Published online 17 November 2012 *Nucleic Acids Research, 2013, Vol. 41, Database issue* **D545–D552**
doi:10.1093/nar/gks1066

Genenames.org: the HGNC resources in 2013

Kristian A. Gray*, Louise C. Daugherty, Susan M. Gordon, Ruth L. Seal, Mathew W. Wright and Elspeth A. Bruford

HUGO Gene Nomenclature Committee, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK

Received September 19, 2012; Accepted October 11, 2012

ABSTRACT

The HUGO Gene Nomenclature Committee situated at the European Bioinformatics Institute assigns unique symbols and names to human genes. Since 2011, the data within our database has expanded largely owing to an increase in naming pseudogenes and non-coding RNA genes, and we now have >33 500 approved symbols. Our gene families and groups have also increased to nearly 500, with ∼45% of our gene entries associated to at least one family or group. We have also redesigned the HUGO Gene Nomenclature Committee website http://www.genenames.org creating a constant look and feel across the site and improving usability and readability for our users. The site provides a public access portal to our database with no restrictions imposed on access or the use of the data. Within this article, we review our online resources and data with particular emphasis on the updates to our website.

INTRODUCTION

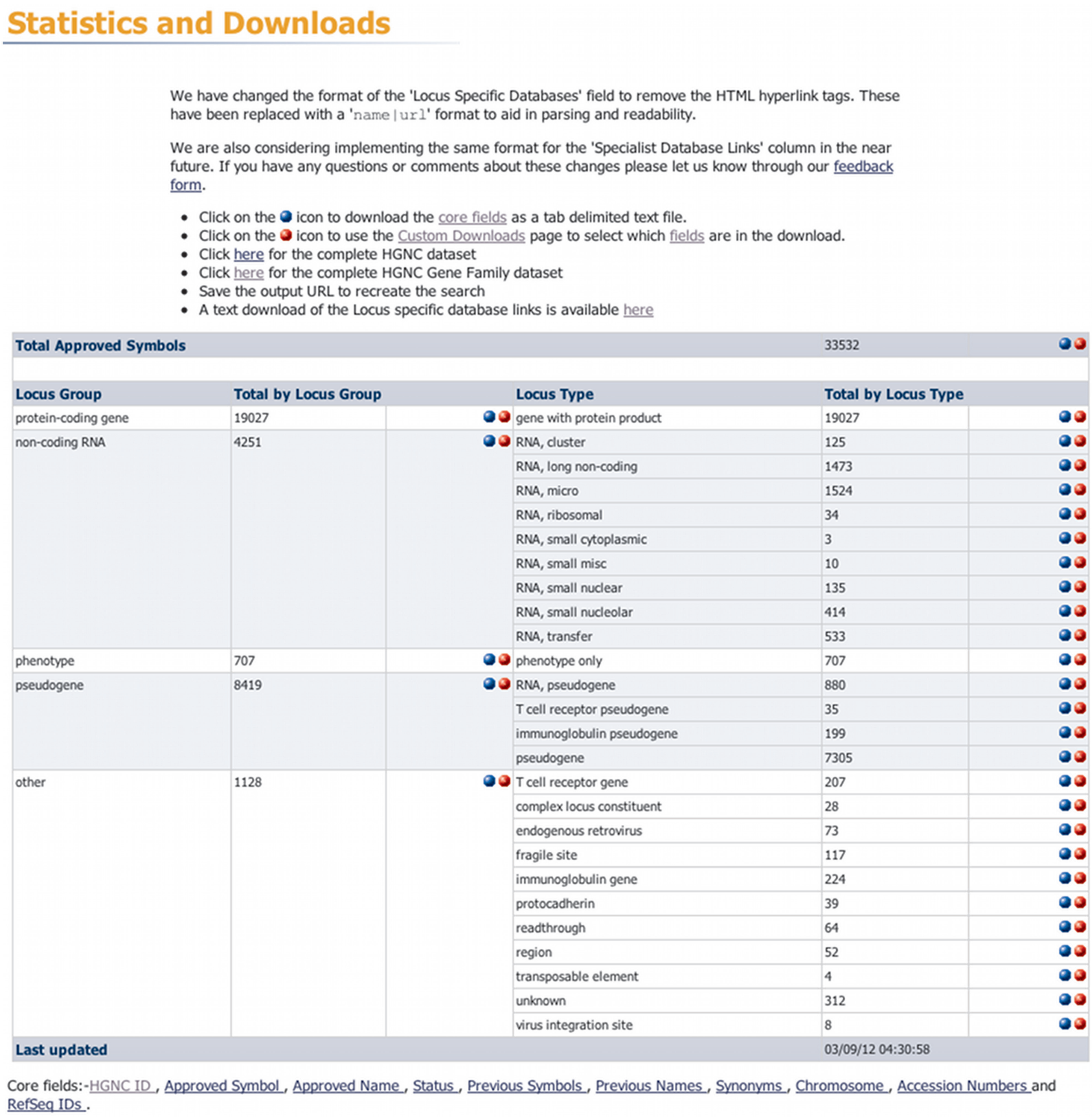
The HUGO Gene Nomenclature Committee (HGNC) maintains a publicly accessible database of unique and approved human names and symbols for protein coding genes and other features found within the human genome (1). The HGNC arose from the Human Gene Mapping community recognizing the need for a single committee with the authority to approve human gene nomenclature in 1977; in 2007, 30 years and >25 000 gene symbols later, the HGNC moved to its current location within the European Bioinformatics Institute on the Wellcome Trust Genome Campus south of Cambridge, UK.

All HGNC entries are manually curated, and the HGNC symbols and names assigned represent a standard, to be used in all publications and databases where a specific gene is discussed or referenced. Our aim is to provide nomenclature that is informative and acceptable to researchers in the field. To accomplish this, we contact researchers working on particular genes and gene families before approving symbols. We also encourage researchers to submit proposals for gene symbols to determine their suitability before publication, and these are considered under strict confidentiality. We also work closely with other nomenclature groups such as those for mouse (2), rat (3) and zebrafish (4) to ensure that orthologous genes are assigned equivalent symbols where possible. Our symbols are used extensively within the human genetics and genomics communities and throughout the databases that concentrate on human genes and proteins, such as Entrez Gene (5), Ensembl (6), Vega (7), ENA/GenBank/DDBJ (8–10), GeneCards (11), UCSC genome browser (12) and UniProt (13), as well as disease-related databases such as Decipher (14), OMIM (15), and COSMIC (16).

DATA

As of the beginning of September 2012, we have 33 532 active entries within our database with 19 027 of the entries being protein coding genes (Figure 1). In May 2003, at Cold Spring Harbor Laboratories, winners were declared for GeneSweep, a competition to estimate how many protein coding genes were present in human, which concluded that <30 000 genes would be found (17). In the intervening years, improvements in genome annotation have led to progressively lower estimates. The Consensus CDS (CCDS) project (18), dedicated to identifying a core set of protein coding regions, contains 18 474 gene IDs as of August 2012. Comparing our figure of 19 027 protein coding genes with that of the CCDS project, we believe we have an almost complete set for the human genome. Our entries are constantly being reviewed and updated with extra information, name and symbol changes and locus type classification changes. We have found that the number of protein coding genes has plateaued, as novel genes have been offset by revision and reclassification of existing entries to non-coding locus types and withdrawal of redundant entries. We also have at present ∼300 entries

*To whom correspondence should be addressed. Tel: +44 1223 496967; Fax: +44 1223 494468; Email: hgnc@genenames.org



Statistics and Downloads

We have changed the format of the 'Locus Specific Databases' field to remove the HTML hyperlink tags. These have been replaced with a 'name|url' format to aid in parsing and readability.

We are also considering implementing the same format for the 'Specialist Database Links' column in the near future. If you have any questions or comments about these changes please let us know through our feedback form.

- Click on the ● icon to download the core fields as a tab delimited text file.
- Click on the ● icon to use the Custom Downloads page to select which fields are in the download.
- Click here for the complete HGNC dataset
- Click here for the complete HGNC Gene Family dataset
- Save the output URL to recreate the search
- A text download of the Locus specific database links is available here

Total Approved Symbols				33532	●●
Locus Group	**Total by Locus Group**		**Locus Type**	**Total by Locus Type**	
protein-coding gene	19027	●●	gene with protein product	19027	●●
non-coding RNA	4251	●●	RNA, cluster	125	●●
			RNA, long non-coding	1473	●●
			RNA, micro	1524	●●
			RNA, ribosomal	34	●●
			RNA, small cytoplasmic	3	●●
			RNA, small misc	10	●●
			RNA, small nuclear	135	●●
			RNA, small nucleolar	414	●●
			RNA, transfer	533	●●
phenotype	707	●●	phenotype only	707	●●
pseudogene	8419	●●	RNA, pseudogene	880	●●
			T cell receptor pseudogene	35	●●
			immunoglobulin pseudogene	199	●●
			pseudogene	7305	●●
other	1128	●●	T cell receptor gene	207	●●
			complex locus constituent	28	●●
			endogenous retrovirus	73	●●
			fragile site	117	●●
			immunoglobulin gene	224	●●
			protocadherin	39	●●
			readthrough	64	●●
			region	52	●●
			transposable element	4	●●
			unknown	312	●●
			virus integration site	8	●●
Last updated				03/09/12 04:30:58	

Core fields:-HGNC ID , Approved Symbol , Approved Name , Status , Previous Symbols , Previous Names , Synonyms , Chromosome , Accession Numbers and RefSeq IDs .

Figure 1. The statistics and downloads page allows users to retrieve plain text, tab-separated data sets. The user may download the entire gene symbol report data set, gene family sets or download data that belong to a certain locus type. The table on the page shows the total number of entries per locus group and type including all of the non-coding RNA locus types and the number of entries associated to them.

with a locus type of 'unknown'. Entries fall in this category when annotation groups disagree on the coding potential of the gene in question. These entries are reviewed regularly, and we are very interested in hearing about new evidence to prove or disprove the coding potential of these genes.

The main areas of growth within our database in the past 2 years have been in naming non-coding RNA genes (19) and pseudogenes, and classifying genes into families and groups. At the start of September 2012, we had 4251 non-coding RNAs, and this figure is expected to increase as more long non-coding RNAs (>200 nucleotides) are annotated (Figure 1). Within our database, the classification of non-coding RNA is a 'locus group' that contains several 'locus types'. Since 2011 (20), we have changed our non-coding RNA locus types to better reflect sequence ontology (21). A full list of current non-coding RNA types can be seen in Figure 1. Pseudogenes account for almost a quarter of our total entries and are the second largest group of genes with approved nomenclature.

The vast majority of pseudogenes are non-functional, but the analysis of these genes can be extremely important for insight into the evolution of gene families, and pseudogenes are discussed frequently in literature; so it is important for us to assign meaningful symbols and names to this class of gene. We have introduced a new addition to our locus type classification, 'readthrough', in which we currently have 64 entries. This class contains loci such as INS-IGF2 (22) where transcription goes beyond the normal termination sequence of a gene (i.e. INS, insulin), and extends into an adjacent gene [i.e. IGF2, insulin growth factor 2 (somatomedin A)].

WEBSITE

Our website www.genenames.org provides a public access portal to our database. There are no restrictions imposed on access or the use of the data provided by the HGNC. In May 2011, we released a new design for our website to give every page a consistent look and feel throughout the site and to improve user experience and interactions. The new design was mainly built using the Drupal content management system for static pages, and Perl common gateway interface (CGI) for our dynamic content that retrieves data from our MySQL database. Drupal offers us many advantages, the main one being that our curators can add and change content easily and efficiently without needing to learn how to build web pages using HTML and so on. The Perl CGI pages use many comprehensive perl archive network (CPAN) modules especially to connect to the database and to create HTML templates. Using templates separates the design and presentation elements away from the core Perl code, which aids development and maintenance. Owing to the consistent design across the site, users will not notice the difference between Drupal and Perl CGI web pages. We are still looking to update sections of our website and are keen to hear from our users to improve their experience in navigating our site. We shall now explore the new design and the main pages of the site that make genenames.org.

Header and footer

Every page of www.genenames.org has a new header/banner that contains everything you will need to browse through the website. One of the main improvements to the website was the addition of the drop-down menus attached to the tabs (Figure 2) in the header. These allow us to break down the sets of pages that we provide and allow the user to find the page they are looking for without clicking through multiple index pages. To activate the drop-down menus, the user need only hover over a tab, and a column of links will appear. The menus are created using cascading style sheets only, and so if the user prefers to disable JavaScript, the menus will still work. The new footer also uses the same colour scheme as the header and appears on every page. It contains links to our terms of use, our privacy and cookies policy and an email link so that the user can email us with any queries about the site or data contained within (Figure 2).

Home page

There are four sections to the main body of our new home page. The four sections are: a 'browse-approved symbols by chromosome' interactive karyotype image, linking to a 'Statistics and Downloads' page specifically for each chromosome (see 'Downloads' section); a 'Quick Gene Search'; a 'latest news' section including a link to our new HGNC Twitter feed (@genenames); and a frequently asked questions portal and website search.

Gene search

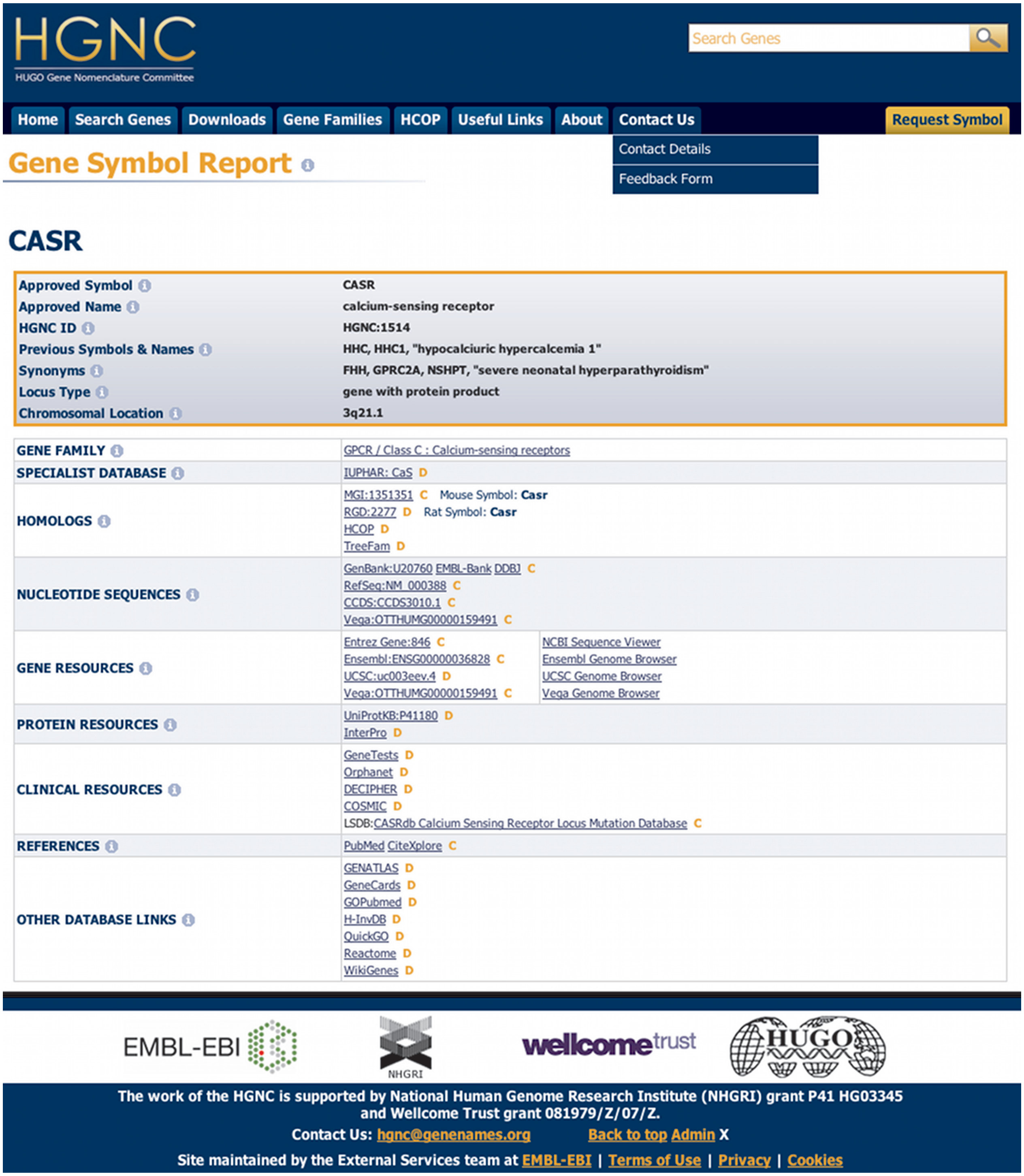
As discussed in 2011 (20), there are three ways to search for gene symbol reports within genenames.org, all of which can be accessed via the 'Search Genes' drop-down menu on the header (Figure 2). The most commonly used search is the 'Quick Gene Search', which can also be found on the home page and within the header (Figure 2). The 'Advanced Gene Search' tool allows the user to specify which fields in the HGNC data set they would like to search and build more complex queries with multiple search terms. The third search tool we provide is the 'List Search', which allows the user to type, paste or upload a list of symbols into a search field.

Gene symbol report

The majority of our users access the genenames.org site to retrieve our gene symbol reports, which are the main interface to our manually curated data and the external database links stored within our database (Figure 2). The gene symbol report page has been completely redesigned to fit our new colour scheme and to make using the page more intuitive. Our core HGNC data are now highlighted at the top of the report within a shaded box. This shaded area features approved nomenclature, a unique HGNC ID, previous nomenclature, synonyms, locus type and the chromosomal location, all of which have been curated manually.

In addition to the HGNC-specific data, we have a wide variety of external links that are relevant to the gene in question. These links are displayed below the shaded box in a table and are grouped together by the type of resource named within the first column. A letter 'C' next to the link indicates that the link has been checked and curated by a member of the HGNC. Links that have a 'D' placed alongside indicate that the link was derived and downloaded from an external source. Data links are organized into the following sections:

- **Gene family**—Only present if the gene is associated to a family or group. The link will navigate the user to the gene family page.
- **Specialist database**—Appears only if a value is present. Contains specialist external database resource links that are specific to a class of genes. To date, we link out to 14 specialist databases, a list of which can be found at www.genenames.org/useful/symbol-report-documentation, with a brief description.
- **Homologs**—A group of homology-related links, including our own HGNC Comparison of Orthology Predictions (HCOP) orthology data-mining tool and



Gene Symbol Report

CASR

Approved Symbol	CASR
Approved Name	calcium-sensing receptor
HGNC ID	HGNC:1514
Previous Symbols & Names	HHC, HHC1, "hypocalciuric hypercalcemia 1"
Synonyms	FHH, GPRC2A, NSHPT, "severe neonatal hyperparathyroidism"
Locus Type	gene with protein product
Chromosomal Location	3q21.1

GENE FAMILY	GPCR / Class C : Calcium-sensing receptors	
SPECIALIST DATABASE	IUPHAR: CaS D	
HOMOLOGS	MGI:1351351 C Mouse Symbol: Casr RGD:2277 D Rat Symbol: Casr HCOP D TreeFam D	
NUCLEOTIDE SEQUENCES	GenBank:U20760 EMBL-Bank DDBJ C RefSeq:NM_000388 C CCDS:CCDS3010.1 C Vega:OTTHUMG00000159491 C	
GENE RESOURCES	Entrez Gene:846 C Ensembl:ENSG00000036828 C UCSC:uc003eev.4 D Vega:OTTHUMG00000159491 C	NCBI Sequence Viewer Ensembl Genome Browser UCSC Genome Browser Vega Genome Browser
PROTEIN RESOURCES	UniProtKB:P41180 D InterPro D	
CLINICAL RESOURCES	GeneTests D Orphanet D DECIPHER D COSMIC D LSDB:CASRdb Calcium Sensing Receptor Locus Mutation Database C	
REFERENCES	PubMed CiteXplore C	
OTHER DATABASE LINKS	GENATLAS D GeneCards D GOPubmed D H-InvDB D QuickGO D Reactome D WikiGenes D	

The work of the HGNC is supported by National Human Genome Research Institute (NHGRI) grant P41 HG03345 and Wellcome Trust grant 081979/Z/07/Z.

Contact Us: hgnc@genenames.org Back to top Admin X

Site maintained by the External Services team at EMBL-EBI | Terms of Use | Privacy | Cookies

Figure 2. A typical gene symbol report showing the shaded box containing the HGNC core data with other information and links below. If the gene is within a gene family, then the gene family field will also be displayed. The new banner featuring a new logo and colour scheme is at the top of the image. The banner contains a simple quick search box to the right and holds tabs that produce drop-down menus, once the mouse cursor hovers over the tab. The website footer containing our affiliations, funding sources and policies is seen at the bottom of the page. The footer also sports an email link so that users can contact us via the link.

links to orthologous gene entries in mouse genome informatics (MGI) and rat genome database (RGD).

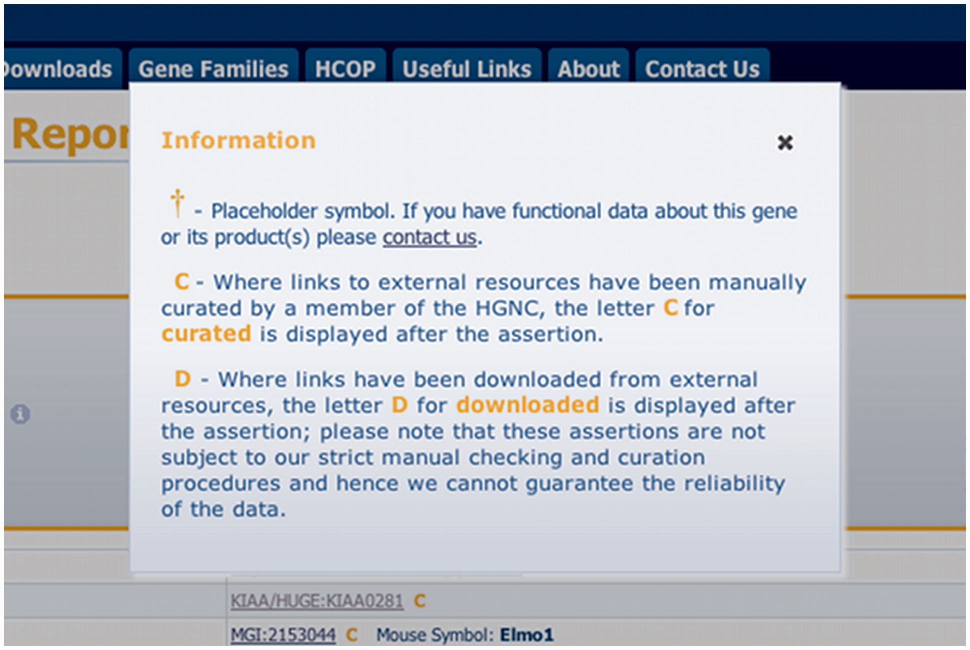
- **Nucleotide sequences**—Links to representative accessions from GenBank/EMBL/DDBJ, RefSeq, the CCDS project and Vega.
- **Gene resources**—Links to the four most popular gene and genome browsers, (Ensembl, Vega, Entrez Gene and UCSC); each resource has two links, one to the gene entry and the other to the genome browser.

- **Protein resources**—Information on proteins encoded by the gene. We include links to UniProt and to InterPro (23), which shows all the domains predicted within the encoded proteins by the InterPro member databases.
- **Clinical resources**—Links to resources for associated phenotypes, diseases and gene mutations.
- **References**—PubMed (24) and CiteXplore (http://www.ebi.ac.uk/citexplore/) hyperlinks, which display the abstracts for references pertinent to the gene. The purpose of the section is not to list all possible published articles for the gene, but to provide links to articles that first describe the gene in question or are particularly relevant to the nomenclature of the gene.
- **Other database links**—New links in this section are to Reactome (25), which contains manually curated, peer-reviewed signalling pathway data, and QuickGo (26) which lists all Gene Ontology terms annotated for the gene product(s).

Gene symbol reports include two new icons that are intended to help the user understand the data that we provide. The first is the dagger symbol that can be found next to the gene symbol in certain entries. This informs the user that this gene symbol is a placeholder, and that we are seeking functional data to rename the gene. In the past 2 years, we have made great strides in replacing these placeholder symbols with more informative symbols, reducing the number of C$orf symbols to ~780 from a peak of 1960, with >200 replaced in 2012 alone. We are always interested in hearing from people who can offer functional information about an entry with a placeholder symbol so that we can reduce these figures further. The second new icon is an 'i' within a circle that links to information about the field. Both these icons and the 'C' and 'D' keys once clicked will create a dialog box within the page (Figure 3). The dialog box provides the user with additional information about the field or key. The dialog boxes are created using jQuery UI and retrieve the text using AJAX. The text is retrieved only when asked and is retrieved only once, storing the text so that subsequent calls for the same text will be instant. If the user's browser has JavaScript switched off, the links will leave the page and fetch the symbol report documentation page that contains all the information about every field (http://www.genenames.org/useful/symbol-report-documentation).

Downloads

The gene symbol report is a good way of browsing individual genes within our database, but many want to retrieve large sets of data. For these users, we offer several ways of downloading data sets. The first of these is via the 'Statistics and Downloads', which allows users to download our entire data set, particular locus types and our complete gene families set (Figure 1). It also provides basic statistics of the number of entries we have for each locus group and type and informs the user when the database was most recently updated (UK time). The second is the 'browse-approved symbols by chromosome' interface on the homepage, as mentioned earlier. The third tool is the custom downloads, which offers a more advanced way to download our data, and is designed for users that want a specific set of data and/or specific fields within their set. The fourth tool is our BioMart (http://www.genenames.org/biomart) that provides another way of performing complex queries and creating bespoke data sets to download. Both the BioMart Central Portal (27)



Information

† - Placeholder symbol. If you have functional data about this gene or its product(s) please contact us.

C - Where links to external resources have been manually curated by a member of the HGNC, the letter C for curated is displayed after the assertion.

D - Where links have been downloaded from external resources, the letter D for downloaded is displayed after the assertion; please note that these assertions are not subject to our strict manual checking and curation procedures and hence we cannot guarantee the reliability of the data.

KIAA/HUGE:KIAA0281 C

MGI:2153044 C Mouse Symbol: Elmo1

Figure 3. An example of an information dialog pop up that appears if the keys 'C', 'D' or the dagger symbol are selected. When the dialog box appears, the background is greyed out, and the dialog comes into focus. The dialog box is scrollable if the text within does not fit into the 400 by 300 px box.

and our HGNC MartView will allow users to not only retrieve data via the web interface but also via a Perl API, RESTful web service and SOAP web interface.

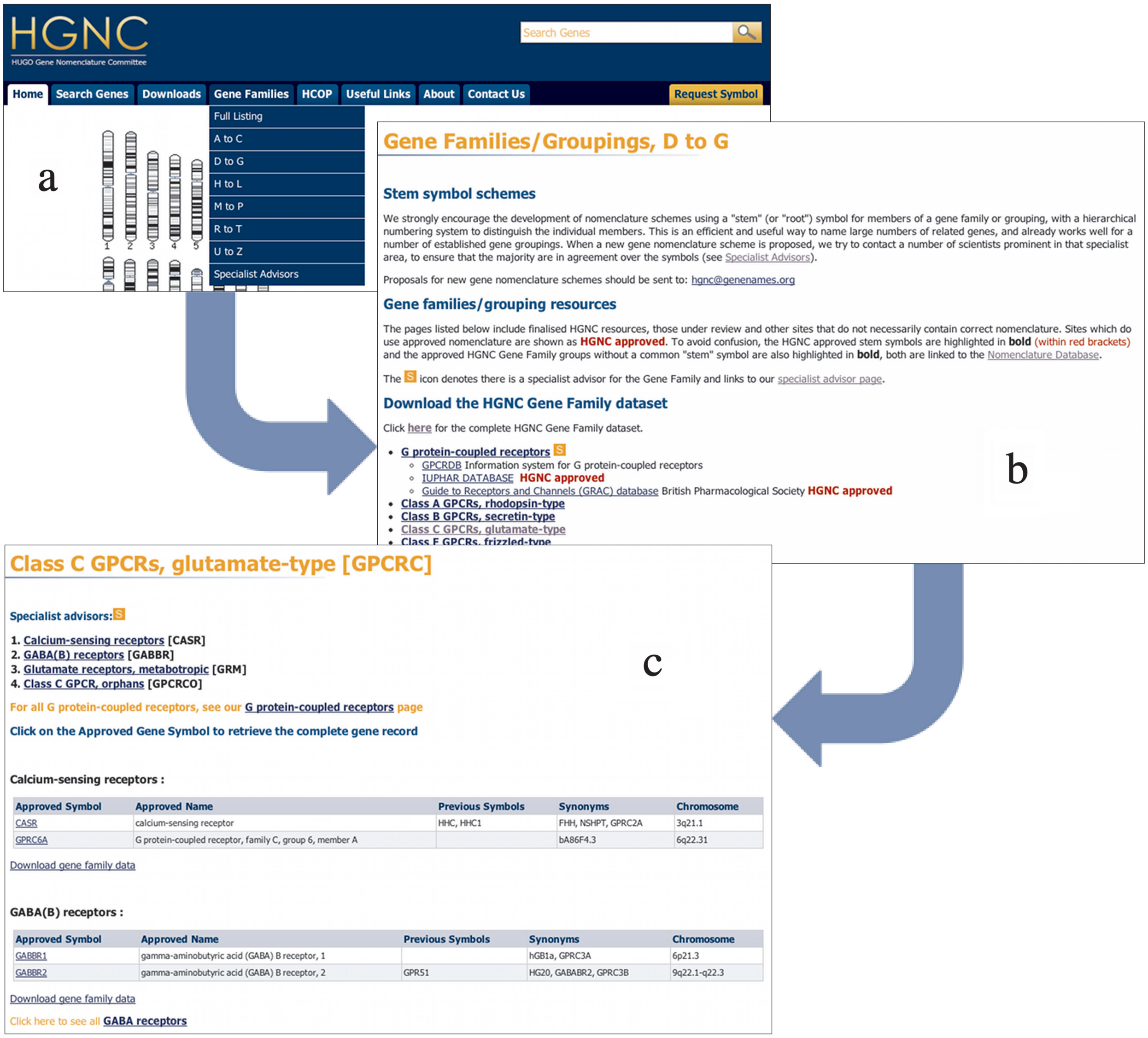
Gene families/groups

In the past 2 years, we have worked on expanding our resources for gene families and groupings (28). As of September 2012, we have ~45% of our entries associated with at least one gene family, and we have 475 families, which equate to ~400 pages of our website dedicated to gene families and sub-families. All families and groups can be found by using the 'Gene Families' tab within the header (Figure 4a), which links to a bullet-pointed list of family names and symbols (Figure 4b). To retrieve information about the family, the user must click on a symbol or hyperlinked name to direct them to the selected family page (Figure 4c), which lists all genes and/or subfamilies that belong to the family.

FUTURE DIRECTION

We are in the process of moving our site and database to two offsite datacentres with multilayer redundancy using virtual machines. Requests to our site will be load balanced between datacentres, which will provide quicker response times during heavy traffic periods. Having a multilayer redundancy infrastructure also allows us to carry out maintenance and repairs without affecting the service for our users, by taking one centre offline while the other serves the content.

We plan to replace our quick gene search with a solr search engine. Many of the world's largest websites have



Gene Families/Groupings, D to G

Stem symbol schemes

We strongly encourage the development of nomenclature schemes using a "stem" (or "root") symbol for members of a gene family or grouping, with a hierarchical numbering system to distinguish the individual members. This is an efficient and useful way to name large numbers of related genes, and already works well for a number of established gene groupings. When a new gene nomenclature scheme is proposed, we try to contact a number of scientists prominent in that specialist area, to ensure that the majority are in agreement over the symbols (see Specialist Advisors).

Proposals for new gene nomenclature schemes should be sent to: hgnc@genenames.org

Gene families/grouping resources

The pages listed below include finalised HGNC resources, those under review and other sites that do not necessarily contain correct nomenclature. Sites which do use approved nomenclature are shown as **HGNC approved**. To avoid confusion, the HGNC approved stem symbols are highlighted in **bold** (within red brackets) and the approved HGNC Gene Family groups without a common "stem" symbol are also highlighted in **bold**, both are linked to the Nomenclature Database.

The S icon denotes there is a specialist advisor for the Gene Family and links to our specialist advisor page.

Download the HGNC Gene Family dataset

Click here for the complete HGNC Gene Family dataset.

- **G protein-coupled receptors** S
  - GPCRDB Information system for G protein-coupled receptors
  - IUPHAR DATABASE **HGNC approved**
  - Guide to Receptors and Channels (GRAC) database British Pharmacological Society **HGNC approved**
- **Class A GPCRs, rhodopsin-type**
- **Class B GPCRs, secretin-type**
- Class C GPCRs, glutamate-type
- **Class F GPCRs, frizzled-type**

Class C GPCRs, glutamate-type [GPCRC]

Specialist advisors: S

1. **Calcium-sensing receptors [CASR]**
2. **GABA(B) receptors [GABBR]**
3. **Glutamate receptors, metabotropic [GRM]**
4. **Class C GPCR, orphans [GPCRCO]**

For all G protein-coupled receptors, see our **G protein-coupled receptors** page

Click on the Approved Gene Symbol to retrieve the complete gene record

Calcium-sensing receptors :

Approved Symbol	Approved Name	Previous Symbols	Synonyms	Chromosome
CASR	calcium-sensing receptor	HHC, HHC1	FHH, NSHPT, GPRC2A	3q21.1
GPRC6A	G protein-coupled receptor, family C, group 6, member A		bA86F4.3	6q22.31

Download gene family data

GABA(B) receptors :

Approved Symbol	Approved Name	Previous Symbols	Synonyms	Chromosome
GABBR1	gamma-aminobutyric acid (GABA) B receptor, 1		hGB1a, GPRC3A	6p21.3
GABBR2	gamma-aminobutyric acid (GABA) B receptor, 2	GPR51	HG20, GABABR2, GPRC3B	9q22.1-q22.3

Download gene family data

Click here to see all **GABA receptors**

Figure 4. The most common way of browsing for gene families is shown in the figure. The flow through the pages is seen by following the arrows from image a to image d. From the header, a user would select an option from the drop-down shown (**a**). This then takes the user to a families list page (**b**) in which the user selects a family, bringing them to the family page (**c**). From the family page, the user may then select a gene symbol, taking them to a gene symbol report (see Figure 2).

adopted solr, as it is open source and offers very quick and powerful full-text searches, hit highlighting and faceted searches, as well as many other features. These search engines are also highly scalable, offering efficient replication to other search servers if needed and are optimized for high volumes of web traffic. We hope that by embracing solr, we will improve our search times and will provide a search that will be more intuitive for users. Maintenance of the search will also be easier, as being part of a large community of solr users, we will be able to find help and advice on many forums, published articles and books.

HCOP (HGNC Comparison of Orthology Predictions http://www.genenames.org/hcop) (29) is a tool that searches and displays predicted orthologs of a particular human gene or set of genes made by multiple orthology resources. The reliability of the prediction can be assumed by the number of databases, which concur and by the presence or absence of synteny between the relevant chromosomes, where known. At present HCOP contains data from 14 genomes that can be compared with the human genome. We are planning to expand HCOP by increasing the number of resources and species, and the interface will also be improved for searching and downloading the data. This will aid us in our future work assigning standardized gene names to orthologous genes across vertebrate species. We will also continue to replace placeholder symbols with more informative symbols and introduce novel entries for loci that are only found on alternative assemblies where annotated by the Genome Reference Consortium (30). For any comments or questions concerning our work, please contact us via hgnc@genenames.org.

ACKNOWLEDGEMENTS

The authors would like to thank all of their past HGNC colleagues for their invaluable contributions to this project.

FUNDING

The Wellcome Trust [081979/Z/07/Z] and [099129/Z/12/Z]; National Human Genome Research Institute [P41 HG03345]. Funding for open access charge: The Wellcome Trust.

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. Eppig,J.T., Blake,J.A., Bult,C.J., Kadin,J.A. and Richardson,J.E. (2012) The mouse genome database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. *Nucleic Acids Res.*, **40**, D881–D886.
3. Dwinell,M.R., Worthey,E.A., Shimoyama,M., Bakir-Gungor,B., DePons,J., Laulederkind,S., Lowry,T., Nigram,R., Petri,V., Smith,J. *et al.* (2009) The rat genome database 2009: variation, ontologies and pathways. *Nucleic Acids Res.*, **37**, D744–D749.
4. Bradford,Y., Conlin,T., Dunn,N., Fashena,D., Frazer,K., Howe,D.G., Knight,J., Mani,P., Martin,R., Moxon,S.A. *et al.* (2011) ZFIN: enhancements and updates to the Zebrafish Model Organism Database. *Nucleic Acids Res.*, **39**, D822–D829.
5. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **39**, D52–D57.
6. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fairley,S., Fitzgerald,S. *et al.* (2012) Ensembl 2012. *Nucleic Acids Res.*, **40**, D84–D90.
7. Wilming,L.G., Gilbert,J.G., Howe,K., Trevanion,S., Hubbard,T. and Harrow,J.L. (2008) The vertebrate genome annotation (Vega) database. *Nucleic Acids Res.*, **36**, D753–D760.
8. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2011) GenBank. *Nucleic Acids Res.*, **39**, D32–D37.
9. Kaminuma,E., Kosuge,T., Kodama,Y., Aono,H., Mashima,J., Gojobori,T., Sugawara,H., Ogasawara,O., Takagi,T., Okubo,K. *et al.* (2011) DDBJ progress report. *Nucleic Acids Res.*, **39**, D22–D27.
10. Leinonen,R., Akhtar,R., Birney,E., Bower,L., Cerdeno-Tarraga,A., Cheng,Y., Cleland,I., Faruque,N., Goodgame,N., Gibson,R. *et al.* (2011) The European nucleotide archive. *Nucleic Acids Res.*, **39**, D28–D31.
11. Safran,M., Dalah,I., Alexander,J., Rosen,N., Iny Stein,T., Shmoish,M., Nativ,N., Bahir,I., Doniger,T., Krug,H. *et al.* (2010) GeneCards Version 3: the human gene integrator. *Database*, **2010**, baq020.
12. Dreszer,T.R., Karolchik,D., Zweig,A.S., Hinrichs,A.S., Raney,B.J., Kuhn,R.M., Meyer,L.R., Wong,M., Sloan,C.A., Rosenbloom,K.R. *et al.* (2012) The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res.*, **40**, D918–D923.
13. The UniProt Consortium. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **40**, D71–D75.
14. Firth,H.V., Richards,S.M., Bevan,A.P., Clayton,S., Corpas,M., Rajan,D., Van Vooren,S., Moreau,Y., Pettett,R.M. and Carter,N.P. (2009) DECIPHER: database of chromosomal imbalance and phenotype in humans using ensembl resources. *Am. J. Hum. Genet.*, **84**, 524–533.
15. Amberger,J., Bocchini,C.A., Scott,A.F. and Hamosh,A. (2009) McKusick's online mendelian inheritance in man (OMIM). *Nucleic Acids Res.*, **37**, D793–D796.
16. Forbes,S.A., Bindal,N., Bamford,S., Cole,C., Kok,C.Y., Beare,D., Jia,M., Shepherd,R., Leung,K., Menzies,A. *et al.* (2011) COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Res.*, **39**, D945–D950.
17. Pennisi,E. (2003) Human genome. a low number wins the GeneSweep Pool. *Science*, **300**, 1484.
18. Pruitt,K.D., Harrow,J., Harte,R.A., Wallin,C., Diekhans,M., Maglott,D.R., Searle,S., Farrell,C.M., Loveland,J.E., Ruef,B.J. *et al.* (2009) The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.*, **19**, 1316–1323.
19. Wright,M.W. and Bruford,E.A. (2011) Naming 'junk': human non-protein coding RNA (ncRNA) gene nomenclature. *Hum. Genomics*, **5**, 90–98.
20. Seal,R.L., Gordon,S.M., Lush,M.J., Wright,M.W. and Bruford,E.A. (2011) genenames.org: the HGNC resources in 2011. *Nucleic Acids Res.*, **39**, D514–D519.
21. Eilbeck,K., Lewis,S.E., Mungall,C.J., Yandell,M., Stein,L., Durbin,R. and Ashburner,M. (2005) The sequence ontology: a tool for the unification of genome annotations. *Genome Biol.*, **6**, R44.
22. Monk,D., Sanches,R., Arnaud,P., Apostolidou,S., Hills,F.A., Abu-Amero,S., Murrell,A., Friess,H., Reik,W., Stanier,P. *et al.* (2006) Imprinting of IGF2 P0 transcript and novel alternatively spliced INS-IGF2 isoforms show differences between mouse and human. *Hum. Mol. Genet.*, **15**, 1259–1269.
23. Hunter,S., Jones,P., Mitchell,A., Apweiler,R., Attwood,T.K., Bateman,A., Bernard,T., Binns,D., Bork,P., Burge,S. *et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*, **40**, D306–D312.

24. Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Federhen,S. *et al*. (2012) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **40**, D13–D25.
25. D'Eustachio,P. (2011) Reactome knowledgebase of human biological pathways and processes. *Methods Mol. Biol*., **694**, 49–61.
26. Binns,D., Dimmer,E., Huntley,R., Barrell,D., O'Donovan,C. and Apweiler,R. (2009) QuickGO: a web-based tool for gene ontology searching. *Bioinformatics*, **25**, 3045–3046.
27. Guberman,J.M., Ai,J., Arnaiz,O., Baran,J., Blake,A., Baldock,R., Chelala,C., Croft,D., Cros,A., Cutts,R.J. *et al*. (2011) BioMart central portal: an open database network for the biological community. *Database*, **2011**, bar041.
28. Daugherty,L.C., Seal,R.L., Wright,M.W. and Bruford,E.A. (2012) Gene family matters: expanding the HGNC resource. *Hum. Genomics*, **6**, 4.
29. Eyre,T.A., Wright,M.W., Lush,M.J. and Bruford,E.A. (2007) HCOP: a searchable database of human orthology predictions. *Brief. Bioinform*., **8**, 2–5.
30. Church,D.M., Schneider,V.A., Graves,T., Auger,K., Cunningham,F., Bouk,N., Chen,H.C., Agarwala,R., McLaren,W.M., Ritchie,G.R. *et al*. (2011) Modernizing reference genome assemblies. *PLoS Biol*., **9**, e1001091.