Nucleic Acids Research, 2006, Vol. 34, Database issue ***D319–D321***
doi:10.1093/nar/gkj147

The HUGO Gene Nomenclature Database, 2006 updates

Tina A. Eyre, Fabrice Ducluzeau, Tam P. Sneddon, Sue Povey, Elspeth A. Bruford and Michael J. Lush*

Department of Biology, HUGO Gene Nomenclature Committee (HGNC), University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, UK

Received September 8, 2005; Revised and Accepted October 28, 2005

ABSTRACT

The HUGO Gene Nomenclature Committee (HGNC) aims to give every human gene a unique and ideally meaningful name and symbol. The HGNC database, previously known as Genew, contains over 22 000 public records with approved human gene nomenclature and associated information. The database has undergone major improvements throughout the last year, is publicly available for online searching at http://www.gene.ucl.ac.uk/cgi-bin/nomenclature/searchgenes.pl and has a new custom downloads interface at http://www.gene.ucl.ac.uk/cgi-bin/nomenclature/gdlw.pl.

OVERVIEW

The HUGO Gene Nomenclature Committee (HGNC) maintains a database of unique and approved human gene names and symbols (1). Current estimates predict the total number of protein coding human genes as 20 000–25 000 (2,3), and over 18 000 of these now have been assigned HGNC approved nomenclature. We also assign nomenclature to other specific features such as fragile sites and disease loci inferred by linkage. This nomenclature is hand-curated and represents the gold standard, to be used in all publications and databases where a specific gene is discussed or referenced.

HGNC data can be accessed in two main ways. First, for specific online searches the HGNC database search engine, Searchgenes, is available at http://www.gene.ucl.ac.uk/cgi-bin/nomenclature/searchgenes.pl with both simple and advanced search options. Second, custom downloads are available, allowing the user to download large volumes of data in their own preferred format using our custom download script (http://www.gene.ucl.ac.uk/cgi-bin/nomenclature/gdlw.pl).

The HGNC database migrated from Microsoft Access to PostgreSQL (http://www.postgresql.org/) at the end of March 2005. This change has meant not only easier curation for the database editors and greatly improved quality control checking, but also increased search speed and flexibility for both editors and users. In addition, custom downloads are now available to the public, allowing retrieval of precise sets of genes and data about those genes.

IMPROVEMENTS SINCE 2004

Renaming the database

Previously the HGNC database was referred to as Genew (1); however, following the change from Microsoft Access to PostgreSQL in March 2005 it was decided to change this to the easily recognized name of the 'HGNC Database'. The term Genew was little known and this move seemed more in line with our policy for assigning unique and meaningful nomenclature. HGNC identification numbers, the unique identifiers associated with each gene record in the HGNC database, should now be referred to using the HGNC: prefix. This syntax has been adopted by all the major genome databases that display HGNC data, including Entrez Gene (4), Ensembl (5) and GeneCards (6).

Database editing

The HGNC database is implemented in PostgreSQL version 8.03. It consists of 28 tables containing in total over 500 000 records. The database now integrates public and confidential data, submitted to the HGNC by independent researchers and from more large-scale projects, such as the Human Genome Sequencing Consortium. This includes the results of our custom BLAST server, making 200 000 sequences searchable and inter-linked with HGNC gene records.

Quality control checking is used to enforce formats on the data entered and to check its integrity, and can now be performed on various levels. First, the database checks for invalid formats or missing required data when an editor attempts to save a modified record. Second, scripts are used to error check records containing newly approved nomenclature prior to release. If an error is found, that record is held back from

*To whom correspondence should be addressed. Tel: +44 20 7679 7410; Fax: +44 20 7387 3496; Email: nome@galton.ucl.ac.uk

release into the public domain and the editor responsible is automatically notified. Third, all data are regularly monitored and any inconsistencies are listed on a quality control web page.

The HGNC editors are now able to curate the database remotely, using a web-based editing tool on a secure server using SSL encryption. All transactions are logged providing an audit trail and SQL triggers are now used to automatically add certain details to the gene records, such as logging the name of the editor and the date on which modifications were made.

Online improvements

The HGNC database front-end and editor are web-based and written in Perl. The HTML::Template perl module is used to allow rapid generation of complex data editing and viewing forms containing multiple gene records from simple repeating units. In addition, special purpose forms can be rapidly generated to support new projects or new applications of HGNC data.

Both Searchgenes and the Symbol Report Form results format have been given a new look using new website templates developed in Macromedia Dreamweaver MX2004. It is now very easy to link to a particular Symbol Report Form via either the HGNC ID or the approved symbol, using URLs such as http://www.gene.ucl.ac.uk/nomenclature/data/get_data.php?hgnc_id=HGNC:29 or http://www.gene.ucl.ac.uk/nomenclature/data/get_data.php?app_sym=ABCA1.

Linking by HGNC ID is preferred and is more reliable in the long term, since HGNC IDs are constant for any given gene whereas approved symbols may change. When one entry has been merged into another entry, the merged entry remains in the database with 'Symbol Withdrawn' status, the text ~withdrawn is added to the symbol and the gene name is replaced with text indicating the entry it has been merged into. On rare occasions when an entry is split, the original HGNC ID remains associated with the most appropriate entry.

Custom data downloads—basic use

Predefined downloads of HGNC data are now available from our custom downloads page (http://www.gene.ucl.ac.uk/nomenclature/data/gdlw_index.html) in both plain text and HTML formats. The previously available static file downloads have been phased out, and the new system has been shown to be more convenient and flexible, and includes improved documentation. A variety of data are available, including approved gene symbol and name, literature and database aliases, chromosomal location, sequence accession numbers and a gene family name (where applicable). Links to relevant entries in other databases, such as Ensembl (5), GENATLAS (7), GeneCards (6), GeneClinics/GeneTests (8), IMGT (9), Entrez Gene (4), MGD (10), PubMed (11), OMIM (11), RefSeq (11), Swiss-Prot (12), UCSC (13) and Vega (14) are also provided.

A particularly important functionality of the custom downloads pages is that the results are generated dynamically so that they are up-to-date whenever the user returns to the saved URL. However, the URL also encodes the format of the data, so that this will be preserved as the database develops and new fields are added.

Custom data downloads—advanced use

More advanced users may use the script directly (http://www.gene.ucl.ac.uk/cgi-bin/nomenclature/gdlw.pl) to select custom views of HGNC data using simple SQL 'WHERE' clauses. This enables data for a particular group of genes to be displayed. The data returned may also be limited by chromosome. Documentation for this feature is available at http://www.gene.ucl.ac.uk/nomenclature/data/gdlw_patmatch.html.

Users may specify the output format of their searches. The 'HTML' option will give a simple HTML table of results with hyperlinks to the HGNC gene symbol reports, as well as to a limited set of relevant entries in external databases. The 'Gene Report Table' format produces a series of tables, each containing data for a single gene with more links. The 'Text' output format is particularly useful for downloading data into a tab-delimited file that may be processed further, injected into other databases or viewed in spreadsheet programs. A valuable debugging option when using the WHERE field is the 'Show SQL' output option which displays the SQL query without executing it.

Users can directly include a particular table of data within their own web pages by using use the 'PHP Code' output option to generate code to be embedded in a PHP document (http://www.php.net/). This technique is used to generate dynamically updated Gene Family Report pages (e.g. http://www.gene.ucl.ac.uk/nomenclature/genefamily/abc.php). Finally, the 'Perl Code' format generates a snippet of code that uses the LWP::Simple module to download the data specified in that search. This option facilitates automatic downloads of HGNC data. Again, the format of the results is specified by the code and will be maintained even when modifications to the database structure are made.

USAGE OF THE HGNC DATABASE

The HGNC custom downloads script received 506 000 hits between January 1 and June 30, 2005, an average of 2800 per day (excluding queries made by HGNC staff and major web crawlers). Searchgenes was queried 290 000 times in this same period.

Nearly all (99%) of our custom downloads users make use of the WHERE clause functionality, rather than downloading the entire data set. Of them 41% selected a plain-text output and 59% requested the Gene Report output, suggesting that the download script is frequently being used as an application program interface (API) to serve specific subsets of HGNC data to external applications. Consistent with this, the most popular searches were for single records specified by HGNC ID (78%) or approved symbol (18%).

Multiple gene records can be returned using inexact query terms with the keywords 'LIKE' or 'ILIKE' or with the 'IN' keyword to identify records matching a list of queries. Less than 1% of searches used these inexact terms, again suggesting the use of the download script as an API. It seems useful to point out that these inexact queries are valuable for concurrently downloading, viewing or linking to a set of records of interest, such as those belonging to a particular group of genes.

FUTURE DIRECTIONS

In the near future the HGNC website will provide an online form for direct submission of sequences to the database to

streamline the flow of data. In addition, Searchgenes will be superseded with an improved search facility, new fields, such as Name Aliases, and further fields, such as locus type, which are currently only available in the downloadable dataset.

CONCLUSIONS

The developments described here have provided much needed automation and opened the way for continued improvements in database flexibility and agility. As a result, the HGNC database is now far more able to respond to the needs of both its editors and the community.

CITATION

Authors are requested to cite this article and the database in the following format: 'The HGNC Database, HUGO Gene Nomenclature Committee (HGNC), Department of Biology, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, UK (URL: http://www.gene.ucl.ac.uk/cgi-bin/nomenclature/searchgenes.pl)'. [Include month and year in which you retrieved the data cited.]

ACKNOWLEDGEMENTS

Many thanks to the HGNC editors Drs Varsha Khodiyar, Ruth Lovering, Kate Sneddon, Mathew Wright, and Connie Talbot Jr, whose accurate curation and attention to detail ensure the validity of the gene records. The work of the HGNC is supported by NHGRI grant P41 HG003345, the UK Medical Research Council and the Wellcome Trust. Funding to pay the Open Access publication charges for this article was provided by JISC.

Conflict of interest statement. None declared.

REFERENCES

1. Wain,H.M., Lush,M.J., Ducluzeau,F., Khodiyar,V.K. and Povey,S. (2004) Genew: the Human Gene Nomenclature Database, 2004 updates. *Nucleic Acids Res*., **32**, D255–D257.
2. Larsson,T.P., Murray,C.G., Hill,T., Fredriksson,R. and Schioth,H.B. (2005) Comparison of the current RefSeq, Ensembl and EST databases for counting genes and gene discovery. *FEBS Lett*., **579**, 690–698.
3. International Human Genome Sequencing Consortium. (2004), Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.
4. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*., **33**, D54–D58.
5. Hubbard,T., Andrews,D., Caccamo,M., Cameron,G., Chen,Y., Clamp,M., Clarke,L., Coates,G., Cox,T., Cunningham,F. *et al*. (2005) Ensembl 2005. *Nucleic Acids Res*., **33**, D447–D453.
6. 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.
7. Frezal,J. (1998) Genatlas database, genes and development defects. *C. R. Acad. Sci. III*, **321**, 805–817.
8. Pagon,R.A., Tarczy-Hornoch,P., Baskin,P.K., Edwards,J.E., Covington,M.L., Espeseth,M., Beahler,C., Bird,T.D., Popovich,B., Nesbitt,C. *et al*. (2002) GeneTests-GeneClinics: genetic testing information for a growing audience. *Hum. Mut*., **19**, 501–509.
9. Lefranc,M.-P. (2003) IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res*., **31**, 307–310.
10. Eppig,J.T., Bult,C.J., Kadin,J.A., Richardson,J.E., Blake,J.A., Anagnostopoulos,A., Baldarelli,R.M., Baya,M., Beal,J.S., Bello,S.M. *et al*. (2005) The Mouse Genome Database (MGD): from genes to mice—a community resource for mouse biology. *Nucleic Acids Res*., **33**, D471–D475.
11. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S., Helmberg,W. *et al*. (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **33**, D39–D45.
12. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*., **31**, 365–370.
13. Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J. *et al*. (2003) The UCSC Genome Browser Database. *Nucleic Acids Res*., **31**, 51–54.
14. Ashurst,J.L., Chen,C.K., Gilbert,J.G.R., Jekosch,K., Keenan,S., Meidl,P., Searle,S.M., Stalker,J., Storey,R., Trevanion,S. *et al*. (2005) The Vertabrate Genome Annotation (Vega) database. *Nucleic Acids Res*., **33**, D459–465.