Published online 1 December 2011 *Nucleic Acids Research, 2012, Vol. 40, Database issue* ***D465–D471***
doi:10.1093/nar/gkr1181

Gene3D: a domain-based resource for comparative genomics, functional annotation and protein network analysis

Jonathan Lees*, Corin Yeats, James Perkins, Ian Sillitoe, Robert Rentzsch, Benoit H. Dessailly and Christine Orengo

Institute of Structural and Molecular Biology, University College London, Darwin Building, Gower St, London, WC1E 6BT, UK

Received September 26, 2011; Revised November 12, 2011; Accepted November 14, 2011

Downloaded from https://academic.oup.com/nar/article/40/D1/D465/2903340 by guest on 25 September 2024

ABSTRACT

Gene3D http://gene3d.biochem.ucl.ac.uk is a comprehensive database of protein domain assignments for sequences from the major sequence databases. Domains are directly mapped from structures in the CATH database or predicted using a library of representative profile HMMs derived from CATH superfamilies. As previously described, Gene3D integrates many other protein family and function databases. These facilitate complex associations of molecular function, structure and evolution. Gene3D now includes a domain functional family (FunFam) level below the homologous superfamily level assignments. Additions have also been made to the interaction data. More significantly, to help with the visualization and interpretation of multi-genome scale data sets, we have developed a new, revamped website. Searching has been simplified with more sophisticated filtering of results, along with new tools based on Cytoscape Web, for visualizing protein–protein interaction networks, differences in domain composition between genomes and the taxonomic distribution of individual superfamilies.

INTRODUCTION

The Gene3D database (1) provides protein domain annotations for sequences from the major sequence databases Ensembl, UniProt and RefSeq (2–4). Proteins are generally composed of one or more discrete independently folding units known as domains and the CATH database (5) uses a combination of manual curation and automated evidence gathering to generate a superfamily classification of such structures in the PDB (6). An accurate HMM and graph theory-based method, DomainFinder (Yeats, Redfern and Orengo, manuscript in revision), is used to identify and resolve the boundaries of predicted domains. The new release of Gene3D (v10.2) provides over 16 million predicted domains from 2549 CATH superfamilies in 60% of approximately 15 million scanned sequences. This is an increase of 5% in domain annotation coverage compared with our last review in NAR (1). Gene3D domain annotations are provided via the Gene3D website (http://gene3d.biochem.ucl.ac.uk), the CATH-Gene3D DAS (http://gene3d.biochem.ucl.ac.uk/Gene3D/Das), RESTful web services at http://gene3d.biochem.ucl.ac.uk/WebServices/ (7) and InterPro (8).

Protein domains, and distinct combinations of them, are considered the primary building blocks of protein function evolution. The assignment of domains to a protein can help identifying functionally important residues from distant homologues (9) provide mechanistic explanations for the effects of sequence polymorphisms (10) and enable the 'inheritance' of interactions from homologues (11,12). To enhance the domain annotations generated by Gene3D, we also integrate many other complementary data sources. These include molecular and pathway function annotations from GO (13), taxonomic information from the NCBI (14) and drug targets from DrugBank (15).

At least 10% of the superfamilies in CATH are functionally highly diverse. Since these superfamilies are also highly populated, accounting for more than half of the domain annotations in Gene3D, we have subdivided them into functionally coherent families, FunFams, derived using

*To whom correspondence should be addressed. Tel: 02076793890; Fax: 02076797193; Email: lees@biochem.ucl.ac.uk

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

© The Author(s) 2011. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

an in-house protocol (16). Due to the well-established correlation between structural and functional similarity, we also include information on structurally similar clusters (SCs) of domains (17). These two new types of annotation have been designed to help homology-based function transfer and the study of function evolution, especially in the larger, functionally diverse superfamilies.

SUMMARY OF PRINCIPAL CHANGES

The major change in the latest release of Gene3D is the completely revamped website, which allows visualization of genome wide family distributions and allows biologists to easily review the known functional information for multiple proteins and examine their interaction partners in the context of domain families. It also allows genome researchers to identify over-represented protein families and their functions. Below, we describe the database updates and give some common use case scenarios.

Database updates

We have extended our previous sequence set (UniProtKb, RefSeq and Ensembl) and we now provide domain assignments for all of Ensembl-Protists, Ensembl-Plants, Ensembl-Fungi, Ensembl-Metazoa, Ensembl-Bacteria and UniProt splice variants—a total of 14 963 305 protein sequences. For these, approximately 16 million domains were found, covering 60% of the sequences.

FunFams are constructed using an in-house protocol (16), which involves profile–profile-based clustering of domain sequences in each superfamily to identify functional families. We have recently improved the speed and assignment quality of the method so that it is possible to identify FunFams in all Gene3D superfamilies. After family identification, an HMM is built for each FunFam using HMMER (18), in conjunction with a model-specific bitscore threshold based on the score attained by the most remote member sequence.

Structural Clusters (SCs) are also useful for suggesting functional annotations for uncharacterized sequences, as they group together relatives with significant structural similarity and therefore likely to have related functions. SCs are generated by clustering highly similar CATH domain structures. Similarities are calculated using CATHEDRAL (19) and clustered with a threshold of 5 Å normalized root mean square deviation (RMSD). The clustering used in the work is based on an in-group clustering algorithm. Like complete-linkage clustering, it is an agglomerative, hierarchical clustering method in which clusters are joined together if and only if their least similar members meet the cutoff (hence avoiding the chaining associated with single-linkage clustering). However, it differs slightly from complete-linkage clustering in that the order in which clusters are joined is based on the most similar pairs rather than the least similar pairs. For each SC, an HMM is built again using HMMER (18), and is used to scan the sequences in the respective Gene3D superfamily. Both the SC and FunFam methods are still under active development by the CATH team and are expected to become better integrated in future releases.

Other expansions to functional data sets include the addition of protein interactions from BIOGRID (20), Reactome (21) and DIP databases (22), to supplement the previous set which comprised of Intact (23), Mint (24) and HPRD (25). Combining interaction data sets is still an important step to maximize coverage of protein interaction networks. We have added in the ability to show if a protein has a knockdown experiment from the popular and comprehensive Genome RNAi (26) database, and provide links to this database for more detailed information.

A more interactive, graphical website

The website has been re-implemented with many new features added and speed-ups provided. We have a new front page to the website providing the main search types that can be carried out. Below, we describe the several different types of pages available in the website and include example use case scenarios for each one.

Protein views

The sequence search box provides the ability to search with many types of identifiers. For several model and medically important organisms, there is an auto-complete function on gene names. There is also a taxonomic filter box with auto-complete for all Ensembl organisms in Gene3D. Searching from this box, the user can retrieve a set of proteins showing their multi-domain architectures (MDAs). The resulting page (Figure 1) shows some default information, with many potential additional annotations available that can be interactively displayed or hidden (Figure 1A). Buttons are available for domain visualization options or to retrieve protein interactions to the displayed proteins (Figure 1B). Tailoring of domain images is made possible by our update to the newer javascript domain graphics library provided by Pfam (27). A common requirement for biologists is to filter by a given sequence motif (28). Entering amino acid motifs in the sequence filter search box, as plain text or regular expressions, filters the proteins displayed in the table. An extensive list of such expressions can be found at the ELM (28) resource, for example. There is also a global search filter, where text is filtered on all columns.

Another commonly desired search query is to retrieve all of the proteins from a genome that contain a particular domain superfamily or MDA of interest. This can be achieved using the superfamily search box on the front page of the website. From here, the user enters a CATH superfamily code or name and an organism taxon ID or scientific name to restrict the sequence results to the genome of interest. If multiple CATH codes are entered, the search returns only those genes containing all of the domain superfamilies searched for, thereby providing a domain composition search. As an example of the use of search term screening, we consider the case of a biologist interested in further elucidating the function of their proteins from domain content. Searching for the genes known to be involved in late anaphase chromosome

Downloaded from https://academic.oup.com/nar/article/40/D1/D465/2903340 by guest on 25 September 2024

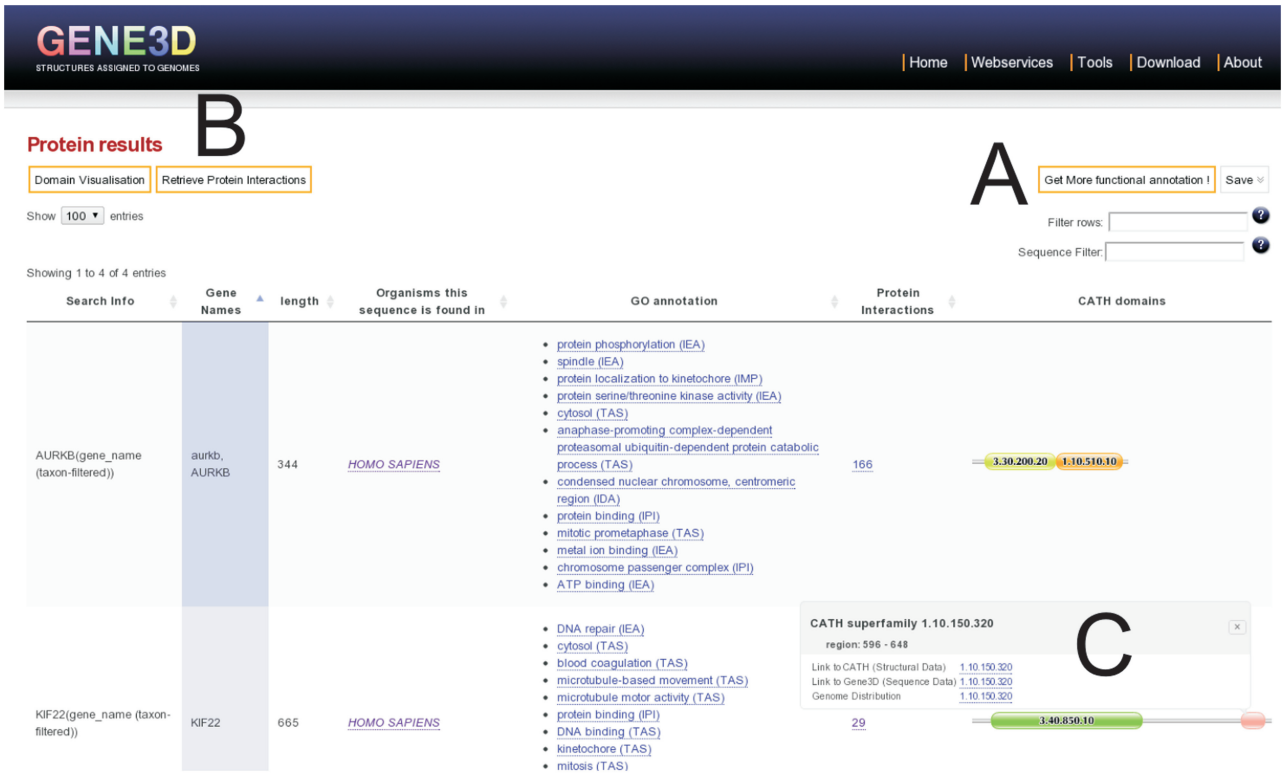

Figure 1. Results of search for genes KIF22 and AURKB. More functional annotation is available from a dropdown menu (**A**). Proteins interacting with these proteins can be displayed (**B**). Clicking on a domain image provides more details of the domain and links to structure, sequence and taxonomic information (**C**).

condensation (AURKB and KIF22), we see the domain architectures of the proteins by default (Figure 1C). More functional information is available from a dropdown menu (Figure 1A). Inspecting this GO annotation shows KIF22 to have both DNA and microtubule binding functions, which is consistent with the functions of its two domains and its possible mechanism in chromosome condensation (29).

Protein interaction data

As the wealth of protein interaction data increases, it is common for researchers to analyse their proteins of interest as a part of a system of dynamic interactions. It is also known that a large portion of protein–protein interactions (PPIs) are mediated by common domain pairings [see (12,13) for resources]. Gene3D integrates the major experimentally defined PPI databases to provide a comprehensive network analysis tool. The network can be obtained from the front page in a search box that mirrors the protein search box, or from any set of protein results (Figure 1B). The aim of this utility is not to provide protein interaction predictions, as this area is well served by popular resources such as STRING (30) and PIPs (31), but instead it focuses on complementing experimentally defined interactions with domain information and vice versa.

The network is displayed using the Cytoscape Web application (32) (Figure 2). The proteins in the network can be analysed in terms of their domain content (Figure 2B) or other protein features. Since networks can quickly become unmanageably large for viewing, the PPI page offers many different filtering options. A summary is provided, giving statistics such as the superfamily frequencies in the network. Various interactive options are available, such as to highlight the proteins in the network with a given domain, only display proteins in the table selected in the network, or select proteins in the network visible in the proteins table. Combining these options with the other filtering options provides a quick and powerful way to explore the data.

As an example, we look at the KSHV virus, a medically highly important human pathogen. By searching with several known oncogenic proteins in KSHV (LANA, vIL-6, vIRF, ORF71, K1 and K12) (33), a set of human–KSHV interactions are retrieved (Figure 2). Several of these proteins have structural protein domains predicted (Figure 2B) and we can inspect the summary of superfamilies found in the network, some of which are involved in apoptotic pathways. KSHV is known to immortalize cell lines (34) and subvert the host cell's molecular machinery. Such networks and domain annotation can provide helpful insights into this process.

CATH superfamily pages

A set of all 2549 CATH superfamilies in Gene3D and associated information with links can be found at http://gene3d.biochem.ucl.ac.uk/superfamily/. The page displays various categories of information including abundance,

Downloaded from https://academic.oup.com/nar/article/40/D1/D465/2903340 by guest on 25 September 2024

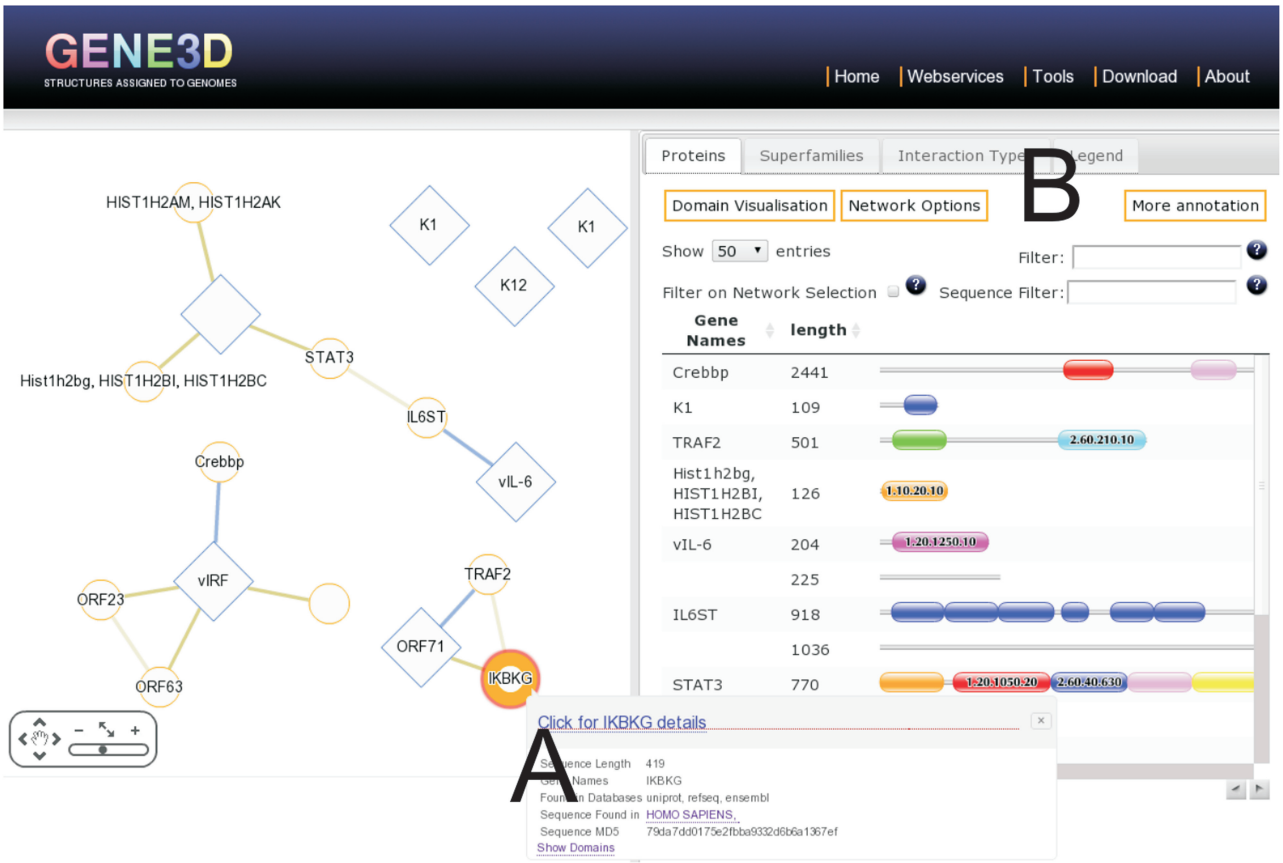

Figure 2. Result of searching for oncogenic viral proteins from KSHV for protein interactions in human (using uniprot accessions Q9DUM3, Q5G851, Q5G850, Q76RF1, Q98823, Q77Q82, D3JNC0). The panel on the left shows the protein interactions. Clicking on an edge or node generates pop-ups with more information on the protein or interaction respectively (**A**). The tabs on the right (**B**) contain information on the proteins, superfamilies and interaction types in the network.

structural diversity, functional diversity and taxonomic distribution. More detailed information on an individual CATH superfamily can be obtained from the front page superfamily search, or by simply adding the CATH code to the end of the above URL. The breakdown of the superfamily into the functional families is shown automatically. Another tab shows structurally similar sets of domains (SCs) likely to have related functions and from here it is possible to link through to CATH to see the structural representative for this SC. A third tab shows the taxonomic distribution of the family in Ensembl genomes for the CATH superfamily. The data are presented in searchable tables which can also be saved as text files.

Genome comparison

With the advent of genome sequencing, several powerful tools have been developed for genome comparison, including looking at domain content (35). Gene3D now provides a comparison tool for visualizing such differences. From the front search page, the user simply inputs the two organisms with Ensembl genomes to be compared and gets a summary of the most differing (in terms of frequency) CATH superfamilies and CATH superfamily combinations in genes between the two genomes (Figure 3), in a network representation. In this network, the nodes corresponding to superfamilies and the edges (links between nodes) indicate superfamilies that co-occur in the same gene. The sizes, colours and thicknesses of the nodes (superfamilies) and edges (superfamily combinations) indicate how different the counts are between the organisms. Contex-specific help on the webpage provides a complete description for interpreting the results.

As an example application, we can compare the non-pathogenic *Escherichia coli* K12 and the pathogenic strain O157:H7 str.TW14359 (Figure 3). We can see an over-represented domain composition 4.10.470.10-3. 40.420.10 highlighted. The context sensitive help for this page tells us that the red-line joining these families indicates that these domains co-occur in the same gene more frequently for the pathogen than non-pathogen. Clicking on the red-line joining the two superfamilies generates a pop-up allowing the user to click through to inspect the genes in the pathogen containing these domains (Figure 3A).

Genome distribution

Superfamilies range from the very ancient—such as the P-loop hydrolases, present in LUCA and found in every organism—to those that appear to have emerged more recently and are taxonomically restricted. Analysis of domain family genome distribution can be a powerful

Downloaded from https://academic.oup.com/nar/article/40/D1/D465/2903340 by guest on 25 September 2024

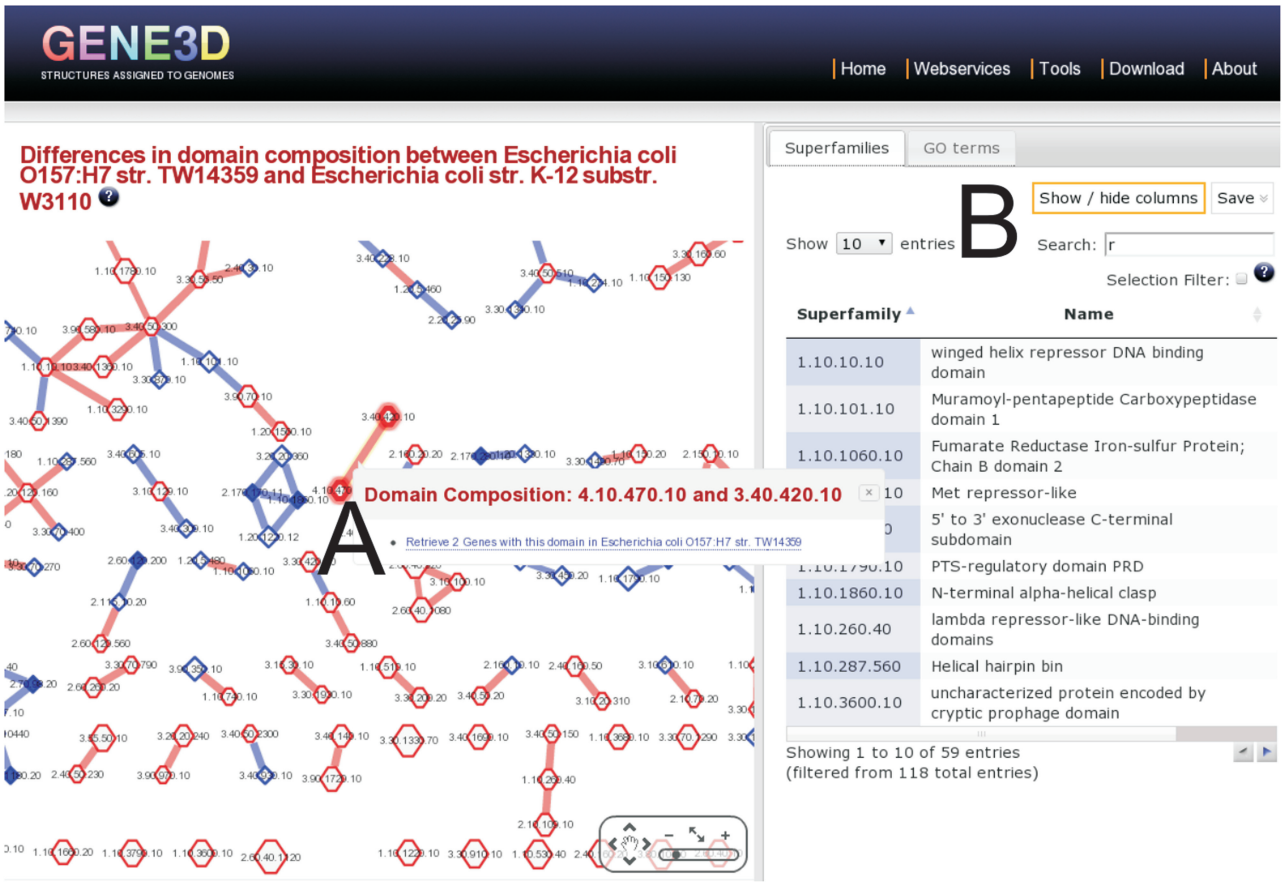

Figure 3. Domain composition comparison between closely related pathogenic and non-pathogenic strains of *E. coli*. Links back to the source genes are available by a tooltip (**A**). The function and superfamily tabs on the right, provide the ability to filter by function to help drill down to the subset of interest (**B**).

tool for evolutionary analyses (35), while superfamily phylogenetic profiling has proven to be an effective tool in identifying functionally linked proteins (36). We now provide a graphical view of a superfamilies genome distribution. Searching for the genome distribution from the front page with a superfamily generates (Figure 4) a tree showing a distribution of the superfamily among Ensembl genomes. Different measures of commonness of a superfamily can be selected from the genome distribution search on the front page of the website. Alternatively as with many of the searches, RESTful alternatives exist, hence http://gene3d.biochem.ucl.ac.uk/superfamily/2.60.40.790/genome-distribution/size/number displays the number of genes with a given superfamily across genomes, while http://gene3d.biochem.ucl.ac.uk/superfamily/2.60.40.790/genome-distribution/size/rate displays the proportion of genes with this superfamily. Cytoscape Web functionality facilitates zooming into the region of interest (Figure 4B). Clicking on a node (Figure 4A) (species or species group) gives a more detailed breakdown of the domain counts and allows the retrieval of genes in the case of species.

Data downloads

Since the last Gene3D update a suite of web services have recently been developed for structural/ functional annotation. The services allow flexible downloading of assignments and associated annotations from Gene3D, and access to some of the computational tools used by Gene3D internally. The services are RESTful, meaning they can be easily accessed from UNIX command line, code or even a browser. For more information, see the recent NAR web services publication (8).

DISCUSSION

Gene3D is an evolving resource adapting to the rapidly emerging fields of molecular biology. In this update, we have developed tools such as the network analysis tab, allowing for protein interaction networks to be analysed within the context of their CATH domains. As new data sets become available, we will continue to integrate them into the database. Among other things, the novel FunFam assignments have been added to provide functional subdivisions of the Gene3D superfamilies and future releases will exploit this resource further. As the FunFam developers expand the tools available, we will continue to improve this aspect of Gene3D. The uniqueness of the Gene3D resource comes from its genome-wide CATH structural domain assignments. This update expands on this uniqueness, by providing more powerful analysis tools and integrating additional, complementary data sources.

Downloaded from https://academic.oup.com/nar/article/40/D1/D465/2903340 by guest on 25 September 2024

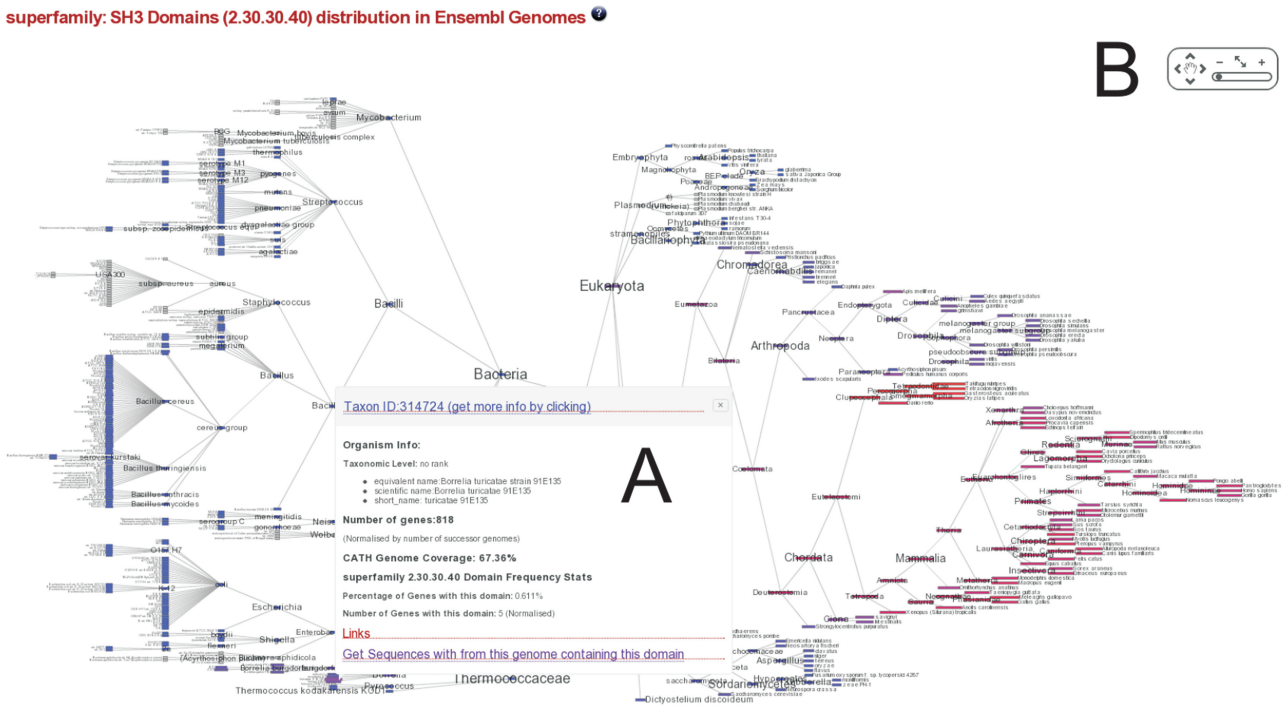

Figure 4. Genome distribution of CATH superfamily SH3-Domains across Ensembl genomes. Wider and redder nodes correspond to higher relative frequencies of the superfamily. Node tooltips (**A**) provide detailed information and links for gene retrieval. Zooming/panning functionality is also available (**B**).

FUNDING

IMI, EU (to J.L.); Wellcome Trust (to J.P. I.S.); EU Impact (to C.Y.); BBSRC (to R.R.); National Institutes of Health (to B.H.D.). Funding for open access charge: The Wellcome Trust.

Conflict of interest statement. None declared.

REFERENCES

1. Lees,J., Yeats,C., Redfern,O., Clegg,A. and Orengo,C. (2010) Gene3D: merging structure and function for a Thousand genomes. *Nucleic Acids Res.*, **38**, D296–D300.
2. Kersey,P.J., Lawson,D., Birney,E., Derwent,P.S., Haimel,M., Herrero,J., Keenan,S., Kerhornou,A., Koscielny,G., Kahari,A. *et al.* (2010) Ensembl Genomes: extending Ensembl across the taxonomic space. *Nucleic Acids Res.*, **38**, D563–D569.
3. Boutet,E., Lieberherr,D., Tognolli,M., Schneider,M. and Bairoch,A. (2007) UniProtKB/Swiss-Prot. *Methods Mol. Biol.*, **406**, 89–112.
4. Pruitt,K.D., Tatusova,T., Klimke,W. and Maglott,D.R. (2009) NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.*, **37**, D32–D36.
5. Cuff,A.L., Sillitoe,I., Lewis,T., Clegg,A.B., Rentzsch,R., Furnham,N., Pellegrini-Calace,M., Jones,D., Thornton,J. and Orengo,C.A. (2011) Extending CATH: increasing coverage of the protein structure universe and linking structure with function. *Nucleic Acids Res.*, **39**, D420–D426.
6. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
7. Yeats,C., Lees,J., Carter,P., Sillitoe,I. and Orengo,C. (2011) The Gene3D Web Services: a platform for identifying, annotating and comparing structural domains in protein sequences. *Nucleic Acids Res.*, **39**, W546–W550.
8. Hunter,S., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Das,U., Daugherty,L., Duquenne,L. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.
9. Redfern,O.C., Dessailly,B.H., Dallman,T.J., Sillitoe,I. and Orengo,C.A. (2009) FLORA: a novel method to predict protein function from structure in diverse superfamilies. *PLoS Comput. Biol.*, **5**, e1000485.
10. Izarzugaza,J.M., Baresic,A., McMillan,L.E., Yeats,C., Clegg,A.B., Orengo,C.A., Martin,A.C. and Valencia,A. (2009) An integrated approach to the interpretation of single amino acid polymorphisms within the framework of CATH and Gene3D. *BMC Bioinformatics*, **10(Suppl. 8)**, S5.
11. Luo,Q., Pagel,P., Vilne,B. and Frishman,D. (2011) DIMA 3.0: Domain Interaction Map. *Nucleic Acids Res.*, **39**, D724–D729.
12. Bjorkholm,P. and Sonnhammer,E.L. (2009) Comparative analysis and unification of domain-domain interaction networks. *Bioinformatics*, **25**, 3020–3025.
13. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
14. Sayers,E.W., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **37**, D5–D15.
15. Knox,C., Law,V., Jewison,T., Liu,P., Ly,S., Frolkis,A., Pon,A., Banco,K., Mak,C., Neveu,V. *et al.* (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.*, **39**, D1035–D1041.
16. Lee,D.A., Rentzsch,R. and Orengo,C. (2010) GeMMA: functional subfamily classification within superfamilies of predicted protein structural domains. *Nucleic Acids Res.*, **38**, 720–737.
17. Cuff,A., Redfern,O.C., Greene,L., Sillitoe,I., Lewis,T., Dibley,M., Reid,A., Pearl,F., Dallman,T., Todd,A. *et al.* (2009) The CATH hierarchy revisited-structural divergence in domain superfamilies and the continuity of fold space. *Structure*, **17**, 1051–1062.
18. Eddy,S.R. (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform*, **23**, 205–211.
19. Redfern,O.C., Harrison,A., Dallman,T., Pearl,F.M. and Orengo,C.A. (2007) CATHEDRAL: a fast and effective algorithm

Downloaded from https://academic.oup.com/nar/article/40/D1/D465/2903340 by guest on 25 September 2024

to predict folds and domain boundaries from multidomain protein structures. *PLoS Comput. Biol.*, **3**, e232.
20. Breitkreutz,B.J., Stark,C., Reguly,T., Boucher,L., Breitkreutz,A., Livstone,M., Oughtred,R., Lackner,D.H., Bahler,J., Wood,V. *et al.* (2008) The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res.*, **36**, D637–D640.
21. Matthews,L., Gopinath,G., Gillespie,M., Caudy,M., Croft,D., de Bono,B., Garapati,P., Hemish,J., Hermjakob,H., Jassal,B. *et al.* (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.*, **37**, D619–D622.
22. Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.
23. Aranda,B., Achuthan,P., Alam-Faruque,Y., Armean,I., Bridge,A., Derow,C., Feuermann,M., Ghanbarian,A.T., Kerrien,S., Khadake,J. *et al.* (2010) The IntAct molecular interaction database in 2010. *Nucleic Acids Res.*, **38**, D525–D531.
24. Chatr-aryamontri,A., Ceol,A., Palazzi,L.M., Nardelli,G., Schneider,M.V., Castagnoli,L. and Cesareni,G. (2007) MINT: the Molecular INTeraction database. *Nucleic Acids Res.*, **35**, D572–D574.
25. Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A. *et al.* (2009) Human Protein Reference Database–2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
26. Gilsdorf,M., Horn,T., Arziman,Z., Pelz,O., Kiner,E. and Boutros,M. (2010) GenomeRNAi: a database for cell-based RNAi phenotypes. 2009 update. *Nucleic Acids Res.*, **38**, D448–D452.
27. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
28. Gould,C.M., Diella,F., Via,A., Puntervoll,P., Gemund,C., Chabanis-Davidson,S., Michael,S., Sayadi,A., Bryne,J.C., Chica,C. *et al.* (2010) ELM: the status of the 2010 eukaryotic linear motif resource. *Nucleic Acids Res.*, **38**, D167–D180.
29. Mora-Bermudez,F., Gerlich,D. and Ellenberg,J. (2007) Maximal chromosome compaction occurs by axial shortening in anaphase and depends on Aurora kinase. *Nat. Cell. Biol.*, **9**, 822–831.
30. Szklarczyk,D., Franceschini,A., Kuhn,M., Simonovic,M., Roth,A., Minguez,P., Doerks,T., Stark,M., Muller,J., Bork,P. *et al.* (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.*, **39**, D561–D568.
31. McDowall,M.D., Scott,M.S. and Barton,G.J. (2009) PIPs: human protein-protein interaction prediction database. *Nucleic Acids Res.*, **37**, D651–D656.
32. Lopes,C.T., Franz,M., Kazi,F., Donaldson,S.L., Morris,Q. and Bader,G.D. (2010) Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, **26**, 2347–2348.
33. Wen,K.W. and Damania,B. (2010) Kaposi sarcoma-associated herpes virus (KSHV): molecular biology and oncogenesis. *Cancer Lett.*, **289**, 140–150.
34. Jenner,R.G., Maillard,K., Cattini,N., Weiss,R.A., Boshoff,C., Wooster,R. and Kellam,P. (2003) Kaposi's sarcoma-associated herpesvirus-infected primary effusion lymphoma has a plasma cell gene expression profile. *Proc. Natl Acad. Sci. USA*, **100**, 10399–10404.
35. Wilson,D., Pethica,R., Zhou,Y., Talbot,C., Vogel,C., Madera,M., Chothia,C. and Gough,J. (2009) SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res.*, **37**, D380–D386.
36. Ranea,J.A., Yeats,C., Grant,A. and Orengo,C.A. (2007) Predicting protein function with hierarchical phylogenetic profiles: the Gene3D Phylo-Tuner method applied to eukaryotic genomes. *PLoS Comput. Biol.*, **3**, e237.

Downloaded from https://academic.oup.com/nar/article/40/D1/D465/2903340 by guest on 25 September 2024