Published online 31 October 2014 Nucleic Acids Research, 2015, Vol. 43, Database issue **D1079–D1085**
doi: 10.1093/nar/gku1071

Genenames.org: the HGNC resources in 2015

Kristian A. Gray, Bethan Yates, Ruth L. Seal, Mathew W. Wright and Elspeth A. Bruford*

HUGO Gene Nomenclature Committee, European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

Received September 15, 2014; Revised October 15, 2014; Accepted October 16, 2014

ABSTRACT

The HUGO Gene Nomenclature Committee (HGNC) based at the European Bioinformatics Institute (EMBL-EBI) assigns unique symbols and names to human genes. To date the HGNC have assigned over 39 000 gene names and, representing an increase of over 5000 entries in the past two years. As well as increasing the size of our database, we have continued redesigning our website http://www.genenames.org and have modified, updated and improved many aspects of the site including a faster and more powerful search, a vastly improved HCOP tool and a REST service to increase the number of ways users can retrieve our data. This article provides an overview of our current online data and resources, and highlights the changes we have made in recent years.

INTRODUCTION

The HUGO Gene Nomenclature Committee (HGNC) is responsible for providing unique and informative nomenclature for all genes within the human genome (1). The majority of our data are manually curated, and we collaborate with researchers working on particular genes and gene families to create names that are acceptable to researchers in the field. HGNC symbols and names are seen as a standard and used in all the major databases that concentrate on human genes and proteins, such as Ensembl (2), UniProt (3), NCBI Gene (4), ENA/GenBank/DDBJ (5–7), Vega (8), GeneCards (9) and the UCSC genome browser (10), as well as disease and phenotype resources including Decipher (11), OMIM (12), Locus Reference Genomics (LRG)(13), ClinVar (14) and GeneTests (15). By working closely with other species' nomenclature groups such as mouse (16), rat (17) and zebrafish (18) we ensure that orthologous genes are assigned equivalent symbols where possible.

We also maintain a website http://www.genenames.org that provides a public access portal without any restrictions to the data, with tools to search, download and discover genes, gene families and possible orthologs. More information about our website and applications will appear later in this article.

DATA

As of the beginning of August 2014, we have 39 135 active entries within our database, with 19 064 of the entries being protein coding genes. Comparing the number of protein coding genes with the figure quoted in our last Nucleic Acids Research database issue article (19) of 19 027, it is clear that the number is plateauing at the low 19 000 mark. In June 2014, Ezkurdia and co-workers released an article (20) in which they describe comparing peptide mass spectroscopy results with GENCODE (21) data and concluded that the number of protein coding genes will approximately be 19 000 within human. The CCDS project (discussed below), whose remit is to identify a core set of protein coding regions, have assigned CCDS IDs to 18 681 genes as of August 2014. Taking these two pieces of information into account, we believe that we have an almost complete set of protein coding genes for the human genome.

This year (2014) the HGNC have become members of the consensus coding sequence (CCDS) project (22), which aims to develop a set of high-quality gene annotations for human and mouse protein-coding genes. The project is an international collaboration between the Ensembl (2), HAVANA (8), NCBI RefSeq (23), MGI (16) and HGNC teams. NCBI provides automatic and manually curated RefSeq annotations, and Ensembl provides an annotation set that includes Ensembl gene model predictions and manual gene annotations by HAVANA. Protein-coding annotations with matching genomic coordinates that pass a set of stringent quality control tests are assigned a unique CCDS identifier (ID). CCDS IDs that are marked for changes of any kind undergo a review process that includes annotators from RefSeq and HAVANA, and curators from the HGNC for human genes and from MGI for mouse genes.

The largest area of growth within our Symbol Reports is currently long non-coding RNA (lncRNA) genes, with entries having increased by 69% since 2012 (19). For a full review of how we name lncRNA genes please see our recent article (24). The number of approved gene symbols for pseudogenes is also rising rapidly with entries having increased by 50% since 2012 (19). Our pseudogene locus type now includes pseudogenes of RNA genes such as *RN7SKP1* (RNA, 7SK small nuclear pseudogene 1) and also functional pseudogenes, which we define as

[*] To whom correspondence should be addressed. Tel: +44 1223 494444; Fax: +44 1223 494468; Email: elspeth@ebi.ac.uk

transcribed pseudogenes that have been reported to have a functional role in a publication. These can be identified in our website by inclusion of '(functional)' at the end of the gene name, e.g. the full name for *MYLKP1* is 'myosin light chain kinase pseudogene 1 (functional)'. A link to the source publication is included in each relevant Symbol Report, and we maintain a webpage where all functional pseudogenes can be viewed (http://www.genenames.org/functional-pseudogenes). Another area of growth within our database has been the increase in the number of gene families, something we commented on in our last article (19). To date, we have 588 families with 51% of the protein coding genes within our database associated to at least one family.

The human genome build is no longer just the 'golden path'. To adequately represent complex structural variations, the Genome Reference Consortium includes alternative loci within their build of the human genome (25). The HGNC aims to name all structural variants annotated on alternative reference loci as described in (26). These genes can be identified within our Symbol Reports by the Chromosomal Location field, which displays the location plus the term 'alternate reference locus' and can be viewed in full using the table provided on our Statistics and Downloads page labelled 'Alternative Loci Statistics'.

WEBSITE

Our website http://www.genenames.org is the public access portal to our data, and since the new design release in May 2011 we have built upon the site, updating and improving many sections including our search and HCOP tool, as well as redesigning the homepage. We have moved the karyotype image to the statistics and downloads page, and replaced it with a word cloud of common root symbols. Each root symbol within the word cloud is a link that retrieves a search result of genes containing the root symbol. Another addition to the design is a new search bar found in the masthead of every page. By default, the search bar will search the symbol reports, but users have the option of searching the whole site by selecting 'site search' in the drop down.

The site and database now run from two offsite data centres with multilayer redundancy using virtual machines. Our user requests are now load balanced between data centres and servers, providing quicker response times during peak times. This infrastructure also allows maintenance without interruption of service, by taking one data centre offline to work on while the other continues to serve content. As well as the infrastructure changing, a number of our pages and applications have been refactored or replaced, and so have the URLs (see http://www.genenames.org/news/new-features-and-changes#URL_changes_-_2014--01--08). Below is a brief tour of the changes and additions we have made to the site within the last two years.

Symbol reports

The symbol reports are our most visited pages as this is our main interface to the manually curated nomenclature and external database links. Below, we briefly outline the key updates we have made to the symbol reports.

Homologs. This section has been reformatted into a table structure to show the rat and mouse homologs of the human gene. The table is split into three columns: organism, symbol and database which displays the gene ID for that species' model organism database. Currently we only show mouse and rat homologs but this may change in the future. In the past we only showed one ortholog per species but now we show multiple paralogous genes per species which better represents homology across the species.

New links. 'Protein resources' include a new link to PDBe (27) in addition to the links to UniProt and Interpro (28). This means we now have links to a resource that describes the protein in general, the domains it contains and the 3D structure of the protein all linked by the protein accession. A new clinical resource link that we have added is to the Genetics Testing Registry (GTR) (29). The GTR is a repository of voluntary submissions of genetic testing methodologies and purpose. We have also added a new '*Other databases*' link to BioGPS (30) which is a gene annotation portal that allows the user to 'plugin' resources into the application for a gene. By default, the BioGPS link will display a graphical representation of gene expression based on GeneAtlas U133A, gcrma data set.

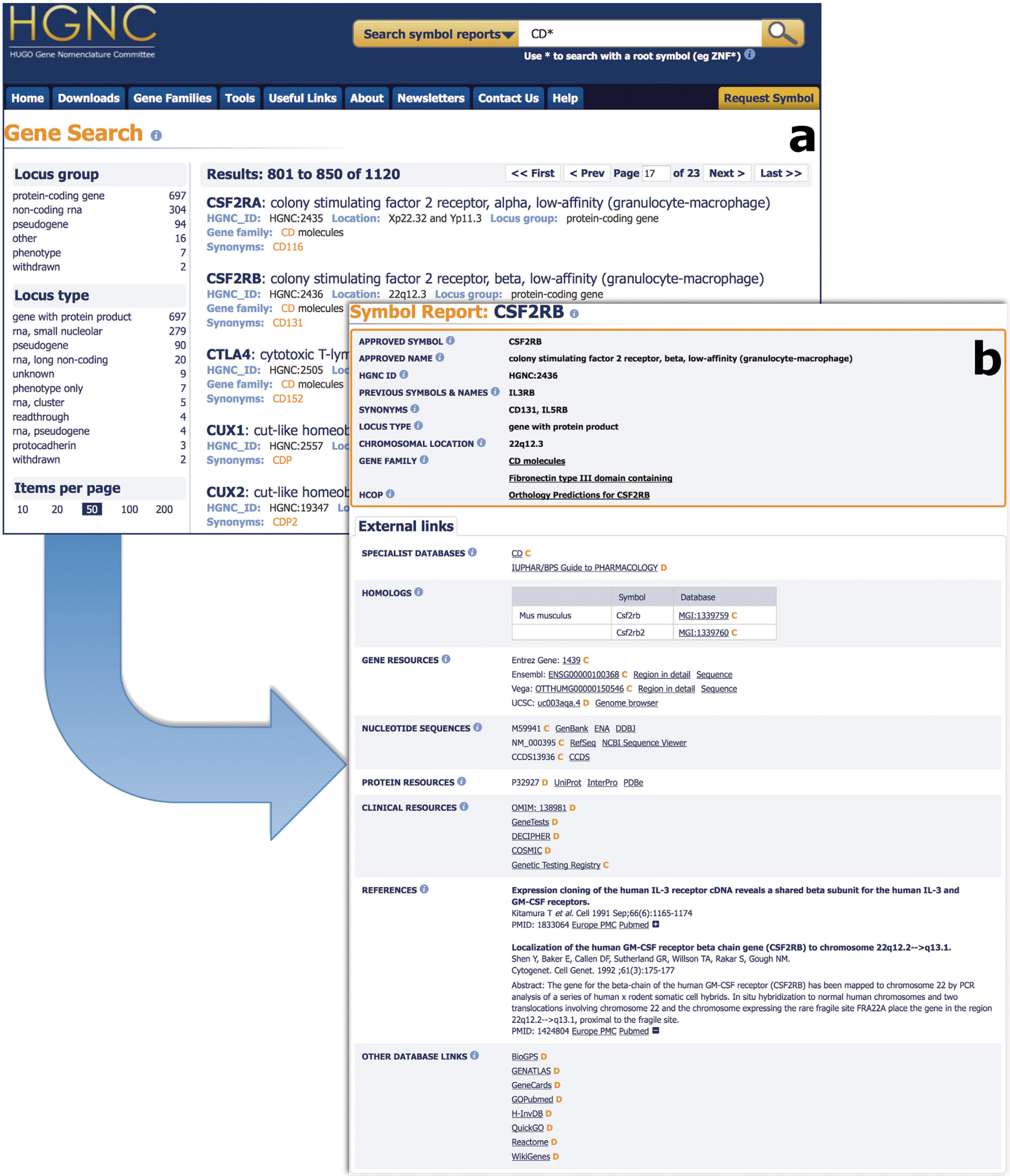
References. The references section of the 'external links' section has changed considerably. In the past we displayed the PMID number, a link to PubMed (31) and a link to CiteXplore which has now become Europe PubMed Central (Europe PMC) (http://europepmc.org/). What we display now by default is the title, article reference, PMID number and links to both Europe PMC and PubMed. Next to the links to PubMed we have added a plus sign which when clicked, alters the reference to include the full author list and abstract. This is all done with an AJAX request to the Europe PMC RESTful web service, retrieving the information needed by PMID number and rendering the information within the symbol report page (Figure 1b).

New faster and smarter gene search

In January 2014 we replaced the 'quick search' application with a Solr search engine to search symbol reports. Solr is an open source search engine, built upon apache lucene, that many of the world's largest websites have adopted since it offers very quick full-text searches. Solr servers are also highly scalable and offer efficient replication to other search servers if needed.

Adding a query term into the input box within the masthead and clicking on the spyglass activates a search. By default, the search term is used for a full-text search over all the indexed fields. Our search also allows users to match records based on a query pattern using the wildcard and logical operators with OR being the default operator. Double quoting the search query term will tell the search to treat the quoted block as one term. Users can search against specific indexed fields by using a colon between the indexed field key and the term.

The results page contains two columns. On the left we utilise Solr's ability to provide facets. The results are faceted by locus group and locus type with counts of the number



Gene Search

Locus group

protein-coding gene	697
non-coding rna	304
pseudogene	94
other	16
phenotype	7
withdrawn	2

Locus type

gene with protein product	697
rna, small nucleolar	279
pseudogene	90
rna, long non-coding	20
unknown	9
phenotype only	7
rna, cluster	5
readthrough	4
rna, pseudogene	4
protocadherin	3
withdrawn	2

Items per page

10 20 50 100 200

Results: 801 to 850 of 1120 << First < Prev Page 17 of 23 Next > Last >>

CSF2RA: colony stimulating factor 2 receptor, alpha, low-affinity (granulocyte-macrophage)
HGNC_ID: HGNC:2435 Location: Xp22.32 and Yp11.3 Locus group: protein-coding gene
Gene family: CD molecules
Synonyms: CD116

CSF2RB: colony stimulating factor 2 receptor, beta, low-affinity (granulocyte-macrophage)
HGNC_ID: HGNC:2436 Location: 22q12.3 Locus group: protein-coding gene
Gene family: CD molecules
Synonyms: CD131

CTLA4: cytotoxic T-lym
HGNC_ID: HGNC:2505 Loc
Gene family: CD molecules
Synonyms: CD152

CUX1: cut-like homeob
HGNC_ID: HGNC:2557 Loc
Synonyms: CDP

CUX2: cut-like homeob
HGNC_ID: HGNC:19347 Lo
Synonyms: CDP2

Symbol Report: CSF2RB

APPROVED SYMBOL	CSF2RB
APPROVED NAME	colony stimulating factor 2 receptor, beta, low-affinity (granulocyte-macrophage)
HGNC ID	HGNC:2436
PREVIOUS SYMBOLS & NAMES	IL3RB
SYNONYMS	CD131, IL5RB
LOCUS TYPE	gene with protein product
CHROMOSOMAL LOCATION	22q12.3
GENE FAMILY	CD molecules Fibronectin type III domain containing
HCOP	Orthology Predictions for CSF2RB

External links

SPECIALIST DATABASES
CD C
IUPHAR/BPS Guide to PHARMACOLOGY D

HOMOLOGS

	Symbol	Database
Mus musculus	Csf2rb	MGI:1339759 C
	Csf2rb2	MGI:1339760 C

GENE RESOURCES
Entrez Gene: 1439 C
Ensembl: ENSG00000100368 C Region in detail Sequence
Vega: OTTHUMG00000150546 C Region in detail Sequence
UCSC: uc003aqa.4 D Genome browser

NUCLEOTIDE SEQUENCES
M59941 C GenBank ENA DDBJ
NM_000395 C RefSeq NCBI Sequence Viewer
CCDS13936 C CCDS

PROTEIN RESOURCES
P32927 D UniProt InterPro PDBe

CLINICAL RESOURCES
OMIM: 138981 D
GeneTests D
DECIPHER D
COSMIC D
Genetic Testing Registry C

REFERENCES
Expression cloning of the human IL-3 receptor cDNA reveals a shared beta subunit for the human IL-3 and GM-CSF receptors.
Kitamura T *et al.* Cell 1991 Sep;66(6):1165-1174
PMID: 1833064 Europe PMC Pubmed ⊞

Localization of the human GM-CSF receptor beta chain gene (CSF2RB) to chromosome 22q12.2-->q13.1.
Shen Y, Baker E, Callen DF, Sutherland GR, Willson TA, Rakar S, Gough NM.
Cytogenet. Cell Genet. 1992 ;61(3):175-177
Abstract: The gene for the beta-chain of the human GM-CSF receptor (CSF2RB) has been mapped to chromosome 22 by PCR analysis of a series of human x rodent somatic cell hybrids. In situ hybridization to normal human chromosomes and two translocations involving chromosome 22 and the chromosome expressing the rare fragile site FRA22A place the gene in the region 22q12.2-->q13.1, proximal to the fragile site.
PMID: 1424804 Europe PMC Pubmed ⊟

OTHER DATABASE LINKS
BioGPS D
GENATLAS D
GeneCards D
GOPubmed D
H-InvDB D
QuickGO D
Reactome D
WikiGenes D

Figure 1. The new search box located within the masthead activates the search and returns the results as seen above in section '**a**'. The results are faceted as seen on the left allowing the user to filter the returned results by locus type and group. The results display some of the core HGNC data such as symbol, name and ID. Each result also highlights the field(s) the term matches. Section '**b**' displays a typical symbol report showing the HGNC core data fields within the shaded box. New additions to the core data are the HCOP and gene families links that have migrated from the external links section below. A slightly changed format to the external links section can be seen including our improved references section. The first reference is the default view while the second displays the full author list and abstract as would be displayed once the plus sign is clicked. Clicking on the minus will return the reference back to the default. Another addition is the support for multiple mouse and rat homologs per human gene within our symbol reports as seen above.

of reports that fit the facets (Figure 1a). Also within this column users can change the default number of results displayed per page from 10 up to 200. In the right column we list the results displayed primarily in order of relevance and then alphanumerically on symbol. The first line of a 'hit' contains the approved gene symbol and the approved gene name as defined by the HGNC. On the second row we display the HGNC ID, the chromosome location (if specified) and the locus group to which the entry belongs. The remaining row reports the field type the keyword/ID matches and highlights the query term within the field. If the number of results is greater than the number of displayed items on a page, a pager will appear, allowing users to navigate through the results. For a list of indexed fields and more information on the advanced features of the search, please see http://www.genenames.org/help/search.

Downloads

As well as our pre-existing BioMart and custom downloads applications we have introduced a new REST service. This is a new addition to the site allowing users to search/query our database and retrieve data in an XML or JSON format from within a calling script or program. The REST service is built upon our Solr search engine and has its own URL of http://rest.genenames.org. The service has three main requests, 'info', 'search' and 'fetch'. The info request is the simplest in that the URL is http://rest.genenames.org/info without any other parameters. The info request notifies the user what fields can be queried and what information is stored. Info also returns exactly when the data were modified and the total number of records including records with an 'Entry Withdrawn' status. Changes to the service such as new or removed fields will be seen using the info request. The other two requests to the REST service are search and fetch. Since the server is built upon Solr, searching for records is far quicker than fetching all stored fields; therefore for a large query the user should request a search and then fetch individual records. The search and fetch requests can utilise the same indexed fields as mentioned within the search section of this article, but the search request will only return the HGNC ID, symbol and score of the search. By default, the search request will search all indexed fields if no field is provided (e.g http://rest.genenames.org/search/ZNF3). If a specific field should be searched then the field should come between the request type and the term (e.g http://rest.genenames.org/search/symbol/ZNF3). The full arsenal of complex query methods is available to the search request such as phrases, wildcards and logical operators. The fetch method requires the user to add a searchable field and the query term to the URL (e.g http://rest.genenames.org/fetch/hgnc_id/HGNC:1097) and will not accept complex queries like wildcards. Fetch will however return all stored fields as described within an info request result and therefore the entire gene symbol report in XML or JSON format. For detailed information on how to use our REST service please visit http://rest.genenames.org.

Multi-symbol Checker

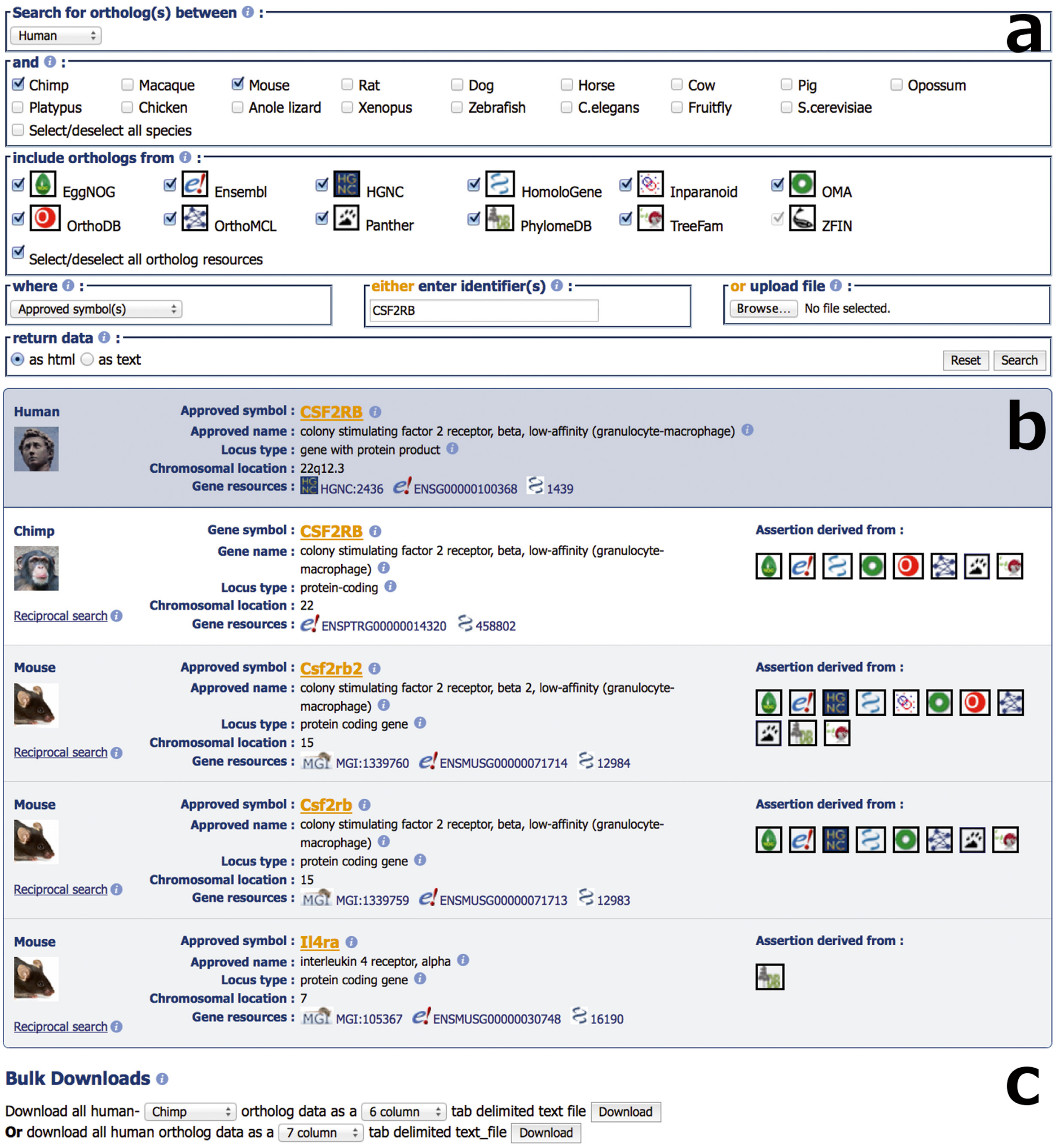
This tool has replaced the 'list search' tool but contains the same functionality as the old application and introduces a sortable results table, filtering by match type, and an increase in speed for large symbol lists. We have redesigned the form for the application to take full advantage of the width of the page and reduce the overall real estate size of the form. Lists of gene symbols can be added to the text box or uploaded to the site within a file. Comma, space, new line or a mixture of the three can separate the symbols within the file or text box. As well as returning a sortable table of results the user can request for the results to be returned as tab delimited plain text. As the name suggests the main functionality of this tool is to check a batch of symbols from the user to see if these symbols are still current; however it can also be used as a bulk symbol search to browse the symbol records within our database for a given set.

HCOP

A new and improved version of our HGNC Comparison of Orthology Predictions (32,33) tool has been released. HCOP currently collates orthology assertions from eggNOG (34), Ensembl Compara (2), HGNC, HomoloGene (31), InParanoid (35), OMA (36), OrthoDB (37), OrthoMCL (38), PANTHER (39), PhylomeDB (40), Treefam (41) and ZFIN (18) into a single tool, enabling comparison of these data to identify a consensus orthology prediction for a specified human gene or set of genes. Data are integrated using pre-existing mappings from the orthology resources to link between model organism databases, NCBI Gene and Ensembl. An indication of the reliability of a prediction is provided by the number of orthology databases which concur. HCOP was originally designed to show orthology predictions between human and mouse, but has since been expanded to include data from chimp, macaque, rat, dog, horse, cow, pig, opossum, platypus, chicken, anole lizard, xenopus, zebrafish, *C. elegans*, fruitfly and *S. cerevisiae,* meaning that there are now 18 genomes available for comparison in HCOP.

The updated version of HCOP involved a complete rewrite of the pipeline used to combine the data from each of the orthology sources. The data behind HCOP are stored in a MySQL database to allow for rapid querying. Each orthology assignment is stored as a pair of genes with a list of associated databases that support that assertion. The data are updated weekly by running a pipeline developed using the eHive (42) production system from Ensembl. In addition to the orthology data we also import gene data from HGNC, MGI, RGD (17), Xenbase (43) and ZFIN to ensure that we have the current approved gene symbols, names, locus types and location information from the appropriate nomenclature or model organism database. For those species without an official nomenclature committee or where a nomenclature committee exists but gene data are not available in a form our pipeline can utilise, we take this information from the NCBI Gene database, or from Ensembl if the gene in question cannot be mapped to an NCBI Gene identifier.

The HCOP user interface (see Figure 2a) has undergone a redesign to improve its usability while still retaining the functionality of the original version. For species with either a model organism or nomenclature database you can search using an approved gene symbol, an approved gene name or



Search for ortholog(s) between : Human

and :
☑ Chimp ☐ Macaque ☑ Mouse ☐ Rat ☐ Dog ☐ Horse ☐ Cow ☐ Pig ☐ Opossum
☐ Platypus ☐ Chicken ☐ Anole lizard ☐ Xenopus ☐ Zebrafish ☐ C.elegans ☐ Fruitfly ☐ S.cerevisiae
☐ Select/deselect all species

include orthologs from :
☑ EggNOG ☑ Ensembl ☑ HGNC ☑ HomoloGene ☑ Inparanoid ☑ OMA
☑ OrthoDB ☑ OrthoMCL ☑ Panther ☑ PhylomeDB ☑ TreeFam ☑ ZFIN
☑ Select/deselect all ortholog resources

where : Approved symbol(s)

either enter identifier(s) : CSF2RB

or upload file : No file selected.

return data : ● as html ○ as text

Human

Approved symbol : CSF2RB
Approved name : colony stimulating factor 2 receptor, beta, low-affinity (granulocyte-macrophage)
Locus type : gene with protein product
Chromosomal location : 22q12.3
Gene resources : HGNC:2436 ENSG00000100368 1439

Chimp

Gene symbol : CSF2RB
Gene name : colony stimulating factor 2 receptor, beta, low-affinity (granulocyte-macrophage)
Locus type : protein-coding
Chromosomal location : 22
Gene resources : ENSPTRG00000014320 458802
Assertion derived from :
Reciprocal search

Mouse

Approved symbol : Csf2rb2
Approved name : colony stimulating factor 2 receptor, beta 2, low-affinity (granulocyte-macrophage)
Locus type : protein coding gene
Chromosomal location : 15
Gene resources : MGI MGI:1339760 ENSMUSG00000071714 12984
Assertion derived from :
Reciprocal search

Mouse

Approved symbol : Csf2rb
Approved name : colony stimulating factor 2 receptor, beta, low-affinity (granulocyte-macrophage)
Locus type : protein coding gene
Chromosomal location : 15
Gene resources : MGI MGI:1339759 ENSMUSG00000071713 12983
Assertion derived from :
Reciprocal search

Mouse

Approved symbol : Il4ra
Approved name : interleukin 4 receptor, alpha
Locus type : protein coding gene
Chromosomal location : 7
Gene resources : MGI MGI:105367 ENSMUSG00000030748 16190
Assertion derived from :
Reciprocal search

Bulk Downloads

Download all human- Chimp ortholog data as a 6 column tab delimited text file
Or download all human ortholog data as a 7 column tab delimited text_file

Figure 2. Section ‘**a**’ shows the updated HCOP search form. Users select a primary species and one or more species that they wish to identify orthologs in. They then select the ortholog resources they wish to include in the search, and the type of search term, e.g. approved symbol, or database identifier, that they are providing. A single search term or list of search terms may be pasted into a text box, or uploaded as a file to be used to run the search. Users can optionally see the results in HTML or plain text format. Section ‘**b**’ displays an example result panel from HCOP. In this case orthologs between the human gene *CSF2RB* in chimp and mouse were requested. Information about the query gene appears in the blue section at the top of the results panel, with each ortholog identified having its own section below this. For both the query and ortholog genes we provide basic information about the gene as well as links to the gene in other resources. Each ortholog section has an additional column labelled ‘Assertion derived from’ that contains a set of icons that represent the orthology sources that support this assignment. Section ‘**c**’ is the bulk downloads section of the HCOP form. This allows users to download pre-computed files of HCOP data from our FTP site, providing fast access to all the data stored in HCOP.

the gene identifier from that database. Orthology assertions can also be obtained for a gene by searching with its Ensembl gene identifier or its NCBI Gene identifier. In the latest version of HCOP you can now specify which species and orthology sources you wish to see represented in the results. The consensus orthology assertions for multiple genes can be viewed simultaneously by searching with a list of query terms.

The display of HCOP results (see Figure 2b) has changed significantly. A scrollable results panel is displayed for each query term that produces a result. At the top of the results panel in the blue section, you will see data relating to the query term you supplied. Below, this you will see a section for each ortholog that has been returned. For both the query and orthologs you should see a gene symbol, a gene name, a locus type, a chromosomal location and a set of gene resources. The gene symbol and gene name will be prefixed with either 'Approved' to indicate that it was assigned by a nomenclature committee or 'Gene' to indicate that it is not an approved symbol or name. A non-approved symbol may in some cases be listed as 'Unknown' when no symbol has been assigned to that gene. Hovering your mouse over the information icons in the results panel will tell you where the data originate. For each orthologous gene you will also see a section labelled 'Assertion derived from:' that contains a series of icons. These icons represent the sources that support this orthology assignment. Clicking on the icon will take you to the entry for that gene in the orthology database, while hovering over the icon will give you the full name of the ortholog data source.

In addition to the new user interface, we also provide precalculated files of HCOP data that can be downloaded from our FTP site or via the 'Bulk downloads' section (see Figure 2c) of the HCOP page. You have the option of getting a file containing human and ortholog data from a single species, or human and ortholog data from all HCOP species in a single file. For the human—single ortholog species files the '6 Column' output returns the raw assertions, Ensembl gene IDs and Entrez Gene IDs for human and one other species, while the '15 Column' output includes additional information such as the chromosomal location, accession numbers and, where possible, references the approved gene nomenclature. The files containing all species' ortholog data have an additional column at the start giving the taxon id for each species.

FUTURE DIRECTIONS

We will continue to improve and add functionality to our website, in particular to the gene search application and to our REST service. One such planned addition is the ability to search/retrieve genes by chromosome. External links within symbol reports will continuously be reviewed and new links to noteworthy resources added, such as ClinVar and GeneReviews (http://www.genereviews.org).

A major change to the site in the near future is a completely new gene families section. As mentioned earlier in the article, our gene families' data are growing in size and complexity. Some of our large gene families contain hierarchies such as 'Class A GPCRs, rhodopsin-type' which has subset families (e.g 'Olfactory receptors') and is itself a subset of the 'G protein-coupled receptors' family. Our new gene family pages will display the relationships between subsets better and allow users to browse through the hierarchy. Many families will contain a brief description and may contain a graphical display of the protein domains of a typical family member. With a better built in hierarchy within the new pages we can also provide a much improved way of downloading gene families as data sets, allowing people to choose between downloading one subset family or the entire family hierarchy. The new gene families will be searchable via our Solr search engine, and the index page for the families will have a searchable and sortable, paged table all of which we hope will improve our users' experience.

In addition the HGNC have recently been funded to expand our remit to naming genes across all vertebrate species that do not have an official gene nomenclature committee. The naming will be based on the names of orthologous and paralogous human genes. These pan-vertebrate data will be accessed from a new website, with reciprocal linking to the human homologs at the current HGNC website (www.genenames.org).

ACKNOWLEDGEMENT

The authors would like to thank their former colleagues, specialist advisors and other researchers who have provided invaluable contributions to this project.

FUNDING

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

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. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Billis,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fitzgerald,S. *et al.* (2014) Ensembl 2014. *Nucleic Acids Res.*, **42**, D749–D755.
3. The UniProt Consortium (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **42**, D191–D198.
4. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2011) Entrez gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **39**, D26–D31.
5. Pakseresht,N., Alako,B., Amid,C., Cerdeño-Tárraga,A., Cleland,I., Gibson,R., Goodgame,N., Gur,T., Jang,M., Kay,S. *et al.* (2014) Assembly information services in the European Nucleotide Archive. *Nucleic Acids Res.*, **42**, D38–D43.
6. Benson,D.A., Clark,K., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2014) GenBank. *Nucleic Acids Res.*, **42**, D32–D37.
7. Kosuge,T., Mashima,J., Kodama,Y., Fujisawa,T., Kaminuma,E., Ogasawara,O., Okubo,K., Takagi,T. and Nakamura,Y. (2014) DDBJ progress report: a new submission system for leading to a correct annotation. *Nucleic Acids Res.*, **42**, D44–D49.
8. Harrow,J.L., Steward,C.A., Frankish,A., Gilbert,J.G., Gonzalez,J.M., Loveland,J.E., Mudge,J., Sheppard,D., Thomas,M., Trevanion,S. *et al.* (2014) The Vertebrate Genome Annotation browser 10 years on. *Nucleic Acids Res.*, **42**, D771–D779.
9. 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.*, doi: 10.1093/database/baq020.

10. Karolchik,D., Barber,G.P., Casper,J., Clawson,H., Cline,M.S., Diekhans,M., Dreszer,T.R., Fujita,P.A., Guruvadoo,L., Haeussler,M. *et al.* (2014) The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.*, **42**, D764–D770.
11. Bragin,E., Chatzimichali,E.A., Wright,C.F., Hurles,M.E., Firth,H. V., Bevan,A.P. and Swaminathan,G.J. (2014) DECIPHER: Database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. *Nucleic Acids Res.*, **42**, D993–D1000.
12. Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and McKusick,V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.
13. MacArthur,J.A.L., Morales,J., Tully,R.E., Astashyn,A., Gil,L., Bruford,E.A., Larsson,P., Flicek,P., Dalgleish,R., Maglott,D.R. *et al.* (2014) Locus Reference Genomic: reference sequences for the reporting of clinically relevant sequence variants. *Nucleic Acids Res.*, **42**, D873–D878.
14. Landrum,M.J., Lee,J.M., Riley,G.R., Jang,W., Rubinstein,W.S., Church,D.M. and Maglott,D.R. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*, **42**, D980–D985.
15. Pagon,R.A. (2006) GeneTests: an online genetic information resource for health care providers. *J. Med. Libr. Assoc.*, **94**, 343–348.
16. Blake,J.A., Bult,C.J., Eppig,J.T., Kadin,J.A. and Richardson,J.E. (2014) The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. *Nucleic Acids Res.*, **42**, D810–D817.
17. Laulederkind,S.J.F., Hayman,G.T., Wang,S.J., Smith,J.R., Lowry,T.F., Nigam,R., Petri,V., De Pons,J., Dwinell,M.R., Shimoyama,M. *et al.* (2013) The Rat Genome Database 2013-data, tools and users. *Brief. Bioinform.*, **14**, 520–526.
18. Bradford,Y., Conlin,T., Dunn,N., Fashena,D., Frazer,K., Howe,D.G., Knight,J., Mani,P., Martin,R., Moxon,S.A.T. *et al.* (2011) ZFIN: enhancements and updates to the zebrafish model organism database. *Nucleic Acids Res.*, **39**, D822–D829.
19. Gray,K.A., Daugherty,L.C., Gordon,S.M., Seal,R.L., Wright,M.W. and Bruford,E.A. (2013) Genenames.org: The HGNC resources in 2013. *Nucleic Acids Res.*, **41**, D514–D519.
20. Ezkurdia,I., Juan,D., Rodriguez,J.M., Frankish,A., Diekhans,M., Harrow,J., Vazquez,J., Valencia,A. and Tress,M.L. (2014) Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Hum. Mol. Genet.*, **23**, 5866–5878
21. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S. *et al.* (2012) GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**, 1760–1774.
22. Farrell,C.M., O'Leary,N.A., Harte,R.A., Loveland,J.E., Wilming,L.G., Wallin,C., Diekhans,M., Barrell,D., Searle,S.M.J., Aken,B. *et al.* (2014) Current status and new features of the Consensus Coding Sequence database. *Nucleic Acids Res.*, **42**, D865–D872.
23. Pruitt,K.D., Tatusova,T., Brown,G.R. and Maglott,D.R. (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, **40**, D130–D135.
24. Wright,M.W. (2014) A short guide to long non-coding RNA gene nomenclature. *Hum. Genomics*, **8**, 7.
25. 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.S. *et al.* (2011) Modernizing reference genome assemblies. *PLoS Biol.*, **9**, e1001091.
26. Seal,R.L., Wright,M.W., Gray,K.A. and Bruford,E.A. (2013) Vive la différence: naming structural variants in the human reference genome. *Hum. Genomics*, **7**, 12.
27. Gutmanas,A., Alhroub,Y., Battle,G.M., Berrisford,J.M., Bochet,E., Conroy,M.J., Dana,J.M., Fernandez Montecelo,M.A., Van Ginkel,G., Gore,S.P. *et al.* (2014) PDBe: protein data bank in Europe. *Nucleic Acids Res.*, **42**, D402–D410.
28. 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.
29. Javitt,G., Katsanis,S., Scott,J. and Hudson,K. (2010) Developing the blueprint for a genetic testing registry. *Public Health Genomics*, **13**, 95–105.
30. Wu,C., MacLeod,I. and Su,A.I. (2013) BioGPS and MyGene.info: organizing online, gene-centric information. *Nucleic Acids Res.*, **41**, D561–D565.
31. Acland,A., Agarwala,R., Barrett,T., Beck,J., Benson,D.A., Bollin,C., Bolton,E., Bryant,S.H., Canese,K., Church,D.M. *et al.* (2014) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **42**, D7–D17.
32. 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.
33. 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.
34. Powell,S., Forslund,K., Szklarczyk,D., Trachana,K., Roth,A., Huerta-Cepas,J., Gabaldón,T., Rattei,T., Creevey,C., Kuhn,M. *et al.* (2014) EggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res.*, **42**, D231–D239.
35. Östlund,G., Schmitt,T., Forslund,K., Köstler,T., Messina,D.N., Roopra,S., Frings,O. and Sonnhammer,E.L.L. (2010) Inparanoid 7: New algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.*, **38**, D196–D203.
36. Altenhoff,A.M., Schneider,A., Gonnet,G.H. and Dessimoz,C. (2011) OMA 2011: Orthology inference among 1000 complete genomes. *Nucleic Acids Res.*, **39**, D289–D294.
37. Waterhouse,R.M., Tegenfeldt,F., Li,J., Zdobnov,E.M. and Kriventseva,E. V. (2013) OrthoDB: A hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res.*, **41**, D358–D365.
38. Chen,F., Mackey,A.J., Stoeckert,C.J. and Roos,D.S. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.*, **34**, D363–D368.
39. Mi,H., Muruganujan,A. and Thomas,P.D. (2013) PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.*, **41**, D377–D386.
40. Huerta-Cepas,J., Capella-Gutiérrez,S., Pryszcz,L.P., Marcet-Houben,M. and Gabaldón,T. (2014) PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res.*, **42**, D897–D902.
41. Schreiber,F., Patricio,M., Muffato,M., Pignatelli,M. and Bateman,A. (2014) TreeFam v9: a new website, more species and orthology-on-the-fly. *Nucleic Acids Res.*, **42**, D922–D925.
42. Severin,J., Beal,K., Vilella,A.J., Fitzgerald,S., Schuster,M., Gordon,L., Ureta-Vidal,A., Flicek,P. and Herrero,J. (2010) eHive: an artificial intelligence workflow system for genomic analysis. *BMC Bioinformatics*, **11**, 240.
43. James-Zorn,C., Ponferrada,V.G., Jarabek,C.J., Burns,K.A., Segerdell,E.J., Lee,J., Snyder,K., Bhattacharyya,B., Karpinka,J.B., Fortriede,J. *et al.* (2013) Xenbase: expansion and updates of the Xenopus model organism database. *Nucleic Acids Res.*, **41**, D865–D870.