Published online 12 June 2012 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* ***W609–W614***
doi:10.1093/nar/gks575

CellBase, a comprehensive collection of RESTful web services for retrieving relevant biological information from heterogeneous sources

Marta Bleda[1,2], Joaquin Tarraga[1,3], Alejandro de Maria[1], Francisco Salavert[1,2], Luz Garcia-Alonso[1], Matilde Celma[4], Ainoha Martin[4], Joaquin Dopazo[1,2,3,*] and Ignacio Medina[1,3,*]

[1]Department of Bioinformatics and Genomics, Centro de Investigación Príncipe Felipe (CIPF), 46012 Valencia, Spain, [2]CIBER de Enfermedades Raras (CIBERER), 46010 Valencia, Spain, [3]Functional Genomics Node (INB) at CIPF, 46012 Valencia, Spain and [4]Research Center on Software Production Methods (ProS). DSIC Universitat Politècnica de València (UPV), 46007 Valencia, Spain

Received March 23, 2012; Revised May 18, 2012; Accepted May 21, 2012

ABSTRACT

During the past years, the advances in high-throughput technologies have produced an unprecedented growth in the number and size of repositories and databases storing relevant biological data. Today, there is more biological information than ever but, unfortunately, the current status of many of these repositories is far from being optimal. Some of the most common problems are that the information is spread out in many small databases; frequently there are different standards among repositories and some databases are no longer supported or they contain too specific and unconnected information. In addition, data size is increasingly becoming an obstacle when accessing or storing biological data. All these issues make very difficult to extract and integrate information from different sources, to analyze experiments or to access and query this information in a programmatic way. CellBase provides a solution to the growing necessity of integration by easing the access to biological data. CellBase implements a set of RESTful web services that query a centralized database containing the most relevant biological data sources. The database is hosted in our servers and is regularly updated. CellBase documentation can be found at http://docs.bioinfo.cipf.es/projects/cellbase.

INTRODUCTION

During the past years, the increase in scientific knowledge and the massive data production have caused an exponential growth in the number and size of biological databases and repositories. However, data size, which can reach hundreds of gigabytes, involves serious problems of data access through Internet and data storage in local disks.

Other challenging issues associated to biological data are that much relevant information is spread out in different databases or repositories, different identifiers or standards are used and data can be very frequently updated as new experiments are conducted. This is a particular problem when analyzing high-throughput experiments such as expression profiling, genotyping or massive sequencing data because much heterogeneous biological information is required for its interpretation.

To address these daily problems, we have developed a comprehensive infrastructure that comprises a relational database containing biological information and a web services application programming interface (API) to query all these data. This relational database integrates biological information from different sources and includes (i) core features such as genes, transcripts and exons or proteins; (ii) regulatory elements such as transcription factors (TFs) and TF binding sites, microRNA (miRNA) and curated and non-curated miRNA targets or CpG islands; (iii) many functional ontologies from Open Biomedical Ontologies (OBO) Foundry (1); (iv) variation data such as single-nucleotide polymorphisms (SNPs), phenotypic-related SNPs, known mutations or structural variation and (v) systems biology information such as pathways or protein interactome.

Web services API have been implemented in representational state transfer (REST) style that allows an easy, lightweight, fast and intuitive way of querying data in the database. Outputs can be obtained in plain tabulated text

*To whom correspondence should be addressed. Tel: +34 96 328 96 80; Fax: +34 96 328 97 01; Email: imedina@cipf.es
Correspondence may also be addressed to Joaquin Dopazo. Tel: +34 96 328 96 80 (ext 1007); Fax: +34 96 328 97 01; Email: jdopazo@cipf.es

or in JSON format. Database and RESTful web services have been designed and implemented to ensure high availability of the servers and to be fast, what results in real-time queries most of the time.

Our results provide a convenient solution to access and retrieve heterogeneous relevant biological information without the need of local databases installations. Data are always available by a high-availability cluster and queries have been tuned to ensure a real-time performance.

BIOLOGICAL SOURCES

CellBase database and web services have been designed to integrate and provide easy and efficient access to the most relevant biological information. This access is provided through a comprehensive and extensible RESTful web services API. Currently, some of the model organisms supported are human, mouse, rat, zebrafish, worm, fruit fly, pig, dog and yeast (and soon new organisms will be added).

CellBase integrates different data types from different sources into a relational database. These data comprise most relevant biological information taken from the main repositories. These data are organized in different sections depending on the type of information as described below.

Core features

We took genome sequences, genes, transcripts, exons, cytobands or cross references (xrefs) identifiers (IDs) from Ensembl (2). Protein information including sequences, xrefs or protein features (natural variants, mutagenesis sites, post-translational modifications, etc.) were imported from UniProt (3).

Regulatory

CellBase imports miRNA from miRBase (4); curated and non-curated miRNA targets from miRecords (5), miRTarBase (6), TargetScan (7) and microRNA.org (8) and CpG islands and conserved regions from the UCSC database (9).

Functional annotation

OBO Foundry (1) develops many biomedical ontologies that are implemented in OBO format. We designed a SQL schema to store these OBO ontologies and >30 ontologies were imported. OBO ontology term annotations were taken from Ensembl (2). InterPro (10) annotations were also imported.

Variation

CellBase includes SNPs from dbSNP (11); SNP population frequencies from HapMap (12), 1000 genomes project (13) and Ensembl (2); phenotypically annotated SNPs were imported from NHRI GWAS Catalog (14), HGMD (15), Open Access GWAS Database (16), UniProt (3) and OMIM (17); mutations from COSMIC (18) and structural variations from Ensembl (2).

Systems biology

We also import systems biology information like interactome information from IntAct (19). Reactome (20) stores pathway and interaction information in BioPAX (21) format. BioPAX data exchange format enables the integration of diverse pathway resources. We successfully solved the problem of storing data released in BioPAX format into a SQL relational schema, which allowed us importing Reactome in CellBase.

TECHNICAL DETAILS

Architecture design and implementation

CellBase database and web services architecture have been designed to be both very fast and fault-tolerant, thus providing a high-availability solution with no single point of failure. This is an important feature that makes CellBase very reliable and scalable if more servers are needed. Figure 1 shows a schema of the architecture. Some of the advantages of this architecture are that no installation of software tools or databases are needed by the user, as the whole API has been implemented using RESTful web services, and thus, database and web services API are always available.

Data from different biological databases and sources were integrated into a normalized relational database, implemented in a MySQL replication cluster to support a high load access as shown in Figure 1. In total, >200 GB were stored in the database. To speed-up queries, indexes and summary tables were created, resulting in runtimes of a few milliseconds for most of the queries. In order to provide a high availability and load balancing to the MySQL replication cluster, a Keepalived (http://www.keepalived.org) and HAProxy (http://haproxy.1wt.eu) services have been configured.

Web services API have been designed and implemented using REST software architectural style. RESTful web services have some advantages over SOAP web services such as that they tend to be more lightweight, scalable and easy to build and consume, thus providing a fast access to data even in low bandwidth conditions. To achieve a higher performance, Java was chosen for the server implementation to both (i) connect to MySQL using JBoss Hibernate (http://www.hibernate.org) and (ii) to implement RESTful web services API using Jersey library (http://jersey.java.net). Apache Tomcat was chosen as Java application server to deploy the web archive (*war* file) with the RESTful web services API implementation. To provide a high availability and load balancing to these web services, a HAProxy (http://haproxy.1wt.eu) was set up to balance Apache Tomcat instances. RESTful web services are registered in BioCatalogue (22) and access is free to all users.

Servers

CellBase MySQL replication cluster database is running in two high-end servers with two Intel Xeon Hexa-Core CPUs each, 96GB of memory and 6 SSD disks configured as a RAID-5 volume giving a 1,2TB of extremely fast

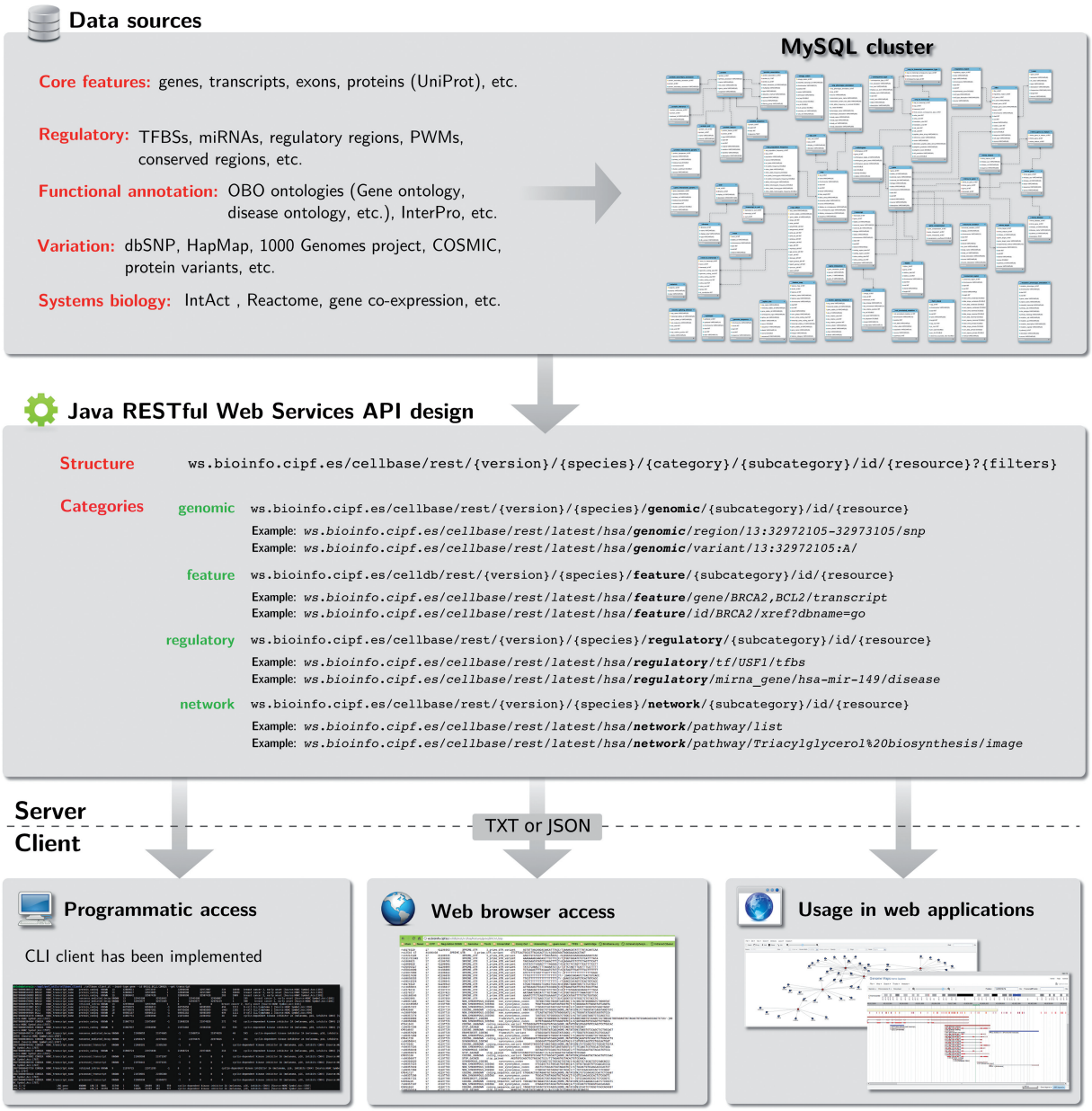

Figure 1. Schema of CellBase architecture of RESTful web services.

access to disk. CellBase Apache Tomcat instances are running in a cluster with three high-end servers with an Intel Xeon Quad-Core and 6GB of memory each.

DATA ISSUES IN BIOLOGY

Biology is experiencing an unprecedented growth of data and resources, what is making very difficult the access to local or conventional databases. New 'cloud computing' paradigm allows storing big data in remote servers and using web services to access and retrieve data efficiently. By doing so, researchers do not need to download, parse or integrate different sources since data are always up-to-date and can be retrieved by different client applications. Here, we follow a philosophy similar to some 'cloud'-based applications or companies in which heavy data are not moved through the web and is always up-to-date available by RESTful web services API.

WEB SERVICES

The CellBase RESTful web services API is an intuitive collection of methods that allow the user searching and retrieving biological data in a user-friendly way. RESTful calls are easily accessed using Universal Resource Locators (URLs), reducing the access to biological information to a simple browser query. Output results can be retrieved in text or JSON formats. Today, all programming languages can handle URLs, what makes CellBase biological information fully accessible in

a programmatic way. In this section, we describe URLs syntax and provide some usage examples and details on how to access data.

Understanding the CellBase web services API structure

Biological information can be accessed easily using URLs. These URLs have been designed to be flexible, neutral and scalable in future extensions, permitting the addition of new methods without altering the structure. The general syntax of CellBase URLs is as follows:

ws.bioinfo.cipf.es/cellbase/rest/version/species/category/subcategory/id/resource?filters

The first part of the URL, ws.bioinfo.cipf.es/cellbase/rest, refers to the host and is fixed for all methods. The remainder part will vary depending on the user's query. 'Versions' are numbered with the letter 'v' followed by a number (i.e. v1) or by the keyword 'latest' to access the current release. The 'species' field can be specified using the three-letter code (hsa, mmu, rno, etc.) or the abbreviated format (hsapiens, mmusculus, rnorvegicus, etc.). Currently, biological information is available for 11 species, as described above. 'Category' field aims to provide a general classification for the input identifier according to its nature. Four main categories are available:

- 'Genomic', which makes reference to genomic coordinates like regions, positions or variants.
- 'Feature', involves all elements that have a defined location on the genome and provides a comfortable way to retrieve cross references for an identifier.
- 'Regulatory', refers to all regulatory features, including interactions that involve TFs and microRNAs.
- 'Network', makes reference to different types of networks and pathways, including the protein interactome, the regulatory network and Reactome.

The 'subcategory' field must indicate the type of the input identifier (gene, transcript, region, pathway, etc.). Users can choose the most suitable one among the predefined subcategories. 'Id' is the query parameter, the feature or term about which the user wants to retrieve the information. Users can query more than one identifier using a comma separated list. The 'resource' field refers to the information the user wants to obtain from the id field. Depending on the category and the subcategory specified, different 'resources' are available. Resources must always be written in singular. Table 1 summarizes a representative selection of the available resources for each category and the corresponding subcategories. Identifier types and formats for each subcategory are also described. Users can add 'filters' to the RESTful query after the question mark. Some of them can be applied to all queries, but some of them are specific for each subcategory. An example of the options that can be applied to all queries are the output format, coded as 'of', and the character used to separate columns in the resulting output, coded as 'separator'.

Additional web services have been designed to provide metadata about CellBase and web services themselves like retrieving information about the available species, genome assemblies or species codes. In addition, column headers and usage about subcategories have been added to provide an online help to users.

Detailed and up-to-date information of currently available methods can be accessed by visiting the documentation page at http://docs.bioinfo.cipf.es/projects/cellbase/wiki.

Examples of CellBase RESTful web services queries

To provide a clarifying idea about what these web services are able to do, here we show some examples of usage.

Example 1. To obtain the SNPs of a particular region (i.e. chromosome 16 from 3698105 to 3701105), we use the genomic category since the specified inputs are genomic coordinates:

http://ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/genomic/region/16:3698105-3701105/snp

Example 2. Retrieving all transcripts for a gene is straightforward using the feature category:

http://ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/feature/gene/BRCA2/transcript

Table 1. Summary of some of the main available categories and subcategories

Category	Subcategory	Identifier format	Resources
Genomic	Region	chr:start-end	gene, transcript, snp, sequence, reverse, tfbs, mirna_target, regulatory
	Variant	chr:position:new allele	consequence_type
	Position	chr:position	gene, snp, mutation, functional
Feature	Gene	All gene ID formats	info, sequence, transcript, tfbs, mirna_target, protein_feature, snp, mutation
	Transcript	Ensembl or RefSeq ID	info, gene, sequence, exon
	Snp	dbSNP or Ensembl ID	info, consequence_type, population_frequency, phenotype, xref
	Exon	Ensembl ID	info, sequence, region, transcript
	Protein	UniProt or Ensembl ID	info, gene, sequence, transcript, feature, xref, variant
	Id	All possible IDs	Xref
Regulatory	mirna_gene	miRBase gene ID	info, gene, mirna_mature, target, disease
	mirna_mature	miRBase mature ID	info, gene, mirna_gene, target, disease
	Tf	TF or gene name	info, tfbs, gene, protein, pwm
Network	Pathway	none	List
		Reactome pathway name	info, subpathway, element, gene, protein, image
	Interactome	UniProt or Ensembl ID	info, element, neighbourhood, adjacent, connected_component

Example 3. To convert a list of Ensembl gene identifiers to their HGNC gene symbol, CellBase implements a way to retrieve cross references for an identifier using the External references (xref) subcategory. We just need to specify the list of identifiers and the name of the database we want to convert these IDs to. The resulting query will look like this:
http://ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/feature/id/ENSG00000011478,ENSG00000008382/xref?dbname=hgnc_symbol
Example 4. To fetch the target genes for a particular miRNA like hsa-mir-150:
http://ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/regulatory/mirna_gene/hsa-mir-150/target
Example 5. To download a diagram of a specific pathway:
http://ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/network/pathway/Triacylglycerol%20biosynthesis/image
Example 6. To obtain all available species, codes and genome assemblies:
http://ws.bioinfo.cipf.es/cellbase/rest/latest/species

Using CellBase RESTful web services

CellBase RESTful web services have been implemented using HTTP GET method, so they can be accessed directly from any web browser. We have also developed a Perl client Command Line Interface that can query CellBase web services and can parse large files of identifiers. CellBase web services are also available through some applications developed in our department such as Genome Maps (http://genomemaps.org), RENATO (http://renato.bioinfo.cipf.es) or VARIANT (http://variant.bioinfo.cipf.es), what proves the benefits and potential of this implementation.

OTHER TOOLS

Despite other solutions could be apparently similar, they do not cover the wide variety of biological information included in CellBase. While some of them are quite restricted to specific biological content, others are more general but do not provide web services and often require a local installation. Several resources use the Distributed Annotation System (DAS) protocol (23) to export their data encoded in XML format, which is slower, larger and more difficult to parse than simple text or JSON. Similar to CellBase RESTful web services, DAS can distribute biological information based on genome and protein annotations. However, DAS protocol cannot handle functional annotations or systems biology data, such as gene ontology terms or protein–protein interactions, as CellBase does. Some databases like Ensembl, HapMap or Interpro export their biological information using tools like Biomart which also provide data through WEB services. This can be useful when querying information from a single database, but becomes a problem when users need to link data from several Biomart sources. CellBase facilitates more complex data queries by integrating and linking together different biological sources.

DISCUSSION

In this work, we have designed and implemented CellBase, a database and a collection of RESTful web services that enable quick and easy access to heterogeneous biological information. By joining data from some of the most relevant resources and integrating them into a single standardized database, CellBase provides users with a homogeneous RESTful web service API, with no need to query or download different sources. The integration of RESTful web services technologies has represented a great advantage to provide a comfortable way to query and retrieve this biological information using URLs.

CellBase can be freely accessed in several ways to accommodate different scenarios or users. Web services can be consumed programmatically from any computing language or from a web browser; moreover, a Perl client has also been developed. Biological information and metadata services have been implemented. Database and web servers' infrastructure have been designed to provide a high-availability and high-performance solution.

The database is maintained by a group of biologists and computer scientists. Regular updates will be carried out every few months as new data appear. More species and additional biological data will be also added since the database schema and RESTful web services have been designed to be scalable and cover all biological information requirements. CellBase has proven to be priceless in some of our projects such as Genome Maps, RENATO or VARIANT, and we are, currently, developing new applications using CellBase as a key part of their implementation.

As more genomes are sequenced and more data are available, the access and transfer will become a critical problem. To solve these issues, we are working on moving database and WEB services to the cloud to be able to increase the physical resources. In addition, by doing this, different images would be installed in different geographic regions for faster data queries. The problem of data size is also affecting relational databases as they are approximating to the limit of data they can store in a single machine, to solve this problem are also exploring different solutions like NoSQL databases that allow to store some terabytes of data in a distributed way. Theses changes will be transparent to users as WEB services will remain unchanged.

ACKNOWLEDGEMENTS

The authors thank National Institute of Bioinformatics (www.inab.org) and the CIBER de Enfermedades Raras (CIBERER), both initiatives of the ISCIII, MICINN.

FUNDING

The Spanish Ministry of Science and Innovation (MICINN) [BIO2011-27069]; the Conselleria de

Educacio of the Valencian Community [PROMETEO/2010/001]; National Institute of Bioinformatics (www.inab.org); CIBER de Enfermedades Raras (CIBERER), ISCIII and MICINN; Red Tematica de Investigacion Cooperativa en Cancer (RTICC) [RD06/0020/1019] ISCIII, MICINN and INNPACTO [IPT-010000-2010-43], MICINN. Funding for open access charge: MICINN [BIO2011-27069].

Conflict of interest statement. None declared.

REFERENCES

1. Smith,B., Ashburner,M., Rosse,C., Bard,J., Bug,W., Ceusters,W., Goldberg,L.J., Eilbeck,K., Ireland,A., Mungall,C.J. *et al.* (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.*, **25**, 1251–1255.
2. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fairley,S., Fitzgerald,S. *et al.* (2012) Ensembl 2012. *Nucleic Acids Res.*, **40**, D84–D90.
3. UniProt Consortium. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **40**, D71–D75.
4. Kozomara,A. and Griffiths-Jones,S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**, D152–D157.
5. Xiao,F., Zuo,Z., Cai,G., Kang,S., Gao,X. and Li,T. (2009) miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res.*, **37**, D105–D110.
6. Hsu,S.D., Lin,F.M., Wu,W.Y., Liang,C., Huang,W.C., Chan,W.L., Tsai,W.T., Chen,G.Z., Lee,C.J., Chiu,C.M. *et al.* (2011) miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res.*, **39**, D163–D169.
7. Friedman,R.C., Farh,K.K., Burge,C.B. and Bartel,D.P. (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, **19**, 92–105.
8. Betel,D., Wilson,M., Gabow,A., Marks,D.S. and Sander,C. (2008) The microRNA.org resource: targets and expression. *Nucleic Acids Res.*, **36**, D149–D153.
9. Dreszer,T.R., Karolchik,D., Zweig,A.S., Hinrichs,A.S., Raney,B.J., Kuhn,R.M., Meyer,L.R., Wong,M., Sloan,C.A., Rosenbloom,K.R. *et al.* (2012) The UCSC genome browser database: extensions and updates 2011. *Nucleic Acids Res.*, **40**, D918–D923.
10. 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.
11. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
12. Altshuler,D.M., Gibbs,R.A., Peltonen,L., Dermitzakis,E., Schaffner,S.F., Yu,F., Bonnen,P.E., de Bakker,P.I., Deloukas,P., Gabriel,S.B. *et al.* (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**, 52–58.
13. 1000 Genomes Project Consortium. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
14. Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S. and Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.
15. Stenson,P.D., Ball,E.V., Mort,M., Phillips,A.D., Shiel,J.A., Thomas,N.S., Abeysinghe,S., Krawczak,M. and Cooper,D.N. (2003) Human gene mutation database (HGMD): 2003 update. *Hum. Mutat.*, **21**, 577–581.
16. Johnson,A.D. and O'Donnell,C.J. (2009) An open access database of genome-wide association results. *BMC Med. Genet*, **10**, 6.
17. McKusick,V. (1998) *A Catalog of Human Genes and Genetic Disorders*, 12th edn. John Hopkins University Press, Baltimore, MD.
18. 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.
19. Kerrien,S., Aranda,B., Breuza,L., Bridge,A., Broackes-Carter,F., Chen,C., Duesbury,M., Dumousseau,M., Feuermann,M., Hinz,U. *et al.* (2012) The Intact molecular interaction database in 2012. *Nucleic Acids Res.*, **40**, D841–D846.
20. Croft,D., O'Kelly,G., Wu,G., Haw,R., Gillespie,M., Matthews,L., Caudy,M., Garapati,P., Gopinath,G., Jassal,B. *et al.* (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, **39**, D691–D697.
21. Demir,E., Cary,M.P., Paley,S., Fukuda,K., Lemer,C., Vastrik,I., Wu,G., D'Eustachio,P., Schaefer,C., Luciano,J. *et al.* (2010) The BioPAX community standard for pathway data sharing. *Nature Biotechnol.*, **28**, 935–942.
22. Bhagat,J., Tanoh,F., Nzuobontane,E., Laurent,T., Orlowski,J., Roos,M., Wolstencroft,K., Aleksejevs,S., Stevens,R., Pettifer,S. *et al.* (2010) BioCatalogue: a universal catalogue of web services for the life sciences. *Nucleic Acids Res.*, **38**, W689–W694.
23. Dowell,R.D., Jokerst,R.M., Day,A., Eddy,S.R. and Stein,L. (2001) The distributed annotation system. *BMC Bioinformatics*, **2**, 7.