The BioMart community portal: an innovative alternative to large, centralized data repositories

Damian Smedley, Syed Haider, Steffen Durinck, Luca Pandini, Paolo Provero, James Allen, Olivier Arnaiz, Mohammad Hamza Awedh, Richard Baldock, Giulia Barbiera, et al.

Nucleic Acids Research Advance Access published April 20, 2015

Nucleic Acids Research, 2015 **1**
doi: 10.1093/nar/gkv350

The BioMart community portal: an innovative alternative to large, centralized data repositories

Damian Smedley[1], Syed Haider[2], Steffen Durinck[3], Luca Pandini[4], Paolo Provero[4,5], James Allen[6], Olivier Arnaiz[7], Mohammad Hamza Awedh[8], Richard Baldock[9], Giulia Barbiera[4], Philippe Bardou[10], Tim Beck[11], Andrew Blake[12], Merideth Bonierbale[13], Anthony J. Brookes[11], Gabriele Bucci[4], Iwan Buetti[4], Sarah Burge[6], Cédric Cabau[10], Joseph W. Carlson[14], Claude Chelala[15], Charalambos Chrysostomou[11], Davide Cittaro[4], Olivier Collin[16], Raul Cordova[13], Rosalind J. Cutts[15], Erik Dassi[17], Alex Di Genova[18], Anis Djari[19], Anthony Esposito[20], Heather Estrella[20], Eduardo Eyras[21,22], Julio Fernandez-Banet[20], Simon Forbes[1], Robert C. Free[11], Takatomo Fujisawa[23], Emanuela Gadaleta[15], Jose M. Garcia-Manteiga[4], David Goodstein[14], Kristian Gray[24], José Afonso Guerra-Assunção[15], Bernard Haggarty[9], Dong-Jin Han[25,26], Byung Woo Han[27,28], Todd Harris[29], Jayson Harshbarger[30], Robert K. Hastings[11], Richard D. Hayes[14], Claire Hoede[19], Shen Hu[31], Zhi-Liang Hu[32], Lucie Hutchins[33], Zhengyan Kan[20], Hideya Kawaji[30,34], Aminah Keliet[35], Arnaud Kerhornou[6], Sunghoon Kim[25,26], Rhoda Kinsella[6], Christophe Klopp[19], Lei Kong[36], Daniel Lawson[37], Dejan Lazarevic[4], Ji-Hyun Lee[25,27,28], Thomas Letellier[35], Chuan-Yun Li[38], Pietro Lio[39], Chu-Jun Liu[38], Jie Luo[6], Alejandro Maass[18,40], Jerome Mariette[19], Thomas Maurel[6], Stefania Merella[4], Azza Mostafa Mohamed[41], Francois Moreews[10], Ibounyamine Nabihoudine[19], Nelson Ndegwa[42], Céline Noirot[19], Cristian Perez-Llamas[22], Michael Primig[43], Alessandro Quattrone[17], Hadi Quesneville[35], Davide Rambaldi[4], James Reecy[32], Michela Riba[4], Steven Rosanoff[6], Amna Ali Saddiq[44], Elisa Salas[13], Olivier Sallou[16], Rebecca Shepherd[1], Reinhard Simon[13], Linda Sperling[7], William Spooner[45,46], Daniel M. Staines[6], Delphine Steinbach[35], Kevin Stone[33], Elia Stupka[4], Jon W. Teague[1], Abu Z. Dayem Ullah[15], Jun Wang[36], Doreen Ware[45], Marie Wong-Erasmus[47], Ken Youens-Clark[45], Amonida Zadissa[6], Shi-Jian Zhang[38] and Arek Kasprzyk[4,48,*]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SD, UK, [2]The Weatherall Institute Of Molecular Medicine, University of Oxford, Oxford, OX3 9DS, UK, [3]Genentech, Inc. 1 DNA Way South San Francisco, CA 94080, USA, [4]Center for Translational Genomics and Bioinformatics San Raffaele Scientific Institute, Via Olgettina 58, 20132 Milan, Italy, [5]Dept of Molecular Biotechnology and Health Sciences University of Turin, Italy, [6]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK, [7]Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Université Paris Sud, 1 avenue de la terrasse, 91198 Gif sur Yvette, France, [8]Department of Electrical and Computer Engineering, Faculty of Engineering, King Abdulaziz University, Jeddah, Saudi Arabia, [9]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, Western General Hospital, Edinburgh, EH4 2XU, UK, [10]Sigenae, INRA, Castanet-Tolosan, France, [11]Department of Genetics, University of Leicester, University Road, Leicester, LE1 7RH, UK, [12]MRC Harwell, Harwell Science and Innovation Campus, Oxfordshire, OX11 0RD, UK, [13]International Potato Center (CIP), Lima, 1558, Peru, [14]Department of Energy, Joint Genome Institute, Walnut Creek, USA, [15]Centre for

[*]To whom correspondence should be addressed. Tel: +39 02 26439139; Fax: +39 02 2643 4153; Email: Arek.Kasprzyk@gmail.com

Molecular Oncology, Barts Cancer Institute, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ, UK, [16]IRISA-INRIA, Campus de Beaulieu 35042 Rennes, France, [17]Laboratory of Translational Genomics, Centre for Integrative Biology, University of Trento, Trento, Italy, [18]Center for Mathematical Modeling and Center for Genome Regulation, University of Chile, Beauchef 851, 7th floor, Chile, [19]Plate-forme bio-informatique Genotoul, Mathématiques et Informatique Appliquées de Toulouse, INRA, Castanet-Tolosan, France, [20]Oncology Computational Biology, Pfizer, La Jolla, USA, [21]Catalan Institute for Research and Advanced Studies (ICREA), Passeig Lluis Companys 23, E-08010 Barcelona, Spain, [22]Universitat Pompeu Fabra, Dr Aiguader 88 E-08003 Barcelona, Spain, [23]Kasuza DNA Research Institute, Chiba, 292–0818, Japan, [24]HUGO Gene Nomenclature Committee (HGNC), European Bioinformatics Institute (EMBL-EBI) Wellcome Trust Genome Campus, Hinxton, CB10 1SD, UK, [25]Medicinal Bioconvergence Research Center, College of Pharmacy, Seoul National University, Seoul 151–742, Republic of Korea, [26]Department of Molecular Medicine and Biopharmaceutical Sciences, Seoul National University, Seoul 151–742, Republic of Korea, [27]Research Institute of Pharmaceutical Sciences, College of Pharmacy, Seoul National University, Seoul 151–742, Republic of Korea, [28]Information Center for Bio-pharmacological Network, Seoul National University, Suwon 443–270, Republic of Korea, [29]Ontario Institute for Cancer Research, Toronto, M5G 0A3, Canada, [30]RIKEN Center for Life Science Technologies (CLST), Division of Genomic Technologies (DGT), Kanagawa, 230–0045, Japan, [31]School of Dentistry and Dental Research Institute, University of California Los Angeles (UCLA), Los Angeles, CA 90095–1668, USA, [32]Iowa State Univeristy, USA, [33]Mouse Genomic Informatics Group, The Jackson Laboratory, Bar Harbor, ME 04609, USA, [34]RIKEN Preventive Medicine and Diagnosis Innovation Program, Saitama 351–0198, Japan, [35]INRA URGI Centre de Versailles, bâtiment 18 Route de Saint Cyr 78026 Versailles, France, [36]Center for Bioinformatics, State Key Laboratory of Protein and Plant Gene Research, College of Life Sciences, Peking University, Beijing, 100871, P.R. China, [37]VectorBase, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SD, UK, [38]Institute of Molecular Medicine, Peking University, Beijing, China, [39]Computer Laboratory, University of Cambridge, Cambridge, CB3 0FD, UK, [40]Department of Mathematical Engineering, University of Chile, Av. Beauchef 851, 5th floor, Santiago, Chile, [41]Departament of Biochemistry, Faculty of Science for Girls, King Abdulaziz University, Jeddah, Saudi Arabia, [42]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, PO Box 281, 17177 Stockholm, Sweden, [43]Inserm U1085 IRSET, University of Rennes 1, 35042 Rennes, France, [44]Department of Biological Sciences, Faculty of Science for Girls, King Abdulaziz University, Jeddah, Saudi Arabia, [45]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA, [46]Eagle Genomics Ltd., Babraham Research Campus, Cambridge, CB22 3AT, UK, [47]Human Longevity, Inc. 10835 Road to the Cure 140 San Diego, CA 92121, USA and [48]Department of Biological Sciences, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia

Received February 09, 2015; Revised March 21, 2015; Accepted April 02, 2015

ABSTRACT

The BioMart Community Portal (www.biomart.org) is a community-driven effort to provide a unified interface to biomedical databases that are distributed worldwide. The portal provides access to numerous database projects supported by 30 scientific organizations. It includes over 800 different biological datasets spanning genomics, proteomics, model organisms, cancer data, ontology information and more. All resources available through the portal are independently administered and funded by their host organizations. The BioMart data federation technology provides a unified interface to all the available data. The latest version of the portal comes with many new databases that have been created by our ever-growing community. It also comes with better support and extensibility for data analysis and visualization tools. A new addition to our toolbox, the enrichment analysis tool is now accessible through graphical and web service interface. The BioMart community portal averages over one million requests per day. Building on this level of service and the wealth of information that has become available, the BioMart Community Portal has introduced a new, more scalable and cheaper alternative to the large data stores maintained by specialized organizations.

INTRODUCTION

The methods of data generation and processing that are utilized in biomedical sciences have radically changed in recent years. With the advancement of new high-throughput technologies, data have grown in terms of quantity as well as complexity. However, the significance of the information that is hidden in the newly generated experimental data can only be deciphered by linking it to other types of biological data that have been accumulated previously. As a result there are already numerous bioinformatics resources and new ones are constantly being created. Typically, each resource comes with its own query interface. This poses a problem for the scientists who want to utilize such resources in their research. Even the simplest task such as compil-

Figure 1. BioMart community databases and their host countries.

ing results from a few existing resources is challenging due to the lack of a complete, up to date catalogue of already existing resources and the necessity of constantly learning how to navigate new query interfaces. A different challenge is faced by collaborating groups of scientists who independently generate or maintain their own data. Such collaborations are seriously hampered by the lack of a simple data management solution that would make it possible to connect their disparate, geographically distributed data sources and present them in a uniform way to other scientists. The BioMart project has been set up to address these challenges.

SOFTWARE

BioMart is an open source data management system, which is based on a data federation model (1). Under this model, each data source is managed, updated and released independently by their host organization while the BioMart software provides a unified view of these sources that are distributed worldwide. The data sources are presented to the user through a unified set of graphical and programmatic interfaces so that they appear to be a single integrated database. To navigate this database and compile a query the user does not have to learn the underlying structure of each data source but instead use a set of simple abstractions: datasets, filters and attributes. Once a user's input is provided, the software distributes parts of the query to individual data sources, collects the data and presents the user with the unified result set.

The BioMart software is data agnostic and its applications are not limited to biological data. It is cross-platform and supports many popular relational database managements systems, including MySQL, Oracle, PostgreSQL. It also supports many third party packages such as Taverna (2), Galaxy (3), Cytoscape (4) and biomaRt (5), which part of the Bioconductor (6) library.

The BioMart project currently maintains two independent code bases: one written in Java and one written in Perl. For more information about the architecture and capabilities of each of the packages please refer to previous publications (1,7). The latest version of the Java based BioMart software has been significantly enhanced with new additions to the existing collection of graphical user interfaces (GUIs). It has also been re-engineered to provide better support and extensibility for data analysis and visualization tools. The first of the BioMart tools based on this new framework has already been implemented and is accessible from the BioMart Community Portal.

The BioMart project adheres to the open source philosophy that promotes collaboration and code reuse. Two good examples of how this philosophy benefits the scientific community are provided by two independent research groups. The INRA group based in Toulouse, France has recently released a software package called RNAbrowse (RNA-Seq De Novo Assembly Results Browser) (8). The Pfizer group based in La Jolla, USA has just announced the release of OASIS: A Web-based Platform for Exploratory Analysis of Cancer Genome and Transcriptome data (www.oasis-genomics.org). Both of these software packages are based on the BioMart software.

DATA

The BioMart community consists of a wide spectrum of different research groups that use the BioMart technology to provide access to their databases. It currently comprises 30 scientific organizations supporting 38 database projects that contain over 800 different biological datasets spanning ge-

nomics, proteomics, model organisms, cancer data, ontology information and more. The BioMart community is constantly growing and since the last publication (9), 11 new database projects have become available. As new BioMart databases become available locally they also become gradually integrated into the BioMart Community Portal. The main function of the portal is to provide a convenient single point of access to all available data that is distributed worldwide (Figure 1). All BioMart databases that are included in the portal are independently administered and funded. Table 1 provides a detailed list of all BioMart community resources as of March 2015.

PORTAL

The current version of the BioMart Community Portal operates two different instances of the web server: one implemented in Perl and the other in Java. Both servers support complex database searches and although they use different types of GUIs, they share the same navigation and query compilation logic based on selection of datasets, filters and attributes (9,10). The Java version of the portal also includes a section for specialized tools, which consists of the following: Sequence retrieval, ID Converter and Enrichment Analysis. Sequence retrieval allows easy querying of sequences while the ID Converter tool allows users to enter or upload a list of identifiers in any format (currently supported by Ensembl), and retrieve the same list converted to any other supported format. The enrichment tool supports enrichment analysis of genes in all species included in the current Ensembl release. For each of those species a broad range of gene identifiers is available. Furthermore, the tool supports cross species analysis using Ensembl homology data. For instance, it is possible to perform a one step enrichment analysis against a human disease dataset using experimental data from any of the species for which human homology data is available. Finally, the enrichment tool facilitates analysis of BED files containing genomic features such as Copy Number Variations or Differentially Methylated Regions. The output is provided in tabular and network graphic format (Figure 2).

WEB SERVICE

The BioMart Community Portal handles queries from several interfaces such as:

- PERL API
- Java API
- Web interfaces
- URL based access
- RESTful web service
- SPARQL

For more detailed description of all the interfaces please refer to earlier publications (1,7). In the section below we provide a description and compare the REST-based web service, which is implemented in Perl and its counterpart, which is implemented in Java. It is worth noting that the web service maintains the same query interface both in Perl and Java implementations. For example, the web service query (Figure 3A) can be run against java-based server as follows:

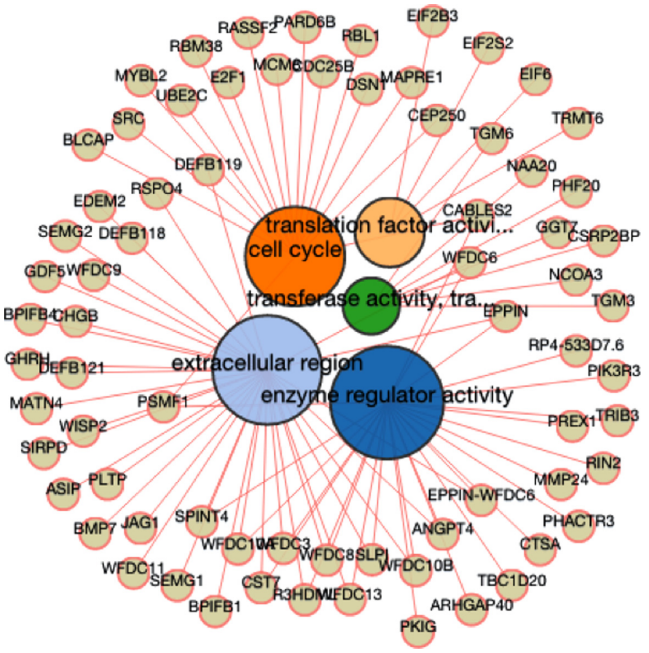

Figure 2. The network graphic output of the BioMart enrichment tool. The Gene Ontology (GO) enrichment analysis was performed using BED file containing human data. This tool is also accessible through web services (Java version only). The programmatic access complies with a standard BioMart interface: dataset, filter and attribute.

curl –data-urlencode query@query.xml http://central.biomart.org/martservice/results

or its Perl-based counter-part as below

curl –data-urlencode query@query.xml http://www.biomart.org/biomart/martservice

By default, query sets the attribute processor to 'TSV' requesting tab-delimited results (Figure 3B). Alternatively, by setting processor to 'JSON', would return JSON formatted results (Figure 3C), which are readily consumable by third-party web-based clients saving overhead of parsing and format translations. Please note that JSON format is only available in the java version.

A simple way to compile a web service query for later programmatic use is to use one of the web GUIs and generate the query XML using REST/SOAP button. After following the steps outlined by the GUI and clicking the 'results' button, the user needs to click the REST/SOAP button, save the query and run it as described above. Alternatively a user can take advantage of the programmatic access to all the metadata defining marts, datasets, filters and attributes. The access to the metadata served by the Java and Perl BioMart servers is provided using the following webservice requests:

Java (central.biomart.org)

- registry information: http://central.biomart.org/martservice/portal
- available marts: http://central.biomart.org/martservice/marts
- datasets available for a config: http://central.biomart.org/martservice/datasets?config=snp_config
- attributes available for a dataset:

Table 1. BioMart community databases and their host organizations

Database	Description	Host	Reference
Animal Genome databases[a,b]	Agriculturally important livestock genomes	Iowa State University, US	NA
Atlas of UTR Regulatory Activity (AURA)[a]	Meta-database centred on mapping post-transcriptional (PTR) interactions of trans-factors with human and mouse untranslated regions (UTRs) of mRNAs	University of Trento, Italy	(36)
BCCTB Bioinformatics Portal[a]	Portal for mining omics data on breast cancer from published literature and experimental datasets	Breast Cancer Campaign/Barts Cancer Institute UK	(37)
Cildb	Database for eukaryotic cilia and centriolar structures, integrating orthology relationships for 44 species with high-throughput studies and OMIM	Centre National de la Recherche Scientifique (CNRS), France	(38)
COSMIC	Somatic mutation information relating to human cancers	Wellcome Trust Sanger Institute (WTSI), UK	(39)
DAPPER[a]	Mass spec identified protein interaction networks in *Drosophila* cell cycle regulation	Department of Genetics, University of Cambridge, Cambridge, UK	NA
EMAGE	In situ gene expression data in the mouse embryo	Medical Research Council, Human Genetics Unit (MRC HGU), UK	(40)
Ensembl	Genome databases for vertebrates and other eukaryotic species	Wellcome Trust Sanger Institute (WTSI), UK	(41)
Ensembl Genomes	Ensembl Fungi, Metazoa, Plants and Protists	European Bioinformatics Institute (EBI), UK	(41)
Euraexpress	Transcriptome atlas database for mouse embryo	Medical Research Council, Human Genetics Unit (MRC HGU), UK	(42)
EuroPhenome	Mouse phenotyping data	Harwell Science and Innovation Campus (MRC Harwell), UK	(15)
FANTOM5[a]	The FANTOM5 project mapped a promoter level expression atlas in human and mouse. The FANTOM5 BioMart instance provides the set of promoters along with annotation.	RIKEN Center for Life Science Technologies (CLST), Japan	(16)
GermOnLine	Cross-species microarray expression database focusing on germline development, meiosis, and gametogenesis as well as the mitotic cell cycle	Institut national de la santé et de la recherche médicale (Inserm), France	(17)
GnpIS[a]	Genetic and Genomic Information System (GnpIS)	Institut Nationale de Recherche Agronomique (INRA), Unité de Recherche en Génomique-Info (URGI), France	(18)
Gramene	Agriculturally important grass genomes	Cold Spring Harbor Laboratory (CSHL), US	(43)
GWAS Central[a]	GWAS Central provides a comprehensive curated collection of summary level findings from genetic association studies	University of Leicester, UK	(19)
HapMap	Multi-country effort to identify and catalog genetic similarities and differences in human beings	National Center for Biotechnology Information (NCBI), US	(20)
HGNC	Repository of human gene nomenclature and associated resources	European Bioinformatics Institute (EBI), UK	(21)
i-Pharm[a]	PharmDB-K is an integrated bio-pharmacological network databases for TKM (Traditional Korean Medicine)	Information Center for Bio-pharmacological Network (i-Pharm), South Korea	(22)
InterPro	Integrated database of predictive protein 'signatures' used for the classification and automatic annotation of proteins and genomes	European Bioinformatics Institute (EBI), UK	(44)
KazusaMart	Cyanobase, rhizobia, and plant genome databases	Kazusa DNA Research Institute (Kazusa), Japan	NA
MGI	Mouse genome features, locations, alleles, and orthologs	Jackson Laboratory, US	(23)
Pancreatic Expression Database	Results from published literature	Barts Cancer Institute UK	(24)
ParameciumDB	Paramecium genome database	Centre National de la Recherche Scientifique (CNRS), France	(25)
Phytozome	Comparative genomics of green plants	Joint Genome Institute (JGI)/Center for Integrative Genomics (CIG), US	(26)

Table 1. Continued

Database	Description	Host	Reference
Potato Database	Potato and sweetpotato phenotypic and genomic information	International Potato Center (CIP), Peru	NA
PRIDE	Repository for protein and peptide identifications	European Bioinformatics Institute (EBI), UK	(45)
Regulatory Genomics Group[a]	Predictive Models of Gene Regulation from High-Throughput Epigenomics Data	Universitat Pompeu Fabra (UPF), Spain	(27)
Rfam[a]	The Rfam database is a collection of RNA families, each represented by multiple sequence alignments, consensus secondary structures and covariance models (CMs).	Wellcome Trust Sanger Institute (WTSI), UK	(28)
RhesusBase[a]	A knowledgebase for the monkey research community	Peking University, China	(29)
Rice-Map	Rice (japonica and indica) genome annotation database	Peking University, China	(30)
SalmonDB	Genomic information for Atlantic salmon, rainbow trout, and related species	Center for Mathematical Modeling and Center for Genome Regulation (CMM), Chile	(31)
sigReannot	Aquaculture and farm animal species microarray probes re-annotation	INRA - French National Institute of Agricultural Research, France	(46)
UniProt	Protein sequence and functional information	European Bioinformatics Institute (EBI), UK	(32)
VectorBase	Genome information for invertebrate vectors of human pathogens	University of Notre Dame, US	(33)
VEGA	Manual annotation of vertebrate genome sequences	Wellcome Trust Sanger Institute (WTSI), UK	(34)
WormBase	*C. elegans* and related nematode genomic information	Cold Spring Harbor Laboratory (CSHL), US	(35)

[a]Denotes new databases that have become available since last publication (9).
[b]Denotes new databases that are not yet integrated into the portal.

http://central.biomart.org/martservice/attributes?datasets=btaurus_snp&config=snp_config
- filters available for a dataset: http://central.biomart.org/martservice/filters?datasets=btaurus_snp&config=snp_config

Perl (www.biomart.org)

- registry information: http://www.biomart.org/biomart/martservice?type=registry
- datasets available for a mart: http://www.biomart.org/biomart/martservice?type=datasets&mart=ensembl
- attributes available for a dataset: http://www.biomart.org/biomart/martservice?type=attributes&dataset=oanatinus_gene_ensembl
- filters available for a dataset: http://www.biomart.org/biomart/martservice?type=filters&dataset=oanatinus_gene_ensembl
- configuration for a dataset: http://www.biomart.org/biomart/martservice?type=configuration&dataset=oanatinus_gene_ensembl

Please note that the granularity between mart and dataset has been improved in the Java version through the introduction of multiple dataset configs. This facilitates the end-users to browse various views of the same dataset, which are presented through the portal either using a different GUI or subsets of data.

QUERY EXAMPLES

Given the coverage of the current BioMart datatsets, many relevant biological questions can be answered. For example, a researcher who has detected potentially pathogenic variants in FGFR2 (ENSG00000066468) from exome sequencing patients may be interested if the same variants have been previously described and if they were associated with the same or similar diseases. To answer this, integrated data from Ensembl can be queried as shown in Table 2 to display all known variants annotated within FGFR2 that are predicted as pathogenic by SIFT (11) and Polyphen (12). The genomic position outputs can be compared to the researcher's variants and the phenotype data used to assess candidacy for their cases. For example, the first batch of results shows a C->G variant at position 121520160 on chromosome 10 that is associated with Apert syndrome (OMIM:176943).

Another common use case that BioMart is used for is to analyse a list of genes to establish whether they are associated with particular protein functions, pathways or diseases more often than would be expected by chance (enrichment analysis). For example, a researcher may have discovered that AURKA, AURKB, AURKC, PLK1, CDK1 and CDK4 are differentially expressed in their experiment and used BioMart's enrichment tool with its default settings to analyse these genes. The results show that these genes are enriched for involvement in the cell cycle, kinase activity and mitotic nuclear division amongst others. Many other real usage examples are documented in our previous paper (10)

A

```
<!DOCTYPE Query>
<Query client="true" processor="TSV" limit="-1" header="1">
    <Dataset name="hsapiens_gene_ensembl" config="gene_ensembl_config">
        <Filter name="chromosome_name" value="1" filter_list=""/>
        <Attribute name="ensembl_gene_id"/>
        <Attribute name="ensembl_transcript_id"/>
    </Dataset>
</Query>
```

B

Ensembl Gene ID	Ensembl Transcript ID
ENSG00000231510	ENST00000443270
ENSG00000162444	ENST00000315901
ENSG00000162444	ENST00000294435

C

```
object {
    "total": number 17022,
    "data": array [
    object {
        "Ensembl Gene ID": string "ENSG00000231510",
        "Ensembl Transcript ID": string "ENST00000443270"
    },
    object {
        "Ensembl Gene ID": string "ENSG00000162444",
        "Ensembl Transcript ID": string "ENST00000315901"
    },
    object {
        "Ensembl Gene ID": string "ENSG00000162444",
        "Ensembl Transcript ID": string "ENST00000294435"
    }
    ]
}
```

Figure 3. The XML web service query (**A**) and the corresponding two types of output: tab delimited following setting a processor to 'TSV' (**B**) and JSON following setting processor to 'JSON'.

Table 2. Query to display phenotypic consequence for known, pathogenic variants in FGFR2

Database and dataset	Filters	Attributes
Ensembl 78 Short Variations (WTSI, UK) Homo sapiens Short Variation (SNPs and indels) (GRCh38)	Ensembl Gene ID(s): ENSG00000066468	Chromosome name
	SIFT Prediction: deleterious	Chromosome position start (bp)
		Chromosome position end (bp)
	PolyPhen Prediction: probably damaging	Strand
		Variant Alleles
		Ensembl Gene ID
		Consequence to transcript
		Associated variation names
		Study External Reference
		Source name
		Associated gene with phenotype
		Phenotype description

and the BioMart special issue in Database: the journal of biological databases and biocuration (www.oxfordjournals.org/our_journals/databa/biomart_virtual_issue.html).

CONCLUSIONS

Since its conception as a data-mining interface for the Human Genome Project (13) BioMart has rapidly grown to become an international collaboration involving a large number of different groups and organizations both in academia and in industry (14). It has been successfully applied to many different types of data including genomics, proteomics, model organisms, cancer data, etc., proving that its generic data model is widely applicable (15–53). BioMart has also provided a first successful solution for the unprecedented data management needs of the International Cancer Genome Consortium proving that the federated model scales well with the amounts of data generated by Next Generation Sequencing (48).

There are a number of important factors that contributed to the BioMart's success and its adoption by many different types of projects around the world as their data management platform. BioMart's ability to quickly deploy a website hosting any type of data, user-friendly GUI, several programmatic interfaces and support for third party tools has proved to be an attractive solution for data managers who were in need of a rapid and reliable solution for their user community. BioMart has also proven to be a platform of choice for many smaller organizations that lack the necessary resources to embark on the development of their own data management solution. As a result, more and more database projects have become accessible through the BioMart interface. The arrival of these new resources coupled with the data federation technology provided by the BioMart software has galvanized the creation of the BioMart Community Portal. The federated model has proven to be very cost-effective since all development and maintenance of individual databases is left to the individual data providers. It also has proven to be very scalable as the internet and database traffic is handled by the local BioMart servers. As a result the BioMart Community Portal service has grown impressively not only in terms of available data but also the level of service. The BioMart community portal now averages over million requests per our services per day. Building on this level of service and the wealth of information that has become accessible through the BioMart interface, the BioMart Community Portal has effectively introduced a new, more scalable and much more cost-effective alternative to the large data stores maintained by specialized organizations.

ACKNOWLEDGEMENT

We are grateful to the following organizations for providing support for the BioMart project: European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK; Ontario Institute for Cancer Research, Toronto, Canada; San Raffaele Scientific Institute, Milan, Italy and King Abdulaziz University, Jeddah, Saudi Arabia.

FUNDING

The BioMart Community Portal is a collaborative, community effort and as such it is the product of the efforts of dozens of different groups and organizations. The individual data sources that the portal comprises are funded separately and independently. In particular: Wellcome Trust [077012/Z/05/Z to COSMIC mart]; Spanish Government [BIO2011–23920 and CSD2009–00080 to BioMart database of the Regulatory Genomics group at Pompeu Fabra University]; Sandra Ibarra Foundation for Cancer [FSI2013]; Breast Cancer Campaign Tissue Bank [09TBBAR to BCCTB bioinformatics portal]; Office of Science of the U.S. Department of Energy [DE-AC02–05CH11231 to Phytozome]; Global Frontier Project (to i-Pharm research) funded by the Ministry of Science, ICT and Future Planning through the National Research Foundation of Korea (NRF-2013M3A6A4043695); Agence National de la Recherche [ANR-10-BLAN-1122, ANR-12-BSV6–0017–03, ANR-14-CE10–0005–03 to ParameciumDB and cilDB]; Centre National de la Recherche Scientifique; Center for Genome Regulation [SalmonDB; Fondap-1509007 to A.M. and A.D.G.]; Center for Mathematical Modelling [Basal-PFB 03 to A.M. and A.D.G.]; Wellcome Trust (WT095908 and WT098051 to R.K., T.M. and A.Z.); European Molecular Biology Laboratory; Japanese Ministry of Education, Culture, Sports, Science and Technology [FANTOM5 BioMart; for RIKEN OSC and RIKEN PMI to Yoshihide Hayashizaki, and for RIKEN CLST]. Deanship of Scientific Research (DSR) King Abdulaziz University (96–130–35-HiCi to M.H.A., A.M.M., A.A.S. and A.K.). Funding for open access charge: King Abdulaziz University.

Conflict of interest statement. None declared.

REFERENCES

1. Zhang,J., Haider,S., Baran,J., Cros,A., Guberman,J.M., Hsu,J., Liang,Y., Yao,L. and Kasprzyk,A. (2011) BioMart: a data federation framework for large collaborative projects. *Database*, bar038.
2. Hull,D., Wolstencroft,K., Stevens,R., Goble,C., Pocock,M.R., Li,P. and Oinn,T. (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Res.*, **34**, W729–W732.
3. Giardine,B., Riemer,C., Hardison,R.C., Burhans,R., Elnitski,L., Shah,P., Zhang,Y., Blankenberg,D., Albert,I., Taylor,J. *et al.* (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.*, **15**, 1451–1455.
4. Cline,M.S., Smoot,M., Cerami,E., Kuchinsky,A., Landys,N., Workman,C., Christmas,R., Avila-Campilo,I., Creech,M., Gross,B. *et al.* (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.*, **2**, 2366–2382.
5. Durinck,S., Moreau,Y., Kasprzyk,A., Davis,S., De Moor,B., Brazma,A. and Huber,W. (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, **21**, 3439–3440.
6. Reimers,M. and Carey,V.J. (2006) Bioconductor: an open source framework for bioinformatics and computational biology. *Methods Enzymol.*, **411**, 119–134.
7. Haider,S., Ballester,B., Smedley,D., Zhang,J., Rice,P. and Kasprzyk,A. (2009) BioMart Central Portal–unified access to biological data. *Nucleic Acids Res.*, **37**, W23–W27.
8. Mariette,J., Noirot,C., Nabihoudine,I., Bardou,P., Hoede,C., Djari,A., Cabau,C. and Klopp,C. (2014) RNAbrowse: RNA-Seq de novo assembly results browser. *PLoS One*, **9**, e96821.
9. Guberman,J.M., Ai,J., Arnaiz,O., Baran,J., Blake,A., Baldock,R., Chelala,C., Croft,D., Cros,A., Cutts,R.J. *et al.* (2011) BioMart

Central Portal: an open database network for the biological community. *Database*, bar041.
10. Smedley,D., Haider,S., Ballester,B., Holland,R., London,D., Thorisson,G. and Kasprzyk,A. (2009) BioMart–biological queries made easy. *BMC Genomics*, **10**, 22.
11. C Ng,Pauline and Henikoff,Steven (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–3814.
12. A Adzhubei,Ivan, Schmidt,Steffen, Peshkin,Leonid, E Ramensky,Vasily, Gerasimova,Anna, Bork,Peer, S Kondrashov,Alexey and R Sunyaev,Shamil (2010) A method and server for predicting damaging missense mutations. *Nature*, **7**, 248–249.
13. Kasprzyk,A., Keefe,D., Smedley,D., London,D., Spooner,W., Melsopp,C., Hammond,M., Rocca-Serra,P., Cox,T. and Birney,E. (2004) EnsMart: a generic system for fast and flexible access to biological data. *Genome Res.*, **14**, 160–169.
14. Kasprzyk,A. (2011) BioMart: driving a paradigm change in biological data management. *Database*, bar049.
15. Mallon,A.M., Iyer,V., Melvin,D., Morgan,H., Parkinson,H., Brown,S.D., Flicek,P. and Skarnes,W.C. (2012) Accessing data from the International Mouse Phenotyping Consortium: state of the art and future plans. *Mamm. Genome*, **23**, 641–652.
16. Lizio,M., Harshbarger,J., Shimoji,H., Severin,J., Kasukawa,T., Sahin,S., Abugessaisa,I., Fukuda,S., Hori,F., Ishikawa-Kato,S. *et al.* (2015) Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol.*, **16**, 22.
17. Lardenois,A., Gattiker,A., Collin,O., Chalmel,F. and Primig,M. (2010) GermOnline 4.0 is a genomics gateway for germline development, meiosis and the mitotic cell cycle. *Database*, baq030.
18. Steinbach,D., Alaux,M., Amselem,J., Choisne,N., Durand,S., Flores,R., Keliet,A.O., Kimmel,E., Lapalu,N., Luyten,I. *et al.* (2013) GnpIS: an information system to integrate genetic and genomic data from plants and fungi. *Database*, bat058.
19. Beck,T., Hastings,R.K., Gollapudi,S., Free,R.C. and Brookes,A.J. (2014) GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. *Eur. J. Hum. Genet.*, **22**, 949–952.
20. International HapMap Consortium. (2003) The International HapMap Project. *Nature*, **426**, 789–796.
21. Povey,S., Lovering,R., Bruford,E., Wright,M., Lush,M. and Wain,H. (2001) The HUGO Gene Nomenclature Committee (HGNC). *Hum. Genet.*, **109**, 678–680.
22. Lee,H.S., Bae,T., Lee,J.H., Kim,D.G., Oh,Y.S., Jang,Y., Kim,J.T., Lee,J.J., Innocenti,A., Supuran,C.T. *et al.* (2012) Rational drug repositioning guided by an integrated pharmacological network of protein, disease and drug. *BMC Syst. Biol.*, **6**, 80.
23. Shaw,D.R. (2009) Searching the Mouse Genome Informatics (MGI) resources for information on mouse biology from genotype to phenotype. *Curr. Protoc. Bioinformatics*, **2009**, doi:10.1002/0471250953.bi0107s25.
24. Dayem Ullah,A.Z., Cutts,R.J., Ghetia,M., Gadaleta,E., Hahn,S.A., Crnogorac-Jurcevic,T., Lemoine,N.R. and Chelala,C. (2014) The pancreatic expression database: recent extensions and updates. *Nucleic Acids Res.*, **42**, D944–D949.
25. Arnaiz,O. and Sperling,L. (2011) ParameciumDB in 2011: new tools and new data for functional and comparative genomics of the model ciliate Paramecium tetraurelia. *Nucleic Acids Res.*, **39**, D632–D636.
26. Goodstein,D.M., Shu,S., Howson,R., Neupane,R., Hayes,R.D., Fazo,J., Mitros,T., Dirks,W., Hellsten,U., Putnam,N. *et al.* (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.*, **40**, D1178–D1186.
27. Althammer,S., Pages,A. and Eyras,E. (2012) Predictive models of gene regulation from high-throughput epigenomics data. *Comp. Funct. Genomics*, **2012**, 284786.
28. Burge,S.W., Daub,J., Eberhardt,R., Tate,J., Barquist,L., Nawrocki,E.P., Eddy,S.R., Gardner,P.P. and Bateman,A. (2013) Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.*, **41**, D226–D232.
29. Zhang,S.J., Liu,C.J., Shi,M., Kong,L., Chen,J.Y., Zhou,W.Z., Zhu,X., Yu,P., Wang,J., Yang,X. *et al.* (2013) RhesusBase: a knowledgebase for the monkey research community. *Nucleic Acids Res.*, **41**, D892–D905.
30. Wang,J., Kong,L., Zhao,S., Zhang,H., Tang,L., Li,Z., Gu,X., Luo,J. and Gao,G. (2011) Rice-Map: a new-generation rice genome browser. *BMC Genomics*, **12**, 165.
31. Di Genova,A., Aravena,A., Zapata,L., Gonzalez,M., Maass,A. and Iturra,P. (2011) SalmonDB: a bioinformatics resource for Salmo salar and Oncorhynchus mykiss. *Database*, bar050.
32. UniProt Consortium. (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **42**, D191–D198.
33. Megy,K., Emrich,S.J., Lawson,D., Campbell,D., Dialynas,E., Hughes,D.S., Koscielny,G., Louis,C., Maccallum,R.M., Redmond,S.N. *et al.* (2012) VectorBase: improvements to a bioinformatics resource for invertebrate vector genomics. *Nucleic Acids Res.*, **40**, D729–D734.
34. Harrow,J.L., Steward,C.A., Frankish,A., Gilbert,J.G., Gonzalez,J.M., Loveland,J.E., Mudge,J., Sheppard,D., Thomas,M., Trevanion,S. *et al.* (2014) The Vertebrate Genome Annotation browser 10 years on. *Nucleic Acids Res.*, **42**, D771–D779.
35. Harris,T.W., Baran,J., Bieri,T., Cabunoc,A., Chan,J., Chen,W.J., Davis,P., Done,J., Grove,C., Howe,K. *et al.* (2014) WormBase 2014: new views of curated biology. *Nucleic Acids Res.*, **42**, D789–D793.
36. Dassi,E., Re,A., Leo,S., Tebaldi,T., Pasini,L., Peroni,D. and Quattrone,A. (2014) AURA 2 Empowering discovery of post-transcriptional networks. *Translation*, **2**, e27738.
37. Cutts,R.J., Guerra-Assuncao,J.A., Gadaleta,E., Dayem Ullah,A.Z. and Chelala,C. (2015) BCCTBbp: the Breast Cancer Campaign Tissue Bank bioinformatics portal. *Nucleic Acids Res.*, **43**, D831–D836.
38. Arnaiz,O., Cohen,J., Tassin,A.M. and Koll,F. (2014) Remodeling Cildb, a popular database for cilia and links for ciliopathies. *Cilia*, **3**, 9.
39. Shepherd,R., Forbes,S.A., Beare,D., Bamford,S., Cole,C.G., Ward,S., Bindal,N., Gunasekaran,P., Jia,M., Kok,C.Y. *et al.* (2011) Data mining using the Catalogue of Somatic Mutations in Cancer BioMart. *Database*, **2011**, bar018.
40. Stevenson,P., Richardson,L., Venkataraman,S., Yang,Y. and Baldock,R. (2011) The BioMart interface to the eMouseAtlas gene expression database EMAGE. *Database*, **2011**, bar029.
41. Kinsella,R.J., Kahari,A., Haider,S., Zamora,J., Proctor,G., Spudich,G., Almeida-King,J., Staines,D., Derwent,P., Kerhornou,A. *et al.* (2011) Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database*, **2011**, bar030.
42. Diez-Roux,G., Banfi,S., Sultan,M., Geffers,L., Anand,S., Rozado,D., Magen,A., Canidio,E., Pagani,M., Peluso,I. *et al.* (2011) A high-resolution anatomical atlas of the transcriptome in the mouse embryo. *PLoS Biol.*, **9**, e1000582.
43. Spooner,W., Youens-Clark,K., Staines,D. and Ware,D. (2012) GrameneMart: the BioMart data portal for the Gramene project. *Database*, **2012**, bar056.
44. Jones,P., Binns,D., McMenamin,C., McAnulla,C. and Hunter,S. (2011) The InterPro BioMart: federated query and web service access to the InterPro Resource. *Database*, **2011**, bar033.
45. Ndegwa,N., Cote,R.G., Ovelleiro,D., D'Eustachio,P., Hermjakob,H., Vizcaino,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.
46. Moreews,F., Rauffet,G., Dehais,P. and Klopp,C. (2011) SigReannot-mart: a query environment for expression microarray probe re-annotations. *Database*, **2011**, bar025.
47. Cutts,R.J., Gadaleta,E., Lemoine,N.R. and Chelala,C. (2011) Using BioMart as a framework to manage and query pancreatic cancer data. *Database*, **2011**, bar024.
48. Zhang,J., Baran,J., Cros,A., Guberman,J.M., Haider,S., Hsu,J., Liang,Y., Rivkin,E., Wang,J., Whitty,B. *et al.* (2011) International Cancer Genome Consortium Data Portal–a one-stop shop for cancer genomics data. *Database*, **2011**, bar026.
49. Oakley,D.J., Iyer,V., Skarnes,W.C. and Smedley,D. (2011) BioMart as an integration solution for the International Knockout Mouse Consortium. *Database*, **2011**, bar028.
50. 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.

51. Perez-Llamas,C., Gundem,G. and Lopez-Bigas,N. (2011) Integrative cancer genomics (IntOGen) in Biomart. *Database*, **2011**, bar039.
52. Koscielny,G., Yaikhom,G., Iyer,V., Meehan,T.F., Morgan,H., Atienza-Herrero,J., Blake,A., Chen,C.K., Easty,R., Di Fenza,A. *et al.* (2014) The International Mouse Phenotyping Consortium Web Portal, a unified point of access for knockout mice and related phenotyping data. *Nucleic Acids Res.*, **42**, D802–D809.
53. Wilkinson,P., Sengerova,J., Matteoni,R., Chen,C.K., Soulat,G., Ureta-Vidal,A., Fessele,S., Hagn,M., Massimi,M., Pickford,K. *et al.* (2010) EMMA–mouse mutant resources for the international scientific community. *Nucleic Acids Res.*, **38**, D570–D576.