Highlights of the 'Gene Nomenclature Across Species' Meeting

*Elspeth A. Bruford**

Project Coordinator, HUGO Gene Nomenclature Committee (HGNC), EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, UK
*Correspondence to: E-mail: elspeth@ebi.ac.uk

Date received (in revised form): 24th February, 2010

Abstract

The first 'Gene Nomenclature Across Species' meeting was held on 12th and 13th October 2009, at the Møller Centre in Cambridge, UK. This meeting, organised and hosted by the HUGO Gene Nomenclature Committee (HGNC), brought together invited experts from the fields of gene nomenclature, phylogenetics and genome assembly and annotation. The central aim of the meeting was to discuss the issues of coordinating gene naming across vertebrates, culminating in the publication of recommendations for assigning nomenclature to genes across multiple species.

Meeting summary

The meeting began with a welcome and outline of the agenda from Elspeth Bruford, one of the meeting organisers and the group coordinator for the HUGO Gene Nomenclature Committee (HGNC). HGNC has been based at the European Bioinformatics Institute (EBI) at Hinxton, UK, since 2007. Since its inception in 1979, the HGNC has been assigning gene symbols and names to all human genes, including pseudogenes and non-coding RNAs.

The first session was chaired by Jennifer Harrow, who leads the Human and Vertebrate Analysis and Annotation (Havana) group from the Wellcome Trust Sanger Institute (WTSI), also located on the Hinxton campus. This session was devoted to introducing the three established gene nomenclature groups for mammals — namely HGNC, the Mouse Genome Nomenclature Committee (MGNC) — based at the Mouse Genome Informatics Database (MGI) at the Jackson Laboratory in Maine, USA — and the Rat Genome and Nomenclature Committee, based at the Rat Genome Database (RGD) in Milwaukee, Wisconsin, USA. Matt Wright from the HGNC, coorganiser of the meeting, kicked off by discussing the current work of the HGNC, 'An Essential Resource for the Human Genome'. Matt outlined the roles of the HGNC, including a summary of the process of symbol assignment, and its current efforts in coordinating gene naming across vertebrates. He also highlighted instances where the lack of approved gene nomenclature for most mammalian genomes has resulted in valuable published data for these species being absent or confused in the genomic databases. He was followed by Janan Eppig, principal investigator of the MGI, who, in her talk, 'What's in a Name', told us about current nomenclature issues and activities for the mouse. As well as genes, the group at MGI also name genetic markers, alleles, mutations and strains. Current efforts are focused on creating a unified gene catalogue for the mouse, by comparing gene models from the National Center for Biotechnology Information (NCBI)'s Entrez Gene database, the Ensembl database and the Havana group's Vega database. The mouse genetics community began naming genes in a standardised way long before the human community, with the first Mouse Nomenclature Guide published in 1940. In

2003, the International Committee on Standardized Genetic Nomenclature for Mice, and the Rat Genome and Nomenclature Committee agreed to unify rules and guidelines for gene, allele and mutation nomenclature in the mouse and rat. It was therefore apt that Janan was followed by Mary Shimoyama from the RGD. Mary talked about 'Nomenclature Assignment, Review and Resolution at the Rat Genome Database', starting with a discussion of the pipelines and software they have established for naming rat genes, quantitative trait loci (QTLs) and strains, and for making nomenclature updates and orthology assignments between rat, mouse and human. The state of the current rat genome assembly can prove problematic, and there is a need to establish a core consensus rat gene set in a manner similar to that of the Consensus CDS (CCDS) projects that are currently in place for the human and mouse genomes (PMID: 19498102). Other issues raised by Mary included problems with synchronising updates between databases, the need for timely adoption of RGD gene nomenclature by some databases, and the lack of requirement for authors to use standardised nomenclature in many journals.

The second session was chaired by Derek Stemple, head of the Vertebrate Development and Genetics group at the WTSI, and focused on the three further vertebrate nomenclature groups, starting with a report from Monte Westerfield, the principal investigator of the Zebrafish Model Organism Database (ZFIN). Zebrafish gene names are based on human names wherever possible, but the symbols are written in lower case to distinguish them from human gene symbols (which are in upper case letters) or mouse/rat symbols (which are lower case except for an initial upper case letter). Monte raised the important point that species-specific mutants can drive the naming of genes, such as the *oep* one-eyed pinhead gene in zebrafish, which is the orthologue of human teratocarcinoma-derived growth factor 1 (*TDGF1*) gene. The next speaker was Erik Segerdell from Xenbase, a *Xenopus laevis* and *tropicalis* resource based at the University of Calgary in Alberta, Canada. As for zebrafish, *Xenopus* gene nomenclature is identical to human where possible, and where a gene has been duplicated in *Xenopus* relative to mammals the gene symbols are appended with a numeral or letter suffix to indicate this. The newest nomenclature group, the Chicken Gene Nomenclature Committee (CGNC; PMID: 19607656), also aims to name chicken genes based on the names assigned to human genes. Alan Archibald from the Roslin Institute, Edinburgh, UK, updated us on the progress of the CGNC, which has begun its naming efforts by transferring the human gene symbols to 1:1 orthologues in chicken. To date, over 8,000 genes with a confirmed 1:1 orthologue in human have been assigned approved names by the CGNC.

After lunch, the third session turned to look at other mammalian genomes that do not have an established nomenclature group. Elizabeth Murchison from the WTSI spoke first on 'Gene Annotation and Nomenclature in Marsupials and Monotremes'. While currently they are only represented by three 'complete' genomes in the public domain (namely those for the opossum, wallaby and platypus), the important positions of these non-eutherian mammals in the vertebrate phylogeny mean that they should be able to teach us some fascinating lessons about the evolution of the mammalian genome. In most cases, marsupial and monotreme genes *do* have clear eutherian orthologues, but Elizabeth also discussed the platypus defensin genes, which have shown us that duplication of these immune genes has independently resulted in the convergent evolution of venom in both monotremes and reptiles.

Chris Elsik and Ross Tellam, the analysis leaders of the Bovine Genome Sequencing and Analysis Consortium, then told us about the 'Annotation of the Bovine Genome — the Easy and the Difficult'. This talk highlighted several common and recurring themes from the meeting: the importance of high coverage and a quality genome assembly; the necessity of producing a consensus gene set that is deposited in a centralised database (in this case the Bovine Genome Database, www.bovinegenome.org); and the need for expert input into specific groups and families of genes. As currently there are

no guidelines for assigning bovine gene symbols, of the 5,757 bovine gene models found in both Ensembl and Entrez Gene, over 60 per cent have different symbols assigned to them in each database, so, clearly, there is a need for standardising the nomenclature for this genome. Jim Reecy, the bioinformatics coordination leader of the USA's National Animal Genome Research Program, then talked to us about porcine gene annotation. To date, over 17,000 gene models have been annotated in the swine genome, of which nearly 10,000 have been projected from other species. Manual annotation, both from the Havana team at WTSI and from community annotation, is now being used to refine these gene models. Jim also mentioned the International Society of Animal Genetics (ISAG), which is an established forum for the livestock genetics community. Its genome sequence workshops could provide an excellent opportunity for gene nomenclature committees to meet. The final speaker of this session was Noelle Cockett, the sheep genome coordinator, based at Utah State University, USA, who updated us on the 'Assembly of the Ovine Whole Genome Reference Sequence'. The sheep genome is still in the early stages of assembly. There is currently a 'virtual sheep genome' available, which is based on a reorganised version of the human, dog and bovine genomes, and provides 70 per cent coverage of the ovine genome with a 0.05 per cent false positive rate. It is anticipated that the eventual Ovine Whole Genome Reference Sequence will be to a depth of 7X and will cover 95 per cent of the unique ovine genome.

Noelle was followed by a telepresentation, courtesy of Lisa Stubbs from the Kruppel Zinc Finger Catalog, based at the University of Illinois at Urbana-Champaign, USA, who discussed the 'Rapidly Evolving Transcription Factor Genes: the KRAB-Family'. She outlined the nomenclature issues raised by these complex tandem gene families that differ significantly in gene content between species. While most zinc fingers are grouped into clusters that are found in syntenic locations, lineage-specific gene duplications and losses mean that 1:1 orthologues are rare. In over 400 human genes and over 400 mouse genes, there are only around 120 sets of 1:1 orthologues, making the direct transfer of gene names between species impossible without extensive manual curation. The afternoon concluded with a lively discussion on nomenclature guidelines across species, chaired by Alan Archibald. All those present at the meeting agreed that it would be useful to have a common set of nomenclature rules that could be applied to any novel vertebrate genome, and that these would be based on human gene nomenclature but also take into account species-specific characteristics. This should prove an invaluable resource for assigning standardised gene names to newly sequenced genomes.

The next day, the proceedings began with two in-depth talks on complex gene families, following on from Lisa's presentation the previous afternoon on zinc fingers. This session was chaired by Vasilis Vasiliou from the University of Denver, Colorado, USA, an expert in the aldehyde dehydrogenase family. The first talk came from Jed Goldstone from Woods Hole Oceanographic Institution in Massachusetts, USA, who studies the evolution of the cytochrome P450 (CYP) superfamily. While there are 57 CYP genes in humans, to date, over 11,500 CYP sequences have been named across species by the CYP Gene Nomenclature Committee. This relies on the dedication of David Nelson at the University of Tennessee Health Science Center (PMID: 19951895), who individually analyses and assigns names to each sequence, and consults with other experts where necessary. The CYP nomenclature divides the genes into families (~40 per cent predicted amino acid identity cut-off) and subfamilies (~55 per cent identity cut-off) — for example, cytochrome P450 family 1, subfamily A, polypeptide 1 is *CYP1A1*. Several other established gene families, such as the aldehyde dehydrogenases (*ALDH*) and aldo-keto reductases (*AKR*) use similar rules for naming. Clear 1:1 CYP orthologues have the same names across species; but where the orthology is unclear, novel genes are given the next available number in the subfamily, which can prove complicated when dealing with incomplete genomes.

Jed was followed by Doron Lancet, the principal investigator of the Human Olfactory Data Explorer (HORDE) and GeneCards databases. Olfactory receptor (OR) genes encode seven-helix G-protein-coupled receptors and comprise the largest gene superfamily in the human genome, with a total of 855 genes. Of these, around 370 are predicted to encode functional proteins, around 60 are segregating pseudogenes (ie can encode both functional and non-functional alleles in the human population) and the remainder are pseudogenes. In humans, this superfamily has been named using a similar nomenclature system to that for the CYPs, with divisions into families and subfamilies. The HORDE database currently contains data on the human, chimp, dog, opossum and platypus olfactory receptor repertoires, and Doron showed us how these repertoires can vary significantly between vertebrate species. Nevertheless, the presence of putative ancestral OR clusters helps in the identification of orthologues between species, and hence could enable the current human nomenclature scheme to be expanded to other species in combination with expert manual curation.

The final presentation session of the meeting concentrated on multi-species databases and orthology resources, and was chaired by Ewan Birney, a senior scientist from the EBI at Hinxton, and one of the principal investigators of the Ensembl database. The first speaker was Donna Maglott from the NCBI in Bethesda, Maryland, USA, who told us about 'Naming Genes at NCBI'. The NCBI's Entrez Gene database contains data from multiple species that do not yet have a nomenclature authority. Currently names are assigned to genes in these species based on their homology to a gene with an informative name (as calculated by HomoloGene, which uses a pairwise gene comparison-based approach). Hence, the HGNC name is projected to non-rodent mammals, the MGNC name to rodents excluding rat (which is given RGD names), the ZFIN name is used across fishes, the CGNC name across birds and the XenBase name for amphibia. These assignments exclude olfactory receptors and genes from other known complex families. To date, this has allowed over 11,000 dog genes and over 12,000 chimp genes in Entrez Gene to be assigned a meaningful name automatically, based on their human orthologue in Homologene.

The next speaker was Albert Villela from the EnsemblCompara group, based at the EBI, who talked about the 'EnsemblCompara GeneTrees: Gene Orthologs and Paralogs in Ensembl'. Albert explained how EnsemblCompara produces complex gene trees that identify both orthologues and paralogues using data from all the species in Ensembl and using the longest translation of each gene. In a similar situation to that at Entrez Gene, any 1:1 orthology assignments produced by Compara are then used to project gene names from human genes to other vertebrates, excluding zebrafish and rodents.

The final speaker of the meeting was Leo Goodstadt from the MRC Functional Genomics Unit at Oxford University, UK. Leo's talk, entitled 'Accurate Inferences of Orthology Among Closely Related Species', began by outlining the different methods of predicting orthology. He stated that different phylogenetic methods often offer comparable accuracy, which can be improved by taking into account conserved gene order (syteny), and that phylogenetic inferences are mostly limited by problems with the genomic data and information content of the sequence. By looking at the human, mouse, dog, opossum, platypus and chicken genomes, he has identified a set of 9,675 1:1 orthologues. He suggested that these comprise a core, conserved non-duplicating gene set that exists between vertebrate species. This set would comprise clear candidates for easily transferring gene names between species.

The meeting concluded with a discussion chaired by David Landsman, Chief of the Computational Biology Branch of the NCBI, on how to implement gene nomenclature across species in the databases. This interesting debate concluded that coordination between databases and orthology resources is required to identify a core set of agreed 1:1 orthologues between any given species. Such a consensus set could then be candidates for automatic transferral of gene names

between species. Everyone also agreed that it is clear that some complex gene families cannot be named in an automated manner, and that expert manual curation is required and should be sought for these families.

Conclusions

The key points agreed as a result of this meeting can be summarised as follows:

- Gene nomenclature should, where possible, reflect homologous relationships across vertebrate species;
- Consensus naming, predominantly based on human gene nomenclature, has already been implemented between six vertebrate species (human, mouse, rat, chicken, zebrafish and *Xenopus*), and this effort should be expanded to other vertebrate genomes;
- Care must be taken when attempting to assign gene names in 'incomplete' genomes, and to avoid 'humanisation' of non-human genomes;
- Guidelines for the naming of genes across vertebrates should be published; these will build on current guidelines and include basic rules for the naming of paralogues;
- A list of complex gene families, which will require expert manual curation for cross-species nomenclature, should be compiled;
- Potential funding should be sought for curation of the nomenclature of these complex gene families and the construction of a database framework for superfamily nomenclature;
- The formation of novel species-specific gene nomenclature committees should be encouraged, with the aim of at least one per order for mammals;
- Automated naming efforts should initially concentrate on consensus 1:1 orthologues as identified by at least two independent and comprehensive orthology resources;
- There is a need to increase community awareness of standardised gene nomenclature, especially in journals.

Acknowledgments

The HGNC would like to acknowledge that this meeting was made possible by funding from NHGRI grant P41 HG03345 and Wellcome Trust grant 081979/Z/07/Z.