The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection

Michael Y. Galperin¹* and Xosé M. Fernández-Suárez²*

¹National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health (NIH), Bethesda, MD 20894, USA and ²EMBL–European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received November 16, 2011; Accepted November 17, 2011

ABSTRACT

The 19th annual Database Issue of Nucleic Acids Research features descriptions of 92 new online databases covering various areas of molecular biology and 100 papers describing recent updates to the databases previously described in NAR and other journals. The highlights of this issue include, among others, a description of neXtProt, a knowledgebase on human proteins; a detailed explanation of the principles behind the NCBI Taxonomy Database; NCBI and EBI papers on the recently launched BioSample databases that store sample information for a variety of database resources; descriptions of the recent developments in the Gene Ontology and UniProt Gene Ontology Annotation projects; updates on Pfam, SMART and InterPro domain databases; update papers on KEGG and TAIR, two universally acclaimed databases that face an uncertain future; and a separate section with 10 wiki-based databases, introduced in an accompanying editorial. The NAR online Molecular Biology Database Collection, available at http://www.oxfordjournals.org/nar/database/a/, has been updated and now lists 1380 databases. Brief machine-readable descriptions of the databases featured in this issue, according to the BioDBcore standards, will be provided at the http://biosharing.org/biodbcore web site. The full content of the Database Issue is freely available online on the Nucleic Acids Research web site (http://nar.oxfordjournals.org/).

COMMENTARY

This current, 19th annual Database Issue of Nucleic Acids Research (NAR) features descriptions of 92 new online databases covering a variety of molecular biology data, 77 update papers on databases that have been previously described in the NAR Database Issue and 23 papers with updates on database resources whose descriptions have previously been published in other journals (Table 1). The accompanying NAR online Molecular Biology Database Collection (http://www.oxfordjournals.org/nar/database/a/) has been revised, which resulted in updating the URLs of more than 30 databases and exclusion of more than 20 obsolete web sites. This list now includes 1380 databases sorted into 14 categories and 41 subcategories.

NEW AND UPDATED DATABASES

This issue contains an unusually high number of papers from the authors’ host institutions, NCBI and EMBL-EBI, respectively. In addition to the annual papers from the International Nucleotide Sequence Database collaboration [INSDC (1), which includes the DNA Data Bank of Japan, GenBank and the European Nucleotide Archive (2–4)], Ensembl (5), UniProtKB (6) and the Protein Data Bank in Europe (7), these include two papers that describe the BioSample database project, recently launched at both institutions. The BioSample databases [http://www.ncbi.nlm.nih.gov/biosample and http://www.ebi.ac.uk/biosamples/, (8) and (9), respectively] aim at capturing essential information about each biological sample used to obtain sequence, gene expression or protein expression data, as well as the relationship between different samples and their sources. The sample

*To whom correspondence should be addressed. Tel: +301 435 5910; Fax: +301 435 7793; Email: galperin@ncbi.nlm.nih.gov
Correspondence may also be addressed to Xosé M. Fernández-Suárez. Tel: +44 (0)1223 494 591; Fax: +44 (0)1223 494 468; Email: xose@ebi.ac.uk

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

Published by Oxford University Press 2011.

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.
### Table 1. New databases featured in the 2012 NAR Database issue

| Database name          | URL                                                                 | Brief description                                                                 |
|------------------------|----------------------------------------------------------------------|-----------------------------------------------------------------------------------|
| ApoHoloDB              | http://ahdb.ee.ncku.edu.tw/                                          | Apo- and Holo- structure pairs of proteins                                        |
| AutismKB               | http://autism.cbi.pku.edu.cn                                         | Autism genetics knowledgebase                                                      |
| BGMUT                  | http://www.ncbi.nlm.nih.gov/projects/gv/mhc/xs1.cgi?cmd=bgmut         | Blood Group antigen gene Mutation database                                         |
| BitterDB               | http://bitterdb.agri.hauji.ac.il/bitterdb/dbbitter.php                | Bitter taste: molecules and receptors                                              |
| canSAR                 | http://cansar.icr.ac.uk                                              | Integrated cancer research and drug discovery resource                            |
| CAPS-DB                | http://www.bioinsilico.org/CAPSDB                                    | Classification of helix cappings in protein structures                             |
| ccPDB                  | http://crdd.osdd.net/raghava/ccpdb/                                  | Compilation and creation of datasets from Protein Data Bank                         |
| CharProtDB             | http://www.jcvi.org/charprotdb/                                      | Experimentally Characterized Protein annotations                                   |
| COLT-Cancer            | http://colt.cbr.utoronto.ca/cancer                                    | Essential gene profiles in human cancer cell lines                                 |
| Crystallography Open Database | http://www.crystallography.net/                                 | Crystal structures of small molecules                                              |
| Cube-DB                | http://epsp.bmad.bii.a-star.edu.sg/cube/db/html/home.html            | Functional divergence in human protein families                                    |
| DARC                   | http://darcsite.genzentrum.lmu.de/darc/                              | Database for Aligned Ribosomal Complexes                                            |
| DBETH                  | http://www.hpppi.iicb.res.in/btx/                                    | Database for Bacterial ExoToxins for Humans                                        |
| Death Domain database  | http://www.deathdomain.org                                            | Protein interaction data for Death Domain superfamily                              |
| DIGIT                  | http://www.biocomputing.it/digit4/                                   | Database of ImmunoGlobulin sequences and Integrated Tools                          |
| Disease Ontology       | http://diseaseontology.sf.net/                                      | Ontology for a variety of human diseases                                           |
| DiseaseMeth            | http://201.97.205.78/diseasemeth/                                    | Human disease methylation database                                                 |
| DistLD                 | http://distild.jensenlab.org/                                       | Diseases and Traits In Linkage Disequilibrium blocks                               |
| DNAtraffic             | http://dnatraffic.ibb.waw.pl/                                        | DNA dynamics during the cell cycle                                                 |
| DOMMINO                | http://dommino.org                                                   | Database of MacroMolecular INteractions                                             |
| doRNA                  | http://dorina.mdc-berlin.de/                                         | Database of RNA interactions in post-transcriptional regulation                    |
| DR. VIS                | http://www.scbit.org/dbmi/drvis                                      | Human Disease-Related Viral Integration Sites                                      |
| EBI BioSample Database | http://www.ebi.ac.uk/biosamples/                                     | Biological samples used as sources of sequence, structure or expression data       |
| EcoliWiki              | http://ecoliwiki.net                                                 | Community-based pages about non-pathogenic *E. coli*                               |
| eQuilibrator           | http://equilibrator.weizmann.ac.il                                   | Thermodynamics calculator for biochemical reactions                                 |
| FungiDB                | http://fungidb.org                                                   | Functional genomics of fungi                                                       |
| FunTree                | http://www.ebi.ac.uk/thornton-srv/databases/FunTree/                 | Evolution of novel enzyme functions in enzyme superfamilies                         |
| GeneWeaver             | http://www.GeneWeaver.org                                            | Functional genomics analysis system                                               |
| GONUTS                 | http://gowiki.tamu.edu                                               | Gene Ontology Normal Usage Tracking System                                          |
| GWASdb                 | http://jijwănglab.org/gwasdb                                        | Human genetic variants identified by genome wide association studies               |
| HaploReg               | http://combio.bio.mit.edu/HaploReg                                  | SNP-centric access to chromatin state information                                  |
| HFV database           | http://hfv.lanl.gov                                                   | Hemorrhagic fever virus sequence database                                           |
| hiPathDB               | http://hipathdb.kobic.re.kr/                                         | Human Integrated Pathway Database                                                   |
| Histome                | http://www.histome.net                                               | Human histone database                                                             |
| HotRegion              | http://prism.ccb.cb.ku.edu.tr/hotregion                              | Database of interaction Hotspots                                                    |
| Human OligoGenome Resource | http://oligogenome.stanford.edu/                                  | Oligonucleotides for targeted resequencing of the human genome                    |
| ICEberg                | http://db-mml.sjtu.edu.cn/ICEBerg/                                   | Integrative and Conjugal Elements in Bacteria                                      |
| IDEAL                  | http://www.ideal.force.cs.is.nagoya-u.ac.jp/IDEAL/                  | Intrinsically Disordered proteins with Extensive Annotations and Literature         |
| IGDB.NSCLC             | http://igdb.nsclc.ibm.snicu.edu.tw                                   | Integrated Genomic Database of Non-Small Cell Lung Cancer                           |
| IndelFR                | http://indel.bioinfo.sdu.edu.cn                                      | Indel Flanking Region database                                                     |
| InterEvol              | http://biodev.cea.fr/interevol                                       | Evolution of protein–protein Interfaces                                             |
| LegumeITP              | http://plantgrn.noble.org/legumeitp/                                | Model Legumes Integrative database Platform                                        |
| MetaBase               | http://metadbase.org                                                | Wiki database of biological databases                                             |
| MethyloMeDB            | http://epigenomics.columbia.edu/methylomedb/                         | DNA methylation profiles in human and mouse brain                                 |
| MINAS                  | http://www.minas.uzh.ch                                             | Metal Ions in Nucleic AcidS                                                         |
| MIPModDB               | http://bioinfo.itk.ac.in/MIPModDB                                   | Major Intrinsic Protein superfamily Models                                          |
| miREX                  | http://bioinfo.amu.edu.pl/mirex                                      | Plant microRNA Expression data                                                     |
| miRNOST                | http://mirnest.amu.edu.pl                                           | microRNAs in animal and plant EST sequences                                       |
| MMMDDB                 | http://mmdb.iab.keio.ac.jp/                                         | Mouse Multiple Tissue Metabolomics Database                                         |
| modMine                | http://intermine.modencode.org                                      | Mining of modENCODE data                                                           |
| MOPED                  | http://moped.proteinspire.org                                       | Model Organism Protein Expression Database                                         |
| NCBI BioSample         | http://www.ncbi.nlm.nih.gov/biosample                                | Biological samples used as sources of sequence, structure or expression data        |
| NCBI BioProject        | http://www.ncbi.nlm.nih.gov/bioproject                              | Linked data related to a single research project                                   |
| Nematodes.org          | http://www.nematodes.org                                           | Wiki for coordinating nematode sequencing projects                                 |
| Newt-omics             | http://newt-omics.mpi-bn.mpg.de                                      | Data on red spotted new *Notophthalmus viridescens*                                |
| neXtProt               | http://www.nextprot.org                                             | A knowledgebase for human proteins                                                 |

(continued)
information includes the name of the source organism (or an environmental isolate), the source material within that species such as e.g. the organ, tissue and the cell type. It will also contain information about the isolation source of the sample, (some or all of) locality, host, collection date, etc. For human sources, BioSample information will include any available—and ethically appropriate—additional data, such as the disease state and clinical information [clinical samples that may raise privacy concerns will continue to be kept at the NCBI's dbGaP database (10) and the EBI's European Genome-phenome Archive (http://www.ebi.ac.uk/ega/), with sanitized versions available in the BioSample databases]. While providing sample information will place additional burden on the submitters, the availability of BioSample data should dramatically improve the experience of a typical user. By consistently recording sample information for various kinds of data stored in the NCBI and EBI databases, the BioSample databases will allow smooth cross-database searching of all available information pertaining to a particular sample source, such as cell type, disease, or a tissue biopsy. Furthermore, since NCBI and EBI agreed to assign shared sample accession numbers, these numbers could now be used to query web sites of both institutions (8,9).

Five papers in this issue describe databases resources of the US Department of Energy’s Joint Genome Institute (JGI, http://www.jgi.doe.gov). These include a description of the JGI Genome Portal (11) with its fungal (MycoCosm), plant (Phytozome), prokaryotic (IMG)
and metagenomic (IMG/M) resources, and the Genomes OnLine Database (GOLD, http://www.genomesonline.org), which lists the ongoing genomic and metagenomic projects (12).

One of the major highlights of this issue is the first description of neXtProt, a knowledgebase on human proteins that has been created at the Swiss Institute of Bioinformatics (SIB) on the basis of the human protein set in the UniProtKB/Swiss-Prot and then expanded by including quality-assessed protein expression, localization, variation and proteomics data (13). Other highlights include CharProtDB, a database of experimentally characterized proteins that is used for genome annotation at the J. Craig Venter Institute (14); a detailed explanation of the basic principles behind the NCBI Taxonomy Database and the ways it ties together various DNA and protein sequence and gene expression data for all organisms and taxonomic groups represented in GenBank (15); the descriptions of the recent developments in the Gene Ontology and UniProt Gene Ontology Annotation projects (16,17), and updates on model organism databases SGD, MGD, FlyBase and WormBase (18–21) and on Pfam, SMART and InterPro domain databases (22–24).

With all the diversity of the databases featured in this issue, the major trend appears to be an increased focus on small molecules (ChEMBL, PubChem, BiterDB, SCRIPDB, Crystallography Open Database) and related topics, such as properties of enzyme-catalyzed reactions (Rhea, MACIE, eQuilibrator, SABIO-RK), protein–ligand binding (Pocketome, PoSSuM, ProtChemSI, STITCH), and the analysis of potential drugs and drug targets for human disease (canSAR, DAMPD, DBETH, SuperTarget, TDR Targets, Therapeutic Target Database). As in previous years, there is a strong representation of structure databases, including descriptions of the European and Japanese Protein Data Banks (PDBe, PDBj), two databases of refined NMR structures (NRG-CING and STAP Refinement of NMR database), and several other databases on protein structure and protein–protein interactions.

An unusually high number of databases, including ChEMBL, FunCoup, MitoMiner, PhosphoSitePlus, Pocketome, SABIO-RK and TDR Targets, are featured in this NAR Database Issue for the first time after having their descriptions published elsewhere (Table 2). All these databases have been available online for several years and have been accepted and valued by the community. Accordingly, they presented few, if any, problems with the database design, although some appeared somewhat less user-friendly than is required for the NAR Database Issue. We consider publication of these papers in the NAR Database Issue a continuation of our efforts to bring the readers the best publicly available molecular biology databases, as well as a reflection of the unique status of this publication that introduces the databases to a very wide audience.

In response to the growing popularity of Wikipedia (http://www.wikipedia.org) and wiki-based approaches to constructing and curating biological databases, this issue includes a special section with 10 papers describing various wiki-based databases. These papers are introduced in an accompanying editorial by Rob Finn, Paul Gardner and Alex Bateman (25), whose very popular Pfam (22) and Rfam (26) databases successfully incorporate wiki elements. It could be argued that the Pfam update paper (22) should have been placed in that section as well.
SUSTAINABILITY OF BIOINFORMATICS DATABASES

A joint paper in this issue from the three INSDC members (27) discusses the progress of the Sequence Read Archive (SRA, previously known as the Short Read Archive), however, without mentioning the controversy that surrounded the SRA in the past year. Established in 2007 as a public repository of raw sequence data from next-generation sequencing platforms, SRA stores sequence data generated for RNA-Seq, ChIP-Seq and genotyping studies, as well as from several large-scale projects, such as the Human Microbiome project (https://commonfund.nih.gov/hmp) and the 1000 Genomes project (http://www.1000genomes.org) (27). In June 2011, its volume surpassed 100 Terabases (10^{12} bases) of DNA. In February, NCBI announced that, due to budget constraints, it would discontinue the SRA within the next 12 months (http://www.ncbi.nlm.nih.gov/About/news/16feb2011). This announcement caused a widespread response (28). One news source even claimed that NCBI ‘announced that it would slowly phase out its DNA archive due to federal budget cuts’. There has been also an extensive online discussion on the http://seqanswers.com wiki web site (which is described in a separate paper in this issue). However, the news of the SRA demise proved largely premature. Within days, EBI and DDBJ announced that they would continue supporting the SRA (http://www.ebi.ac.uk/ena/SRA_announcement_Feb_2011.pdf, http://www.ddbj.nig.ac.jp/whatsnew/2011/DRA20110222.html), and the NIH provided support to enable the continuation of the SRA (http://www.ncbi.nlm.nih.gov/About/news/13Oct2011.html). Still, given that the SRA keeps growing at a rapid pace and handling the data becomes increasingly complicated, the INSDC paper carefully states that ‘SRA partners actively discuss and pursue approaches together with user communities to maximize the benefit gained from archiving next-generation sequencing data while minimizing the infrastructure costs’ (27).

Despite its successful resolution, the SRA story highlights an important problem of whether public database providers should try keeping all sequence-related data or make certain choices about the kind of resources that they would like to maintain. The same news release in February 2011 announced the closure of Peptidome, the NCBI resource for tandem mass spectrometry peptide and protein identification data (29). The closure of Peptidome attracted far less attention than of SRA, probably because of the continued operation of EBI’s PRIDE (30), Seattle Proteome Center’s PeptideAtlas (31), the recently created MOPED (32) and other proteomics resources. Still, it is definitely a sign of things to come, as is the recently announced closure of the International Protein Index, which is to be replaced by the complete proteome sets in UniProtKB (33).

Most importantly, the worldwide attention to the SRA story illuminates the deep concern that exists in the community with regard to the stability (viability) of the online databases that have become key resources enabling all kinds of biomedical research. Previously, we have seen a natural selection of databases that led to a relatively orderly succession: as some databases have grown obsolete, they were replaced by similar but more robust databases maintained elsewhere. For example, after termination of IRESDb, a database of the internal ribosome entry sites (34), the same data were still available through the IRESite database (35). Among the databases featured in this issue, MitoZoa provides the same coverage of metazoan mitochondrial genomes as the now-defunct AMmtDB, Gene3D fully replaces the no-longer-maintained 3D-Genomics, and Ensembl (5) provides the alternative splicing data that have previously been available through ASHE5db, EBI’s ASD/ATD/ATSD and several other recently discontinued databases.

Unfortunately, owing to the difficult economic times, budget constraints are now leading to the termination (or commercialization) of truly unique resources, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.genome.jp/kegg) and The Arabidopsis Information Resource (TAIR, http://arabidopsis.org), both featured in this issue (36,37). The KEGG database, maintained by Minoru Kanehisa and his colleagues at the Bioinformatics Center of the Kyoto University Institute for Chemical Research, has been a permanent feature of the NAR Database Issue since 1997 and is now in its 60th release (36), see http://www.genome.jp/en/release.html. However, after Kanehisa, who was one of the founders of GenBank and has been at the forefront of bioinformatics research ever since, has reached the mandatory retirement age; the future of KEGG has suddenly become uncertain (see http://www.genome.jp/kegg/docs/plea.html). Right now, KEGG continues to be publicly available but its funding mechanisms support a narrow focus on translational research (36), which is certainly important but is only a minor part of the enormous contribution of this database to the progress of genomics and bioinformatics around the world.

The case of TAIR is even more troubling. Over the past 12 years, TAIR enjoyed generous support from the US National Science Foundation (NSF, http://www.nsf.gov) that helped it grow into a recognized source of sequence data and curated annotation of the model plant Arabidopsis thaliana. Three previous publications on TAIR in the NAR Database Issue in 2001, 2003 and 2008 were all extremely well cited, confirming the widespread use of this resource. With the completion of the Arabidopsis sequencing project, the focus of TAIR shifted from providing new annotation to improving the existing genome annotation, making it the ultimate source of gene annotation and expression data for A. thaliana. Unfortunately, this new focus failed to win the NSF support and the funding for a project that until recently has been heralded as one of the NSF best success stories will end in August of 2013. This will likely mean termination of TAIR as we know it; the existing plans for corporate sponsorship of TAIR and/or for its shift to an International Arabidopsis Informatics Consortium (see http://www.arabidopsis.org/doc/about/tair_funding/410) are not going to prevent the demise of this useful genomic resource.
These recent developments show that the importance of the public database resources, which is obvious to any biologist, needs to be constantly highlighted to the national and international financing bodies. We all remember the financial difficulties encountered in the 1990s by the Swiss-Prot database after it failed to secure sufficient support from the European Union (http://web.expasy.org/docs/crisis96/help-sprot.html) (38). Fortunately, in the end, Swiss government recognized the value of that unique resource and provided funding to support Swiss-Prot (39). It now supports the UniProtKB/Swiss-Prot activities at the SIB, whereas funding for the UniProtKB activities at the EBI and PIR is provided by the NIH, NSF and the European Commission (6).

The stories of Swiss-Prot, KEGG and TAIR also illustrate the need clearly articulated in a recent paper by Julian Parkhill, Ewan Birney and Paul Kersey, (40) for a comprehensive infrastructure that would (i) support the key bioinformatics resources, (ii) extend to the model organism databases and (iii) bring the genomic information into every biological lab. In the USA, such infrastructure includes the NCBI, the JGI and associated DOE labs, the NIH-funded Bioinformatics Resource Centers (this issue includes papers on VectorBase and ViPR, as well as on EuPathDB-associated databases, such as GeneDB, FungiDB, and TDR Targets) and comprehensive resources on model organisms, such as FlyBase, WormBase, SGD and MGD (18–21). In Europe, coordination of the bioinformatics infrastructure is planned through the EU-sponsored ELIXIR (European Life Sciences Infrastructure for Biological Information, http://www.elixir-europe.org) project, which aims at guaranteeing seamless access to biological information by integrating data generators and data centers throughout Europe.

AN ECOSYSTEM OF DATABASES

Although this issue looks like a simple catalog, it is important to note that we are not dealing with isolated resources: many listed databases interact in a variety of ways, forming a network of interconnected (or at least hyperlinked) data resources. Obviously, UniProtKB provides a plethora of links to all kinds of databases, including ENA, GenBank, DDBJ, RefSeq, PDBe, PDBj, IntAct, MINT, Ensembl, KEGG, UCSC Genome Browser, neXtProt, SGD, FlyBase, WormBase, MGD, TAIR, eggNOG, MetaCyc, InterPro, Gene3D, Pfam, SMART and ProtoNet, which are featured in this issue. However, many database interactions are more subtle: for example, BioMart has been recently used to link protein annotation data from the Reactome database of metabolic networks (41) to phosphoproteomics data in PRIDE (30) and somatic mutations in COSMIC (42), which allowed putting cancer-related mutation data into a functional context (43).

We believe that establishing connections between databases is an important way of improving the databases themselves, providing the user with additional search tools and, more generally, creating a live ecosystem that stores and expands knowledge. Accordingly, we consider it essential that the databases featured in the NAR Database Issue do their best in creating links to outside resources and providing an easy and straightforward way for the authors of other databases to link to their database content.

Last year, we published a paper by the BioDBcore Working Group that proposed creating a resource of ‘minimal information about a biological database’, a community-defined, uniform, generic description of the core attributes of biological databases (44). Accordingly, submitters to this year’s NAR Database Issue were asked to fill out a checklist of core attributes (available at http://www.biodbcore.org) of their databases and provide it as supplementary material to their manuscripts. Most of the authors complied with this request, which resulted in a stand-alone resource that contains machine-readable descriptions of the databases featured in this issue and is available from the BioSharing website (http://biosharing.org/biodbcore). We hope that this effort would illuminate the scope and general features of every listed database resource, including the community standards that these systems support, forge better contacts between their authors, simplify linking various data sets, and, eventually, bring greater clarity and integration to the whole field of molecular biology databases.

ACKNOWLEDGEMENTS

The authors thank Sir Richard Roberts and Drs Alex Bateman, David Landsman, Ilene Mizrachi and David Roos for helpful comments; Drs Philippe Rocca-Serra, Susanna-Assunta Sansone (University of Oxford) and Pascale Gaudet (SIB) for processing the BioDBcore submissions; Dr Martine Bernardes-Silva, Patricia Anderson and Ingrid Sjolund for excellent editorial assistance; Sheila Plaister for help with the NAR online Database Collection, and the Oxford University Press team led by Jennifer Boyd, Michael Evans, Andrew Malvern and Kate Puttick for their help in compiling this issue.

FUNDING

Intramural Research Program of the US National Institutes of Health at the National Library of Medicine (to M.Y.G.); European Molecular Biology Laboratory (to X.M.F.S.). Funding for open access charge: waived by Oxford University Press.

Conflict of interest statement. The authors’ opinions do not necessarily reflect the views of their respective institutions.

REFERENCES

1. Karseh-Mizrachi, I., Nakamura, Y., Cochrane, G. and on behalf of the International Nucleotide Sequence Database Collaboration (2012) The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res., 40, D33–D37.
Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, 39, D691–D697.

42. Forbes,S.A., Bindal,N., Bamford,S., Cole,C., Kok,C.Y., Beare,D., Jia,M., Shepherd,R., Leung,K., Menzies,A. *et al.* (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.*, 39, D945–D950.

43. Ndegwa,N., Cote,R.G., Ovelleiro,D., D’Eustachio,P., Hermjakob,H., Vizcaíno,J.A. and Croft,D. (2011) Critical amino acid residues in proteins: a BioMart integration of reactome protein annotations with PRIDE mass spectrometry data and COSMIC somatic mutations. *Database*, 2011, bar047.

44. Gaudet,P., Bairoch,A., Field,D., Sansone,S.A., Taylor,C., Attwood,T.K., Bateman,A., Blake,J.A., Bult,C.J., Cherry,J.M. *et al.* (2011) Towards BioDBcore: a community-defined information specification for biological databases. *Nucleic Acids Res.*, 39, D7–D10.