Published online 4 November 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D445–D448***
doi:10.1093/nar/gkm881

The HGNC Database in 2008: a resource for the human genome

Elspeth A. Bruford[1,*], Michael J. Lush[1], Mathew W. Wright[1], Tam P. Sneddon[2], Sue Povey[2] and Ewan Birney[1]

[1]European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA and [2]Department of Biology, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, UK

Received September 13, 2007; Revised September 28, 2007; Accepted October 1, 2007

ABSTRACT

The HUGO Gene Nomenclature Committee (HGNC) aims to assign a unique and ideally meaningful name and symbol to every human gene. The HGNC database currently comprises over 24 000 public records containing approved human gene nomenclature and associated gene information. Following our recent relocation to the European Bioinformatics Institute our homepage can now be found at http://www.genenames.org, with direct links to the searchable HGNC database and other related database resources, such as the HCOP orthology search tool and manually curated gene family webpages.

INTRODUCTION

The HUGO Gene Nomenclature Committee (HGNC) was founded in 1977 by the Human Gene Mapping community to provide a single worldwide authority to assign unique, standardized and user-friendly gene symbols to human genes. Since 1996 the HGNC has been based at University College London, UK, but in 2007 the Committee relocated to the European Bioinformatics Institute on the Wellcome Trust Genome Campus near Cambridge, UK. The website for the HGNC can now be found at http://www.genenames.org; we ask all users to update their bookmarks to this new URL as our old website at www.gene.ucl.ac.uk/nomenclature is now offline. This site provides direct links to enable users to search and download information from the HGNC database (1), which currently contains over 24 000 public gene records, or 'symbol reports'. The majority of these records represent protein-coding genes, though there are also records for pseudogenes, non-protein-coding RNA genes, phenotypes and a limited number of genomic features such as fragile sites. The primary identifier for each record is the current approved gene symbol, which is an acronym or abbreviation of the associated gene name. Each entry is also assigned a unique 'HGNC ID', which enables easy data tracking regardless of updates in the nomenclature of any given entry. Further data contained in each record include the chromosomal location of the gene, defining nucleotide sequences and publications, other symbols and names for the gene (aliases) and links to a variety of external resources.

ACCESSING THE HGNC DATABASE

The HGNC dataset can be accessed in a number of ways. Many users search and retrieve gene records using the online search facility; a simple search can be found on the new homepage at www.genenames.org. In addition an advanced search feature is located at http://www.genenames.org/cgi-bin/hgnc_search.pl, and allows the user to define up to four search terms from a variety of data fields, including approved symbol, approved name, alias symbol, alias name, previously approved name, chromosome and HGNC ID. Results can be displayed in html or text format, and sorted by approved symbol or chromosome. The HGNC dataset can also be accessed using our data downloads facility (http://www.genenames.org/data/gdlw_index.html).

Along with providing standard 'Core data', 'Core Data by Chromosome' and 'All Data' datasets, the custom downloads feature is a web-based interface that allows users to: select columns of data for output as text or html; execute limited SQL queries; generate PHP and perl code; and save searches for future reference. Two new fields have been added to the public dataset recently: 'Gene Family Name' that indicates the name of the family or families a gene has been assigned to; and 'Ensembl ID (mapped)' derived from the current build of the Ensembl database (2) and provided by the Ensembl team.

*To whom correspondence should be addressed. Tel: +44 (0)1223 494 444; Fax: +44 (0)1223 494468; Email: hgnc@genenames.org

LINKING TO THE HGNC DATABASE

It is still very easy to link directly to a specific HGNC symbol report. In line with our new domain name, URLs of the form http://www.genenames.org/data/hgnc_data.php?match=ABCA1 link directly via the approved gene symbol; however, we recommend users link directly to records using the HGNC ID, in the format http://www.genenames.org/data/hgnc_data.php?hgnc_id=29, as this will allow links to be maintained if the approved gene symbol changes. Standard symbol reports include nine fields: approved symbol, approved name, HGNC ID, status of the record ('approved', 'symbol withdrawn' for previously approved entries, or 'entry withdrawn' for entries that are no longer thought to exist), chromosomal location, previous symbols, previous names, aliases and name aliases.

LINKS TO EXTERNAL DATABASES FROM THE HGNC DATABASE

Symbol reports also contain links to established genome resources via both HGNC-curated data and mapped data provided by the external database; each field is labelled to distinguish curated from mapped data. RefSeq (3) IDs and International Nucleotide Sequence Database accessions are used to link out to GenBank (4) and the UCSC Browser and Gene Index (5). Entrez Gene (6) IDs take the user to the relevant entry in the NCBI's Gene database or Map Viewer (7). Curated PubMed (7) IDs link to specific publications in PubMed, and OMIM (8) IDs to OMIM records. Mapped UniProt (9) IDs link out to SwissProt and UniProt, and recently included Ensembl IDs take the user directly to the Ensembl GeneView (2) for the gene in question. Basic links that query external databases using the approved gene symbol are also provided at the bottom of each symbol report, and these link to GENATLAS (10), GeneCards (11), HCOP (12), GeneClinics/GeneTests (13), Vega (14) and Treefam (15).

Over the last two years, the HGNC has been actively developing reciprocal links with databases specializing in specific gene (or RNA) families or groupings. This both broadens the range of resources available to the community via our symbol report pages, and additionally provides publicity for useful resources that may otherwise be overlooked in a casual search. The majority of our specialist database links, listed in Table 1, are manually curated by the HGNC team, though some (e.g. the KZNF Gene Catalog and IUPHAR) are automatically mapped from download files provided by the specialist database.

GENE FAMILY RESOURCES

Since 2006 we have significantly expanded our resources for specific subsets of genes, either related by function, location or phenotype (gene groupings) or by sequence similarity (gene families) (see http://www.genenames.org/genefamily.html). Assignment of genes into gene families or groupings is based on sequence analyses, publications, information from specialist advisors for specific families and from other databases. For gene family members, we strongly encourage the use of a stem (or root) symbol as a basis for a hierarchical series that allows the easy identification of other related members in both database searches and the literature. HGNC currently provide over 170 manually curated webpages dedicated to individual gene families or groupings, as well as listing over 60 links to externally managed family/grouping resources. If you would like us to create webpages for other specific gene families, or include links to external gene family pages or resources, please contact us.

ORTHOLOGY RESOURCES

Orthologs are genes in different species that derive from a common ancestor and generally share the same function. The utility of standardized orthologous gene names is perhaps one of the strongest arguments for approved nomenclature and cooperation between nomenclature committees, and without this resource the analysis of genomes would be made far more difficult. We closely coordinate our efforts with the Mouse Genome Informatics (MGI) (23) Nomenclature Group and endeavour to approve the equivalent gene symbol for each human/mouse ortholog pair (e.g. human *ACOT1* and mouse *Acot1*). As part of the nomenclature assignment we research the orthology for each human gene and then add the corresponding MGI ID for the orthologous mouse gene to our database, thereby associating each human gene with its mouse ortholog. These hand-curated MGI IDs are displayed in the gene symbol report as a hyperlink direct to the relevant MGI database (23) entry.

The HGNC Comparison of Orthology Predictions search tool, HCOP (http://www.genenames.org/hcop), enables users to compare orthologs predicted for a specified human gene, or set of human genes (12,24). HCOP shows orthology predictions between human and seven other genomes (mouse, rat, chimp, dog, chicken, zebrafish and fruitfly), and currently includes data from Ensembl (2), Evola from the H-Invitational database (25), HGNC, HomoloGene (6), Inparanoid (26), MGI (23), PhIGS (27), PhyOP (28), Treefam (15) and ZFIN (29). Users can assess the reliability of the prediction from the number of these different sources that identify a particular orthologous pair. For ease of use, search terms can be either an approved symbol (e.g. *ACOT1*), a term from an approved gene name (e.g. 'thioesterase'), an Entrez Gene ID (e.g. 641371), HGNC ID or MGI ID (e.g. HGNC: 33128 or MGI: 1349396), or a RefSeq accession (e.g. NM_001037161). We recently updated HCOP to include a reciprocal orthology search link, using the Entrez Gene ID from the orthologous gene to identify human orthologs. In addition to the orthology predictions, the data returned includes the official nomenclatures, DNA sequences, database identifiers, aliases and chromosomal locations for each putative ortholog pair. We plan to expand this resource to include other species and orthology prediction databases.

Table 1. List of specialist database links in the HGNC database

Database	URL	Number of links
microRNA sequence database (miRBase) (16)	http://microrna.sanger.ac.uk/	472
Human Olfactory Receptor Data Exploratorium (HORDE) (11)	http://bioportal.weizmann.ac.il/HORDE/	857
Human Cell Differentiation Molecules (17)	http://www.hcdm.org/	363
RNA families database (Rfam) (18)	http://www.sanger.ac.uk/Software/Rfam/	62
snoRNABase (Database of human H/ACA and C/D box snoRNAs) (19)	http://www-snorna.biotoul.fr/	372
Lawrence Livermore National Laboratory Human KZNF Gene Catalog (LLNL) (20)	http://znf.llnl.gov/catalog/	462
Intermediate Filament Database	http://www.interfil.org/	69
IUPHAR Database	http://www.iuphar-db.org/	188
ImMunoGeneTics information system (IMGT) (21)	http://imgt.cines.fr/	660
MEROPS (the peptidase database) (22)	http://merops.sanger.ac.uk/	648

VARIATION RESOURCES

In recent years, it has been shown that an increasing number of genes that were originally assumed to be single copy in the human genome are actually copy number variant (CNV) between individuals. Following consultation with the research community, and to complement the introduction of these data into the major genome databases, the HGNC decided it was vital to establish a copy number variant gene nomenclature system that would be flexible, dynamic and most importantly accepted and used by the research community. Hence we are in the process of populating our database with CNV genes and associated nomenclature, using published data taken from the Database of Genomics Variants (30). To display this information in a useful and easily accessed format, we will be implementing a hierarchical structure within the HGNC database that will be public by 2008. This will allow users to link from a standard symbol report to sub-entries containing nomenclature and sequence data for each CNV copy. In addition to copy number variant genes, this new hierarchical database structure will also allow us to capture and represent information concerning other types of genomic variation, including complex allelic gene loci such as the immunoglobulins, T-cell receptors and protocadherins, and read-through/chimaeric transcripts.

FUTURE DIRECTIONS

We are planning to develop an HGNC data mining interface based on the BioMart (31) infrastructure. This will allow standalone data mining of the HGNC dataset and will be easily linked to other BioMart instances, including HapMap (32), Reactome (33) and Ensembl (2). We are also aiming to increase the proportion of curated links to external resources and welcome suggestions for further resources we could be linking to. To be notified of future developments in the HGNC database and website, please subscribe to our newsletter by emailing hgnc@genenames.org with the subject line 'subscribe'.

FEEDBACK

We welcome your feedback on any aspect of our work, including specific gene symbols and names. Please click on the 'feedback' link on our homepage to send us your comments and/or suggestions. Users can now also submit data directly to the HGNC using our online Gene Symbol Request Form (http://www.genenames.org/cgi-bin/hgnc_request.pl). This facility can be used to enquire if approved nomenclature has been assigned to a gene, to request an update in the nomenclature of a named gene, or to request nomenclature for a gene or copy number variant that currently does not yet have an approved gene nomenclature.

CITATION

Authors are requested to cite this article and the database in the following format: 'The HGNC Database, HUGO Gene Nomenclature Committee (HGNC), European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK (URL: http://www.genenames.org/)'. [Include month and year in which you retrieved the data cited.]

ACKNOWLEDGEMENTS

The HGNC is funded by the Wellcome Trust (grant 081979/Z/07/Z) and the National Human Genome Research Institute (grant P41 HG03345). E.A.B., M.J.L. and M.W.W. were previously affiliated to the Department of Biology, University College London. We would like to thank all of our collaborators and past members of the HGNC, in particular Fabrice Ducluzeau, for their invaluable help. Funding to pay the Open Access publication charges for this article was provided by the Wellcome Trust.

Conflict of interest statement. None declared.

REFERENCES

1. Eyre,T.A., Ducluzeau,F., Sneddon,T.P., Povey,S., Bruford,E.A. and Lush,M.J. (2006) The HUGO Gene Nomenclature Database, 2006 updates. *Nucleic Acids Res.*, **34**, D319–D321.

2. Hubbard,T.J., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res.*, **35**, D610–D617.
3. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
4. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2007) GenBank. *Nucleic Acids Res.*, **35**, D21–D25.
5. Kuhn,R.M., Karolchik,D., Zweig,A.S., Trumbower,H., Thomas,D.J., Thakkapallayil,A., Sugnet,C.W., Stanke,M., Smith,K.E. *et al.* (2007) The UCSC genome browser database: update 2007. *Nucleic Acids Res.*, **35**, D668–D673.
6. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **35**, D26–D31.
7. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R. *et al.* (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **35**, D5–D12.
8. McKusick,V.A. (2007) Mendelian Inheritance in Man and its online version, OMIM. *Am. J. Hum. Genet.*, **80**, 588–604.
9. UniProt Consortium (2007) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **35**, D193–D197.
10. Frezal,J. (1998) Genatlas database, genes and development defects. *C. R. Acad. Sci. III*, **321**, 805–817.
11. Safran,M., Chalifa-Caspi,V., Shmueli,O., Lapidot,M., Rosen,N., Shmoish,M., Adato,A., Peter,I. and Lancet,D. (2003) Human Gene-Centric Databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE. *Nucleic Acids Res.*, **31**, 142–146.
12. Eyre,T.A., Wright,M.W., Lush,M.J. and Bruford,E.A. (2007) HCOP: a searchable database of human orthology predictions. *Brief. Bioinformatics*, **8**, 2–5.
13. Pagon,R.A. (2006) GeneTests: an online genetic information resource for health care providers. *J. Med. Libr. Assoc.*, **94**, 343–348.
14. Ashurst,J.L., Chen,C.K., Gilbert,J.G.R., Jekosch,K., Keenan,S., Meidl,P., Searle,S.M., Stalker,J., Storey,R. *et al.* (2005) The Vertabrate Genome Annotation (Vega) database. *Nucleic Acids Res.*, **33**, D459–D465.
15. Li,H., Coghlan,A., Ruan,J., Coin,L.J., Hériché,J.K., Osmotherly,L., Li,R., Liu,T., Zhang,Z. *et al.* (2006) TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.*, **34**, D572–D580.
16. Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.
17. Zola,H., Swart,B., Nicholson,I., Aasted,B., Bensussan,A., Boumsell,L., Buckley,C., Clark,G., Drbal,K. *et al.* (2005) CD molecules 2005: human cell differentiation molecules. *Blood*, **106**, 3123–3126.
18. Griffiths-Jones,S., Moxon,S., Marshall,M., Khanna,A., Eddy,S.R. and Bateman,A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, D121–D124.
19. Lestrade,L. and Weber,M.J. (2006) snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res.*, **34**, D158–D162.
20. Huntley,S., Baggott,D.M., Hamilton,A.T., Tran-Gyamfi,M., Yang,S., Kim,J., Gordon,L., Branscomb,E. and Stubbs,L. (2006) A comprehensive catalog of human KRAB-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors. *Genome Res.*, **16**, 669–677.
21. Giudicelli,V., Duroux,P., Ginestoux,C., Folch,G., Jabado-Michaloud,J., Chaume,D. and Lefranc,M.P. (2006) IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucleic Acids Res.*, **34**, D781–D784.
22. Rawlings,N.D., Morton,F.R. and Barrett,A.J. (2006) MEROPS: the peptidase database. *Nucleic Acids Res.*, **34**, D270–D272.
23. Eppig,J.T., Blake,J.A., Bult,C.J., Kadin,J.A. and Richardson,J.E. Mouse Genome Database Group (2007) The mouse genome database (MGD): new features facilitating a model system. *Nucleic Acids Res.*, **35**, D630–D637.
24. Wright,M.W., Eyre,T.A., Lush,M.J., Povey,S. and Bruford,E.A. (2005) HCOP: the HGNC comparison of orthology predictions search tool. *Mamm. Genome*, **16**, 827–828.
25. Imanishi,T., Itoh,T., Suzuki,Y., O'Donovan,C., Fukuchi,S., Koyanagi,K.O., Barrero,R.A., Tamura,T., Yamaguchi-Kabata,Y. *et al.* (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.*, **2**, 856–875.
26. O'Brien,K.P., Remm,M. and Sonnhammer,E.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.*, **33**, D476–D480.
27. Dehal,P.S. and Boore,J.L. (2006) A phylogenomic gene cluster resource: the Phylogenetically Inferred Groups (PhIGs) database. *BMC Bioinformatics*, **7**, 201.
28. Goodstadt,L. and Ponting,C.P. (2006) Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS Comput. Biol.*, **2**, e133.
29. Sprague,J., Bayraktaroglu,L., Clements,D., Conlin,T., Fashena,D., Frazer,K., Haendel,M., Howe,D.G., Mani,P. *et al.* (2006) The Zebrafish Information Network: the zebrafish model organism database. *Nucleic Acids Res.*, **34**, D581–D585.
30. Zhang,J., Feuk,L., Duggan,G.E., Khaja,R. and Scherer,S.W. (2006) Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. *Cytogenet. Genome Res.*, **115**, 205–214.
31. Kasprzyk,A., Keefe,D., Smedley,D., London,D., Spooner,W., Melsopp,C., Hammond,M., Rocca-Serra,P., Cox,T. *et al.* (2004) EnsMart: a generic system for fast and flexible access to biological data. *Genome Res.*, **14**, 160–169.
32. Thorisson,G.A., Smith,A.V., Krishnan,L. and Stein,L.D. (2005) The International HapMap Project Web site. *Genome Res.*, **15**, 1592–1593.
33. Vastrik,I., D'Eustachio,P., Schmidt,E., Joshi-Tope,G., Gopinath,G., Croft,D., de Bono,B., Gillespie,M., Jassal,B. *et al.* (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.*, **8**, R39.