Human Genomics

GENOME DATABASE Open Access

Gene family matters: expanding the HGNC resource

Louise C Daugherty*, Ruth L Seal, Mathew W Wright and Elspeth A Bruford

Abstract

The HUGO Gene Nomenclature Committee (HGNC) assigns approved gene symbols to human loci. There are currently over 33,000 approved gene symbols, the majority of which represent protein-coding genes, but we also name other locus types such as non-coding RNAs, pseudogenes and phenotypic loci. Where relevant, the HGNC organise these genes into gene families and groups. The HGNC website http://www.genenames.org/ is an online repository of HGNC-approved gene nomenclature and associated resources for human genes, and includes links to genomic, proteomic and phenotypic information. In addition to this, we also have dedicated gene family web pages and are currently expanding and generating more of these pages using data curated by the HGNC and from information derived from external resources that focus on particular gene families. Here, we review our current online resources with a particular focus on our gene family data, using it to highlight our new Gene Symbol Report and gene family data downloads.

The HGNC: background and relevance

The HUGO Gene Nomenclature Committee (HGNC) has been responsible for approving unique and informative gene names and symbols to every human gene for over 30 years. Approved gene names and symbols preferably describe the structure, function or homology of a gene and its products. The provision of approved nomenclature allows researchers to discuss genes unambiguously, and this is reflected by HGNC symbol usage in scientific papers describing human genes, hence aiding the dissemination and interpretation of the associated data by the scientific community.

The HGNC website [1,2] provides direct links to genomic, proteomic and phenotypic information that is held in the HGNC database and enables users to search and download current data associated to their gene(s) of interest. As of February 2012, there are over 33,000 approved human gene symbols (including protein-coding genes, pseudogenes, ncRNA genes and phenotypes), each with a publicly available Gene Symbol Report. It is important to note that although the main focus of HGNC concerns human genes, there are coordinated efforts with other nomenclature committees [3], in particular the Mouse Genomic Nomenclature Committee (MGNC) [4] and Rat Genome Database (RGD) [5], and any large new gene family reorganisation or assignment is usually coordinated among these three nomenclature groups. The HGNC also regularly works with specialist advisors and publish scientific papers concerning gene family nomenclature and gene grouping [6-9]. The adoption of HGNC-approved gene names/symbols by the many genome browsers and databases reduces any uncertainty when referring to genes; for example, Ensembl [10], Entrez Gene [11], GeneCards [12], OMIM [13], UCSC [14], UniProt [15] and Vega [16] all use HGNC names/symbols. The data supplied by HGNC have been applied to a range of studies, such as assisting tools to identify candidate genes for further study [17], quantitative assessments of gene annotation status [18] and projects involving 'mashups' of bioinformatics data to explore genes involved in a particular disease [19] to name just a few.

Website updates: the revamped genenames.org website

Recently, we undertook a revamp of our website [1] to improve the navigation and access to our tools and data. The new HGNC homepage (Figure 1a) now has informative tabs with drop-down menus, enabling users to navigate to specific HGNC pages. The banner includes a simplified 'Quick Gene Search' that searches for terms that contain any part of the term and is present at the top of all pages.

* Correspondence: hgnc@genenames.org
European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK

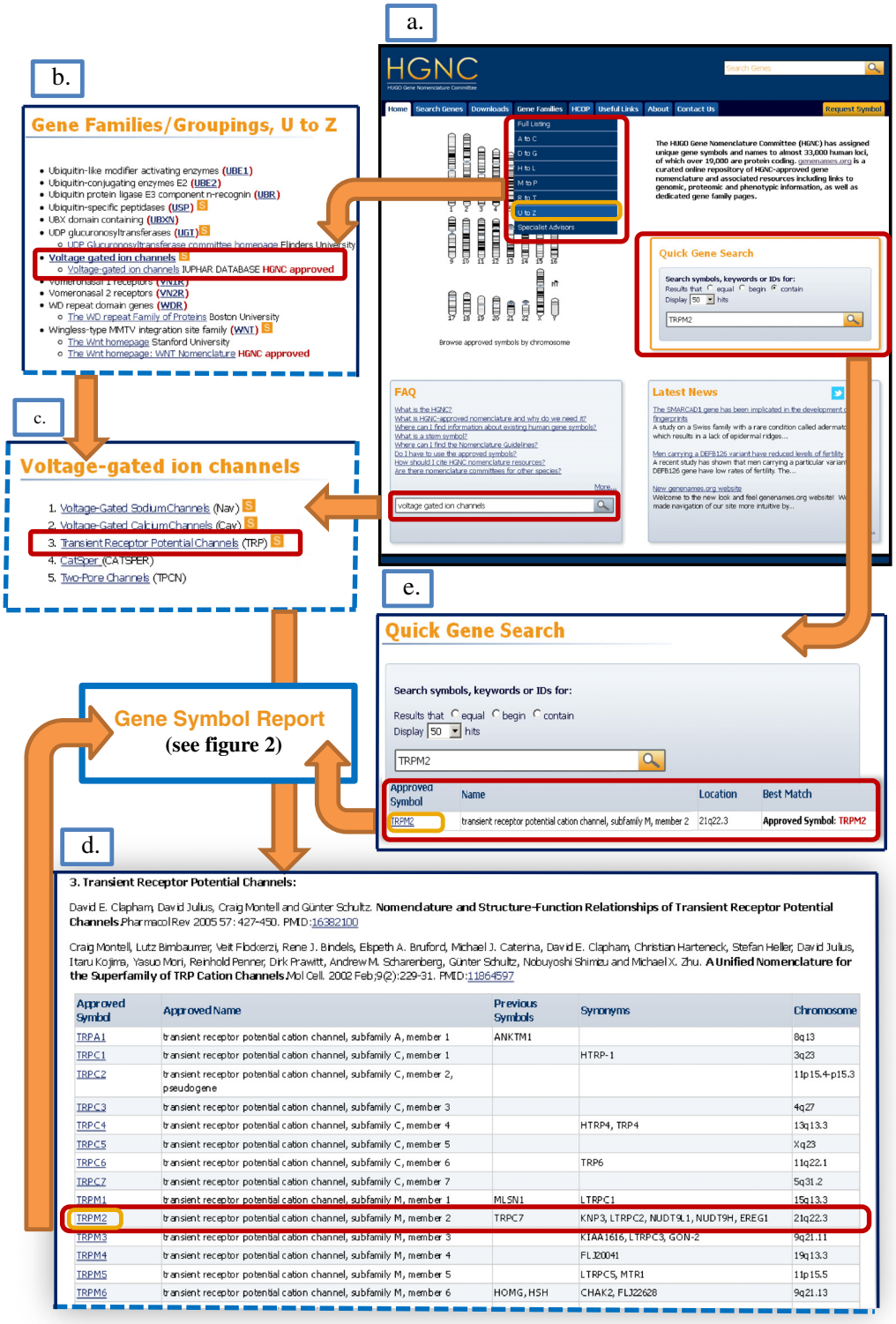

Figure 1 Finding a gene family at genenames.org. Dotted lines indicate where the webpage screenshot has been cut for brevity. (**1a**) Searching for a gene family on the HGNC homepage. (**1b**) Alphabetical listing of gene family names with links to the associated gene family pages. (**1c**) Large gene families are further divided into subgroups. (**1d**) Gene family pages and associated information are organised and represented by a table. (1e) Quick Gene Search tool is a quick way to find whether gene of interest is associated to a gene family.

The 'Quick Gene Search' is also found on the homepage, and this allows users to search for terms that equal, begin or contain the search term and to select the number of returned results displayed [2]. At the bottom left of the homepage, there is now a HGNC website search box that allows searching of gene family pages and HGNC documentation and information pages. The homepage now has an interactive graphic of human chromosomes and mitochondrion, which allows users to 'browse approved symbols by chromosome'. Each chromosome in the display is linked to the HGNC 'Statistics and Downloads' page, and the data returned are associated to the chromosome selected. This feature enables users to download the total number of approved symbols for the chromosome or browse symbols by locus type. Another recent addition to the homepage is the 'News' section that highlights any significant updates to the HGNC project and links to recent media articles that use approved gene symbols in their reports.

One of the main updates is the redesign of the Gene Symbol Report, which contains HGNC-curated data and data derived from external resources. All sections and corresponding links are described in full in our Symbol Report Documentation [20]. In this report, we will use *TRPM2* as an example to highlight the new features found in the report (Figure 2) and gene family pages (Figure 1).

We have restructured the Gene Symbol Report, so that the HGNC 'core' data are now more prominent and are presented in a separate table at the top of each report. The Gene Symbol Report for *TRPM2* (Figure 2) displays data in all the core fields. One of the most significant changes to our Gene Symbol Report is that we now provide access to the gene family page(s) linking from the 'Gene Family' name on the report to the associated gene family page. As shown in Figure 2, *TRPM2* is associated to two gene families: the subgroup 'Transient receptor potential channels' and the 'Nudix motif-containing family'. Gene families will be discussed in more detail later. Gene Symbol Reports also contain links to external biomedical resources. We have grouped related resources into the following sections: 'Specialist Database', 'Homologs', 'Nucleotide Sequences', 'Gene Resources', 'Protein Resources', 'Clinical Resources', 'References' and 'Other Database Links'. External links can either be manually 'curated' by an HGNC curator, which is denoted by the letter 'C', or 'downloaded' from external sources, which is denoted by the letter 'D'. The 'Specialist Database' section provides links to databases that are relevant to only certain classes of gene, and we now link to 14 specialist databases. In the example Gene Symbol Report for *TRPM2* (Figure 2), the specialist receptor database we link to is IUPHAR [21]. In the 'Homologs' section, in addition to linking to the mouse MGI [4] and rat RGD [5] databases, we now display the symbols approved by the two nomenclature committees. The nomenclature committees for human, mouse and rat aim to approve equivalent gene symbols and names for orthologous genes; for example, the human *TRPM2* Gene Symbol Report (Figure 2) shows that the approved symbols for mouse and rat are both *Trpm2*. Links to nucleotide sequences are grouped in the 'Nucleotide Sequences' section and include recently added links to Vertebrate Genome Annotation [16] gene sequence curated by the Havana project. The 'Gene Resources' section groups together links to the gene annotation pages at Entrez Gene [11], Ensembl [10], UCSC [14] and Vega [16]. As part of the Symbol Report redesign, we now also provide direct links to the Genome Browsers supported by these four projects. The 'Protein Resources' section still includes links to the UniProt project, but as part of the update, we have added a link to the InterPro [22] Protein Match page; this shows all predicted protein signatures (integrated and unintegrated) for the encoded protein by the InterPro member databases. All the mutation and variation-related data links are displayed in the 'Clinical Resources' section, while our curated links to references in PubMed [23] and CiteXplore [24] are shown in the 'References' section. Finally, the 'Other Database Links' section includes links to relevant biomedical resources that cannot be grouped into the categories above. For example, we now also link to the Reactome signalling pathway database [25] and to a list of all Gene Ontology terms annotated for the gene product at the QuickGO project [26].

Gene families and groupings

Gene families are generally defined as a group of genes descended by duplication from a common ancestor. The degree of divergence from the ancestral gene can vary considerably between members, and they may or may not have a conserved function. In many cases, the homology between the family members may be restricted to a specific highly conserved region or domain(s) of the encoded protein, and genes can therefore belong to more than one gene family, e.g. *TRPM2* belongs to two gene families: the 'Transient receptor potential channels' and the 'Nudix motif-containing family' (see Figure 2). Large gene families may also be subdivided into smaller subfamilies, which often equate to functional groups. The 'Transient receptor potential channels' are a good example of this as they are subdivided into subfamily A, subfamily C, subfamily M and subfamily V [27].

When naming gene family members, the HGNC aims to use a common root (or stem) symbol which allows easy identification of the members, e.g. TRP is the root symbol for Transient receptor potential channels. As well as homologous gene families, the HGNC also organises human genes into gene groupings, which correspond to sets of genes that are not necessarily related by sequence homology but do have another shared feature: for example, a common function (e.g. 'class I' and 'class II aminoacyl tRNA synthetases'), a specific chromosomal location (e.g. genes

Figure 2 Gene Symbol Report for *TRPM2*.

from the 'pseudoautosomal region'), secondary structure (e.g. 'micro RNAs') or a grouping that provides a useful community resource (e.g. all genes encoding 'blood group' antigens). These groupings do not usually share a common root symbol, and again, one gene can be a member of more than one gene family and/or grouping. When assigning genes as members of families and groupings, the HGNC look at all available data including sequence similarity and conserved domain structure, publications and other databases, and where possible, take advice from specialist advisors who are experts working on that specific family or group. For ease of discussion, all gene families and groupings are referred to simply as gene families throughout this paper.

As of June 2012, close to 45% of the 33,000 HGNC database entries are associated with a gene family. The HGNC website [1] also currently displays over 237 curated webpages dedicated to individual gene families. However, if all the gene subfamilies are considered, this would give a larger total of around 400 pages due to instances where large gene families have been subdivided into smaller subgroups. The 'Voltage-gated ion channels'

family (Figure 1c) is an example of a large gene family which has the following 11 subgroupings that make up the gene family: 'Voltage-Gated Sodium Channels', 'Voltage-Gated Calcium Channels', 'Voltage-Gated Sodium Channels', 'CatSper', 'Two-Pore Channels', 'Cyclic Nucleotide-Regulated Channels', 'Calcium-Activated Potassium Channels', 'Voltage-Gated Potassium Channels', 'Inwardly Rectifying Potassium Channels', 'Two-P Potassium Channels' and 'Hydrogen Voltage-Gated Ion Channels').

Gene family resources: finding the information

HGNC is now in the process of actively expanding our gene family pages using internally curated data and data from the growing number of external resources and publications that focus on particular gene families. The HGNC homepage [1] (Figure 1a) features a website search box at the bottom of the page that allows searching for gene family pages, HGNC documentation and information pages. Furthermore, there are two main ways of ascertaining whether a gene is associated to a particular gene family: by browsing the 'Gene Families' pages [28] from the homepage or by directly querying the database by using one of the 'Search Genes' tools [29].

The 'Gene Families' [28] section from the drop-down menu on the homepage (Figure 1a.) gives an alphabetical listing of all the gene family names (Figure 1b) and links to the associated gene family pages. The information grouped on the gene family page is organised and represented by a table (Figure 1d), which lists all of the associated genes with the following data: 'Approved Symbol', 'Approved Name', 'Previous Symbols', 'Synonyms' and 'Chromosome'. The 'Approved Symbol' links to the 'Gene Symbol Report', so selecting 'TRPM2' from the table will take the user to the specific Gene Symbol Report (Figure 2). The gene family page (Figure 1c,d) also indicates if there is a specialist advisor associated to the gene family, and their contact details, where applicable, are linked from each relevant gene family page by an orange 'S' icon. There are currently 115 specialist advisors that help with the content and approval of any new gene family members, and a separate page on the website lists all of HGNC's specialist advisors [30].

Users can also query the HGNC data by using the tools listed on the 'Search Genes' page [29]: the 'Quick Gene Search', 'Advanced Gene Search' and the 'List Search'. These tools can all be used to obtain the Gene Symbol Report, so if the gene of interest has a gene family associated, it will link to the relevant gene family page from the 'Gene Family' name in the Gene Symbol Report. The 'Quick Gene Search' tool [31] (see Figure 1e) is a quick and easy way to check whether a gene of interest is associated to a gene family; this is done by querying the gene symbol, name or ID to locate a Gene Symbol Report. Using the 'Advanced Gene Search' [32] will allow the user to build a more specific query by choosing to query within specific datasets, e.g. only for those genes that are approved. The 'List Search' [33] also allows users to access the Gene Symbol Report but enables users to search multiple genes by gene symbol.

Gene family resources: downloading the information

We have recently developed new and additional ways to download our gene family data, depending on the scope of data required. From the 'Downloads' tab, users can access the following:

- The 'Statistics and Downloads' page [34], which has a direct link to the 'complete HGNC dataset'. This file includes all the HGNC core fields and now includes the new gene family data fields, 'Gene Family Tag' and 'Gene Family Description' (discussed below). Alternatively, if it is just the information relating to the gene family data that is required, select 'complete HGNC Gene Family dataset'. This gives a file with the following fields: URL for gene family page, Gene Family Tag, Gene Family Description, Symbol and HGNC ID.
- The 'Custom Downloads' page, on the other hand, allows the specification of the exact fields required. The selection should include the 'Gene Family Description' and the 'Gene Family Tag' fields to retrieve the gene family data in the output.
- Finally, to obtain the data for the gene family associated to a particular gene of interest, select 'Download gene family data' beneath the gene family table on the relevant gene family page.

The data from our previously established 'complete HGNC dataset' and the newly created 'complete HGNC Gene Family dataset' can be associated together by the 'Gene Family Tag' field and the 'Gene Family Description' field. The 'Gene Family Tag' is used to generate gene family or group specific pages at the HGNC website [1] and does not necessarily reflect an official nomenclature. The 'Gene Family Description' is the name given to a particular gene family. Each 'Gene Family Description' has an associated 'Gene Family Tag' and vice versa. If a particular gene is a member of more than one gene family, the tags and the descriptions will be shown in the same order. Like all HGNC data, the 'Gene Family' datasets are updated daily, so the user will always get the most recent and up-to-date data when clicking on the download links or accessing custom downloads.

Future directions

In the future, we plan to increase the number of gene families and the assignment of human genes to gene families. We will also arrange the current alphabetical gene family list into more meaningful categories, for

example, grouping them by domain structure, function or disease associations. Another area we are considering is integrating the graphical display in InterPro [22] to represent the domains that are encoded by each gene family member.

Updates regarding the gene family pages are now also mentioned in our Twitter feed and quarterly 'Newsletter'. If you subscribe to this, you will be notified of any new gene families; see our feedback form [35] and tick the box to receive the Newsletter. Alternatively, newsletters are also available on the website. If you have a gene family you think should be represented or you would like to be considered as one our specialist advisors, please contact us via hgnc@genenames.org.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

Louise Daugherty would like to thank HGNC for the opportunity to contribute to and work with the team. This work was supported by the Wellcome Trust (081979/Z/07/Z) and the National Human Genome Research Institute (P41 HG03345).

Authors' contributions

LCD drafted the manuscript and created the new gene family pages. MWW, RLS and EAB curated the data in the HGNC dataset. All authors read and approved the final manuscript.

Received: 11 August 2011 Accepted: 5 July 2012
Published: 5 July 2012

References

1. *HGNC Database*. http://www.genenames.org.
2. Seal RL, Gordon SM, Lush MJ, Wright MW, Bruford EA: **genenames.org: the HGNC resources in 2011.** *Nucleic Acids Res* 2011, **39**(Database issue):D514–519.
3. Lefranc MP: **From IMGT-ONTOLOGY CLASSIFICATION Axiom to IMGT standardized gene and allele nomenclature: for immunoglobulins (IG) and T cell receptors (TR).** *Cold Spring Harb Protoc* 2011, **2011**(6):627–632.
4. Blake JA, Bult CJ, Kadin JA, Richardson JE, Eppig JT: **The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics.** *Nucleic Acids Res* 2011, **39**(Database issue):D842–848.
5. Twigger SN, Shimoyama M, Bromberg S, Kwitek AE, Jacob HJ: **The Rat Genome Database, update 2007--easing the path from disease to data and back again.** *Nucleic Acids Res* 2007, **35**(Database issue):D658–662.
6. Mayer J, Blomberg J, Seal RL: **A revised nomenclature for transcribed human endogenous retroviral loci.** *Mob DNA* 2011, **2(1):**7.
7. Bingle CD, Seal RL, Craven CJ: **Systematic nomenclature for the PLUNC/PSP/BSP30/SMGB proteins as a subfamily of the BPI fold-containing superfamily.** *Biochem Soc Trans* 2011, **39**(4):977–983.
8. Holmes RS, Wright MW, Laulederkind SJ, Cox LA, Hosokawa M, Imai T, Ishibashi S, Lehner R, Miyazaki M, Perkins EJ, Potter PM, Redinbo MR, Robert J, Satoh T, Yamashita T, Yan B, Yokoi T, Zechner R, Maltais LJ: **Recommended nomenclature for five mammalian carboxylesterase gene families: human, mouse, and rat genes and proteins.** *Mamm Genome* 2010, **21**(9–10):427–441.
9. Persson B, Kallberg Y, Bray JE, Bruford E, Dellaporta SL, Favia AD, Duarte RG, Jornvall H, Kavanagh KL, Kedishvili N, Kisiela M, Maser E, Mindnich R, Orchard S, Penning TM, Thornton JM, Adamski J, Oppermann U: **The SDR (short-chain dehydrogenase/reductase and related enzymes) nomenclature initiative.** *Chem Biol Interact* 2009, **178**(1–3):94–98.
10. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Gordon L, Hendrix M, Hourlier T, Johnson N, Kahari A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Larsson P, Longden I, McLaren W, Overduin B, Pritchard B, Riat HS, Rios D, Ritchie GR, Ruffier M, Schuster M, *et al*: **Ensembl 2011.** *Nucleic Acids Res* 2011, **39**(Database issue):D800–806.
11. Maglott D, Ostell J, Pruitt KD, Tatusova T: **Entrez Gene: gene-centered information at NCBI.** *Nucleic Acids Res* 2011, **39**(Database issue):D52–57.
12. Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, Nativ N, Bahir I, Doniger T, Krug H, Sirota-Madi A, Olender T, Golan Y, Stelzer G, Harel A, Lancet D: **GeneCards Version 3: the human gene integrator.** *Database (Oxford)* 2010, **2010:**baq020.
13. Amberger J, Bocchini C, Hamosh A: **A new face and new challenges for Online Mendelian Inheritance in Man (OMIM(R)).** *Hum Mutat* 2011, **32**(5):564–567.
14. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, Diekhans M, Dreszer TR, Giardine BM, Harte RA, Hillman-Jackson J, Hsu F, Kirkup V, Kuhn RM, Learned K, Li CH, Meyer LR, Pohl A, Raney BJ, Rosenbloom KR, Smith KE, Haussler D, Kent WJ: **The UCSC Genome Browser database: update 2011.** *Nucleic Acids Res* 2011, **39**(Database issue):D876–882.
15. UniProt Consortium: **The Universal Protein Resource (UniProt) in 2010.** *Nucleic Acids Res* 2010, **38**(Database issue):D142–148.
16. Wilming LG, Gilbert JG, Howe K, Trevanion S, Hubbard T, Harrow JL: **The vertebrate genome annotation (Vega) database.** *Nucleic Acids Res* 2008, **36**(Database issue):D753–760.
17. Lecerf F, Bretaudeau A, Sallou O, Desert C, Blum Y, Lagarrigue S, Demeure O: **AnnotQTL: a new tool to gather functional and comparative information on a genomic region.** *Nucleic Acids Res* 2011, **39**(Web Server issue):W328–333.
18. Harel A, Inger A, Stelzer G, Strichman-Almashanu L, Dalah I, Safran M, Lancet D: **GIFtS: annotation landscape analysis with GeneCards.** *BMC Bioinformatics* 2009, **10:**348.
19. Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J: **Bio2RDF: towards a mashup to build bioinformatics knowledge systems.** *J Biomed Inform* 2008, **41**(5):706–716.
20. *HGNC Symbol Report Documentation*. http://www.genenames.org/useful/symbol-report-documentation.
21. Sharman JL, Mpamhanga CP, Spedding M, Germain P, Staels B, Dacquet C, Laudet V, Harmar AJ: **IUPHAR-DB: new receptors and tools for easy searching and visualization of pharmacological data.** *Nucleic Acids Res* 2011, **39**(Database issue):D534–538.
22. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, *et al*: **InterPro: the integrative protein signature database.** *Nucleic Acids Res* 2009, **37**(Database issue):D211–215.
23. *PubMed*. http://www.ncbi.nlm.nih.gov/pubmed.
24. *CiteXplore*. http://www.ebi.ac.uk/citexplore.
25. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal B, Jupe S, Kalatskaya I, Mahajan S, May B, Ndegwa N, Schmidt E, Shamovsky V, Yung C, Birney E, Hermjakob H, D'Eustachio P, Stein L: **Reactome: a database of reactions, pathways and biological processes.** *Nucleic Acids Res* 2011, **39**(Database issue):D691–697.
26. Barrell D, Dimmer E, Huntley RP, Binns D, O'Donovan C, Apweiler R: **The GOA database in 2009--an integrated Gene Ontology Annotation resource.** *Nucleic Acids Res* 2009, **37**(Database issue):D396–403.
27. Montell C, Birnbaumer L, Flockerzi V, Bindels RJ, Bruford EA, Caterina MJ, Clapham DE, Harteneck C, Heller S, Julius D, Kojima I, Mori Y, Penner R, Prawitt D, Scharenberg AM, Schultz G, Shimizu N, Zhu MX: **A unified nomenclature for the superfamily of TRP cation channels.** *Mol Cell* 2002, **9**(2):229–231.
28. *HGNC Gene Families*. http://www.genenames.org/genefamilies.
29. *HGNC Searches*. http://www.genenames.org/hgnc-searches.
30. *HGNC Specialist Advisors*. http://www.genenames.org/genefamilies/hgnc-specialist-advisors.
31. *HGNC Quick Gene Search*. http://www.genenames.org/cgi-bin/quick_search.pl.
32. *HGNC Advanced Gene Search*. http://www.genenames.org/cgi-bin/advanced_search.pl.
33. *HGNC List Search*. http://www.genenames.org/cgi-bin/hgnc_bulkcheck.pl.
34. *HGNC Statistics and Downloads*. http://www.genenames.org/cgi-bin/hgnc_stats.pl.
35. *HGNC Feedback Form*. http://www.genenames.org/cgi-bin/hgnc_feedback.pl.

doi:10.1186/1479-7364-6-4
Cite this article as: Daugherty *et al.*: **Gene family matters: expanding the HGNC resource.** *Human Genomics* 2012 **6**:4.