner (The European Bioinformatics Institute, UK); Dr. Alan J. Barrett (MRC Molecular Enzymology Laboratory, UK); Dr. Richard Balderelli (The Jackson Laboratory, USA); Ms. Carol Bocchini (OMIM, USA); Dr. T. E. (Tom) Broad (AgResearch Invermay, New Zealand); Dr. Elspeth Bruford (HUGO Nomenclature Committee, UK); Dr. R. G. H. Cotton (St. Vincent's Hospital, Australia); Dr. Nicole Créau (CNRS, France); Dr. Kirill Degtyarenko (The European Bioinformatics Institute, UK); Dr. Gaudenz Dolf (University of Berne, Switzerland); Dr. Richard Durbin (Sanger Centre, UK); Dr. Selina S. Dwight (*Saccharomyces* Genome Database, USA); Dr. Jean Frézal (GenAtlas, France); Dr. Jennifer A. Marshall Graves (La Trobe University, Australia); Dr. David Hewett-Emmett (University of Texas, Houston–SPH, USA); Dr. Jonathan Hodgkin (MRC Laboratory of Molecular Biology, UK); Dr. Pat Humphrey (University of Cambridge, UK); Dr. Ian J. Jackson (Western General Hospital, UK); Dr. Anne E. Kwitek-Black (Medical College of Wisconsin, USA); Dr. Marie-Paule Lefranc (CNRS, France); Dr. Goran Levan (Göteborg University, Sweden); Dr. Bronwen Loder (HUGO, UK); Dr. David Lonsdale (John Innes Centre, UK); Dr. Philippe Lory (CNRS, France); Dr. Donna Maglott (NCBI, USA); Dr. Kathy Matthews (Indiana University, USA); Dr. Reginald O. Morgan (University of Oviedo, Spain); Dr. Daniel W. Nebert (University of Cincinnati Medical Center, USA); Dr. Rebekah Rasooly (OMIM, USA); Dr. Roger Reeves (Johns Hopkins University, USA); Dr. Kate Rice (SmithKline-Beecham Pharmaceuticals, UK); Dr. Magali Roux-Rouquie (GenAtlas, France); Dr. Philippe Sanseau (Glaxo Wellcome, UK); Dr. Fredrik Ståhl (Göteborg University, Sweden); Dr. Alan F. Scott (OMIM, USA); Mr. C. Conover Talbot (Genome Data Base, USA); Dr. Vasilis Vasiliou (University of Colorado Health Sciences Center, USA); Dr. Hester Wain (HUGO Nomenclature Committee, UK); Dr. Monte Westerfield (University of Oregon, USA); Dr. Liora Yaar (Weizmann Institute of Science, Israel).

*Note.* The following quote was provided by Jonathan Hodgkin (modified from a recent (1997) translation of *The Analects of Confucius*, by Simon Leys):

> *A disciple of Confucius once asked his Master: "If you were to be entrusted with the government of a country, what would be your first initiative?"*
>
> *Confucius replied: "It would certainly be to rectify the names . . . . If the names are not correct, language is without an object. When the language is without an object, no affair can be effected." (The Analects of Confucius (551–479 B.C.) Chap. 13, Sect. 3.)*

## ACKNOWLEDGMENTS

The Second International Nomenclature Workshop was made possible by grants from the European Union (BMH4-CT97-2031) and National Institutes of Health (HG01590). The organizers also acknowledge with gratitude the generous contributions made by *Nature Genetics*, Glaxo-Wellcome, and SmithKline-Beecham.