Nucleic Acids Research, 2004, Vol. 32, Database issue ***D255–D257***
DOI: 10.1093/nar/gkh072

Genew: the Human Gene Nomenclature Database, 2004 updates

Hester M. Wain*, Michael J. Lush, Fabrice Ducluzeau, Varsha K. Khodiyar and Sue Povey

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

Received September 15, 2003; Revised and Accepted September 30, 2003

ABSTRACT

Genew, the Human Gene Nomenclature Database http://www.gene.ucl.ac.uk/cgi-bin/nomenclature/searchgenes.pl is the only resource that provides data for all human genes that have approved symbols. It is managed by the HUGO Gene Nomenclature Committee (HGNC) as a confidential database, containing over 22 000 records, 75% of which are represented online by a publicly searchable text file. Since 2002, there have been significant improvements to the Genew search engine. Additionally we have increased our capacity to analyse confidential sequence data, which has enabled us to manage the large numbers of gene symbol requests that we receive from the chromosome sequencing consortia.

OVERVIEW

The Genew database (1) is the primary resource for approved gene symbols for all other human genetic databases. We exchange information with many databases and organizations throughout the world to update new gene symbols and encourage their use.

IMPROVEMENTS SINCE 2002

New search engine

The new version of the Genew search engine was made available in 2002. This can be found at the same URL: http://www.gene.ucl.ac.uk/cgi-bin/nomenclature/searchgenes.pl and now provides direct links from the search results to individually curated gene records. Both quick and advanced search options are available, with 93% of users opting for the quick gene search option, indicating that this resolves most user queries. However, the advanced search options can be very useful in resolving more complex queries. We have significantly increased the variety of search terms, so now any term within the data file searchdata.txt can be used. This file is available directly online (http://www.gene.ucl.ac.uk/public-files/nomen/searchdata.txt) and by FTP (http://www.gene.ucl.ac.uk/nomenclature/code/ftpaccess.html).

Each online gene record contains 23 fields, with 14 links to other relevant resources including: Ensembl (2), GENATLAS (3), GeneCards (4), GeneClinics/GeneTests (http://www.genetests.org), the international ImMunoGeneTics database® (IMGT) (5), LocusLink (6), MGD (7), OMIM (8), Ref_Seq (6) and Swiss-Prot (9).

Each gene record is available by querying either the approved gene symbol or the HGNC ID number, thus enabling other databases to link directly to the Genew record, even if the symbol changes. For example the gene record for CFTR, using the approved symbol, is at URL: http://www.gene.ucl.ac.uk/cgi-bin/nomenclature/get_data.pl?match=CFTR and using the HGNC ID number is at URL: http://www.gene.ucl.ac.uk/cgi-bin/nomenclature/get_data.pl?hgnc_id=1884.

The new Genew search engine has received a total of 422 113 hits (since July 2002), with an average of 31 038 hits per month. Table 1 gives an indication of how many of these hits are followed by searches of the database.

We also monitor the top 20 search terms used, as this assists us in developing both a more user-friendly search engine and a better understanding of commonly used (but possibly not approved) gene symbols. Table 2 shows the total number of searches for the top 20 search terms and their approved symbols (which are the same in all bar one case: TP53 is the approved symbol for 'p53').

Non-human orthologues

With increased requests for gene symbols in other species, we have added a new gene status, 'Approved Non-Human'. This currently includes 98 entries that we have approved in order to maintain the orthologous symbol in the human gene family series. It is quite likely that most of these genes will ultimately be found in the human genome. Each 'Approved Non-Human' gene symbol has links to the appropriate non-human sequence accession ID where possible. The orthologous species currently include: mouse, cow, rat, African clawed toad, pig, zebrafish and dog.

LocusLink updates

In order to update correctly the LocusLink entries with approved gene symbols we have added a new field designated

*To whom correspondence should be addressed. Tel: +44 20 7679 5027; Fax: +44 20 7387 3496; Email: nome@galton.ucl.ac.uk

Table 1. Use of the Genew search engine (July 2002–August 2003)

Genew search engine	Number
Searches run	353 581
Search terms used	62 454
Individual users (measured by IP address)	60 862

Table 2. Top 20 search terms used in the Genew search engine

Search term	Approved symbol	Total number of searches
CFTR	CFTR	1482
TP53	TP53	1048
BRCA1	BRCA1	1035
SLC	SLC[a]	890
BCL2	BCL2	742
TNF	TNF[a]	740
HD	HD	677
ABCB1	ABCB1	635
ABC	ABC[a]	623
MYC	MYC	579
GPR	GPR[a]	561
HBB	HBB	539
SPG	SPG[a]	534
APC	APC	531
ABCA1	ABCA1	506
p53	TP53	490
APOE	APOE	478
CDKN2A	CDKN2A	462
HFE	HFE	461
ERBB2	ERBB2	448

[a]Indicates root symbol for a gene family.

'Locus Type'. This includes designations such as:
(i) gene with no protein product;
(ii) model, supported by EST alignments;
(iii) phenotype only;
(iv) pseudogene;
(v) RNA, ribosomal.

Genew updates are exported twice a week as the text file: http://www.gene.ucl.ac.uk/public-files/nomen/ncbi2.txt, which is automatically imported into the LocusLink database.

Confidential gene records

Unnamed genes are placed into the confidential section of Genew (known previously as 'pending'). This includes those genes that have been submitted by authors and/or journals for symbol approval prior to publication. In addition, we have further increased this resource with unnamed genes from two major public data sets: the 'Interim' human genes from LocusLink and the interim mouse genes from MGD which are updated once a week. There are now just over 3000 unnamed gene records awaiting approval.

Downloads/FTP

A variety of files is available online or via FTP from: http://www.gene.ucl.ac.uk/public-files/nomen/. These include chromosome-specific files with any nomenclature changes highlighted.

GENEW UPGRADE

We have been working towards transferring Genew to PostgreSQL and creating a more dynamic web interface. However, the large numbers of symbol requests from chromosome sequencing consortia have altered our priorities, so in the last year we have focused our bioinformatics resources on a more comprehensive sequence database termed LBlast.

LBlast

Our LBlast database system comprises a set of Perl scripts that provide active maintenance of sequence annotation and automatic sequence importation into the LBlast database, thus reflecting sequence additions to the Genew database on an ongoing basis from three diverse sources of confidential sequence data:

(i) raw sequence data from Genew records (4608 DNA and 1660 protein sequences);
(ii) sequence accession numbers from Genew records (28 771 sequences);
(iii) raw sequence data from Editors and chromosome projects (24 110 sequences).

Each gene sequence is now tracked via a unique HGNC sequence accession number (HSeq), which is added to the confidential gene record. The LBlast system has been set up in such a way that any sequence used to search the database is immediately assigned an HSeq ID and added to user_contrib, which consists of sequences that have been searched against the database in the previous 4 weeks. Thus, the submitted sequences are added to the LBlast database before the BLAST (10) search is run, allowing duplicate submissions to be identified immediately.

Sequence analysis

All sequences submitted to the HGNC are analysed initially using NCBI's BLAST. This searches our confidential sequences, sequence data imported from LocusLink, the non-redundant DNA and protein sequences and patent sequences [from GenBank (10) and EMBL (11)]. In addition, all sequences are also analysed for the presence of domains and motifs via InterProScan (12). All InterProScan and BLAST results are stored permanently in the database.

The LBlast sequence data are managed in a PostgreSQL database (http://www.postgresql.org/), via a collection of Perl scripts (http://www.perl.com/) using BioPerl (http://bioperl.org/) with a PHP interface (http://www.php.net). This has been developed with the intention of adding the Genew interface at a later date.

Our capacity to process sequence data increased significantly in 2003 with the development and installation of our Beowulf Cluster. The cluster contains 16 Athlon MP 2000+ CPUs, 32 Gb of RAM and 520 Gb of disk space, and enables us to process 500 LBlast searches, or 37 InterProScans, an hour. Previously, our Sun E250 could only manage one or two LBlast searches an hour and was unable to complete InterProScans in a reasonable time. Details of the cluster construction will be available from our website http://www.gene.ucl.ac.uk/nomenclature/ by January 2004.

IMPLEMENTATION

Genew is currently implemented in the Microsoft Access 97 relational database management system. The database consists of 13 tables containing over 170 fields and 22 000 gene records.

The Genew search engine, http://www.gene.ucl.ac.uk/cgi-bin/nomenclature/searchgenes.pl, is based on a Perl front-end querying a PostgreSQL database, derived from text files exported from the off-line database.

CITATION

Authors are requested to cite this article and the database in the following format: 'Genew, 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 Elspeth Bruford, Ruth Lovering, Mathew Wright and Connie Talbot Jr whose accurate curation and attention to detail ensure the validity of the gene records. The HGNC is supported by NIH contract N01-LM-9-3533 and by the UK Medical Research Council.

REFERENCES

1. Wain,H.M., Lush,M., Ducluzeau,F. and Povey,S. (2002) Genew: The Human Nomenclature Database. *Nucleic Acids Res.*, **30**, 169–171.
2. Clamp,M., Andrews,D., Barker,D., Bevan,P., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V. *et al.* (2003) Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.*, **31**, 38–42.
3. Frezal,J. (1998) Genatlas database, genes and development defects. *C. R. Acad. Sci. III*, **321**, 805–817.
4. 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.
5. Lefranc,M.-P. (2003) IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res.*, **31**, 307–310.
6. Pruitt,K.D. and. Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.
7. Blake,J.A., Richardson,J.E., Bult,C.J., Kadin,J.A., Eppig,J.T.; Mouse Genome Database Group (2003) MGD: the Mouse Genome Database. *Nucleic Acids Res.*, **31**, 193–195.
8. Wheeler,D.L., Church,D.M., Federhen, S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E., Tatusova,T.A. *et al.* (2003) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **31**, 28–33.
9. 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.
10. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
11. Stoesser,G., Baker,W., van den Broek,A., Garcia-Pastor,M., Kanz,C., Kulikova,T., Leinonen,R., Lin,Q., Lombard,V., Lopez,R. *et al.* (2003) The EMBL Nucleotide Sequence Database: major new developments. *Nucleic Acids Res.*, **31**, 17–22.
12. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P., Bucher,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.