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

To cite this version:

HAL Id: hal-01146849
https://hal-univ-rennes1.archives-ouvertes.fr/hal-01146849
Submitted on 28 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
The BioMart community portal: an innovative alternative to large, centralized data repositories


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

1To whom correspondence should be addressed. Tel: +39 02 26439139; Fax: +39 02 2643 4153; Email: Arek.Kasprzyk@gmail.com

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic Acids Research. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
Molecular Oncology, Barts Cancer Institute, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ, UK, 16IRISA-INRIA, Campus de Beaulieu 35042 Rennes, France, 17Laboratory of Translational Genomics, Centre for Integrative Biology, University of Trento, Trento, Italy, 18Center for Mathematical Modeling and Center for Genome Regulation, University of Chile, Beauchef 851, 7th floor, Chile, 19Plate-forme bio-informatique Genotoul, Mathématiques et Informatique Appliquées de Toulouse, INRA, Castanet-Tolosan, France, 20Oncology Computational Biology, Pfizer, La Jolla, USA, 21Catalan Institute for Research and Advanced Studies (ICREA), Passeig Lluis Companys 23, E-08010 Barcelona, Spain, 22Universitat Pompeu Fabra, Dr Aiguader 88 E-08003 Barcelona, Spain, 23Kasuza DNA Research Institute, Chiba, 292–0818, Japan, 24HUGO Gene Nomenclature Committee (HGNC), European Bioinformatics Institute (EMBL-EBI) Wellcome Trust Genome Campus, Hinxton, CB10 1SD, UK, 25Medicinal Bioconvergence Research Center, College of Pharmacy, Seoul National University, Seoul 151–742, Republic of Korea, 26Department of Molecular Medicine and Biopharmaceutical Sciences, Seoul National University, Seoul 151–742, Republic of Korea, 27Research Institute of Pharmaceutical Sciences, College of Pharmacy, Seoul National University, Seoul 151–742, Republic of Korea, 28Information Center for Bio-pharmacological Network, Seoul National University, Suwon 443–270, Republic of Korea, 29Ontario Institute for Cancer Research, Toronto, M5G 0A3, Canada, 30RIKEN Center for Life Science Technologies (CLST), Division of Genomic Technologies (DGT), Kanagawa, 230–0045, Japan, 31School of Dentistry and Dental Research Institute, University of California Los Angeles (UCLA), Los Angeles, CA 90095–1668, USA, 32Iowa State Univeristy, USA, 33Mouse Genomic Informatics Group, The Jackson Laboratory, Bar Harbor, ME 04609, USA, 34RIKEN Preventive Medicine and Diagnosis Innovation Program, Saitama 351–0198, Japan, 35INRA URGI Centre de Versailles, bâtiment 18 Route de Saint Cyr 78026 Versailles, France, 36Center for Bioinformatics, State Key Laboratory of Protein and Plant Gene Research, College of Life Sciences, Peking University, Beijing, 100871, PR. China, 37VectorBase, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SD, UK, 38Institute of Molecular Medicine, Peking University, Beijing, China, 39Computer Laboratory, University of Cambridge, Cambridge, CB3 0FD, UK, 40Department of Mathematical Engineering, University of Chile, Av. Beauchef 851, 5th floor, Santiago, Chile, 41Department of Biochemistry, Faculty of Science for Girls, King Abdulaziz University, Jeddah, Saudi Arabia, 42Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, PO Box 281, 17177 Stockholm, Sweden, 43Inserm U1085 IRSET, University of Rennes 1, 35042 Rennes, France, 44Department of Biological Sciences, Faculty of Science for Girls, King Abdulaziz University, Jeddah, Saudi Arabia, 45Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA, 46Eagle Genomics Ltd., Babraham Research Campus, Cambridge, CB2 3AT, UK, 47Human Longevity, Inc. 10835 Road to the Cure 140 San Diego, CA 92121, USA and 48Department of Biological Sciences, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia

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-
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 management 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-
omics, 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:

```
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:
```

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.
<table>
<thead>
<tr>
<th>Database</th>
<th>Description</th>
<th>Host</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Animal Genome databases&lt;sup&gt;a,b&lt;/sup&gt;</td>
<td>Agriculturally important livestock genomes</td>
<td>Iowa State University, US</td>
<td>NA</td>
</tr>
<tr>
<td>Atlas of UTR Regulatory Activity (AURA)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Meta-database centred on mapping post-transcriptional (PTR) interactions of trans-factors with human and mouse untranslated regions (UTRs) of mRNAs</td>
<td>University of Trento, Italy</td>
<td>(36)</td>
</tr>
<tr>
<td>BCCTB Bioinformatics Portal&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Portal for mining omics data on breast cancer from published literature and experimental datasets</td>
<td>Breast Cancer Campaign/Barts Cancer Institute UK</td>
<td>(37)</td>
</tr>
<tr>
<td>Cildb</td>
<td>Database for eukaryotic cilia and centriolar structures, integrating orthology relationships for 44 species with high-throughput studies and OMIM</td>
<td>Centre National de la Recherche Scientifique (CNRS), France</td>
<td>(38)</td>
</tr>
<tr>
<td>COSMIC</td>
<td>Somatic mutation information relating to human cancers</td>
<td>Wellcome Trust Sanger Institute (WTSI), UK</td>
<td>(39)</td>
</tr>
<tr>
<td>DAPPER&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Mass spec identified protein interaction networks in <em>Drosophila</em> cell cycle regulation</td>
<td>Department of Genetics, University of Cambridge, Cambridge, UK</td>
<td>NA</td>
</tr>
<tr>
<td>EMAGE</td>
<td>In situ gene expression data in the mouse embryo</td>
<td>Medical Research Council, Human Genetics Unit (MRC HGU), UK</td>
<td>(40)</td>
</tr>
<tr>
<td>Ensembl</td>
<td>Genome databases for vertebrates and other eukaryotic species</td>
<td>Wellcome Trust Sanger Institute (WTSI), UK</td>
<td>(41)</td>
</tr>
<tr>
<td>Ensembl Genomes</td>
<td>Ensembl Fungi, Metazoa, Plants and Protists</td>
<td>European Bioinformatics Institute (EBI), UK</td>
<td>(41)</td>
</tr>
<tr>
<td>Euraexpress Genomes</td>
<td>Transcriptome atlas database for mouse embryo</td>
<td>Medical Research Council, Human Genetics Unit (MRC HGU), UK</td>
<td>(42)</td>
</tr>
<tr>
<td>EuroPhenome</td>
<td>Mouse phenotyping data</td>
<td>Harwell Science and Innovation Campus (MRC Harwell), UK</td>
<td>(15)</td>
</tr>
<tr>
<td>FANTOM5&lt;sup&gt;a&lt;/sup&gt;</td>
<td>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.</td>
<td>RIKEN Center for Life Science Technologies (CLST), Japan</td>
<td>(16)</td>
</tr>
<tr>
<td>GermOnLine</td>
<td>Cross-species microarray expression database focusing on germline development, meiosis, and gametogenesis as well as the mitotic cell cycle</td>
<td>Institut national de la santé et de la recherche médicale (Inserm), France</td>
<td>(17)</td>
</tr>
<tr>
<td>GnpIS&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Genetic and Genomic Information System (GnpIS)</td>
<td>Institut Nationale de Recherche Agronomique (INRA), Unité de Recherche en Génomique-Info (URGI), France</td>
<td>(18)</td>
</tr>
<tr>
<td>Gramene</td>
<td>Agriculturally important grass genomes</td>
<td>Cold Spring Harbor Laboratory (CSHL), US</td>
<td>(43)</td>
</tr>
<tr>
<td>GWAS Central&lt;sup&gt;b&lt;/sup&gt;</td>
<td>GWAS Central provides a comprehensive curated collection of summary level findings from genetic association studies</td>
<td>University of Leicester, UK</td>
<td>(19)</td>
</tr>
<tr>
<td>HapMap</td>
<td>Multi-country effort to identify and catalog genetic similarities and differences in human beings</td>
<td>National Center for Biotechnology Information (NCBI), US</td>
<td>(20)</td>
</tr>
<tr>
<td>HGNC</td>
<td>Repository of human gene nomenclature and associated resources</td>
<td>European Bioinformatics Institute (EBI), UK</td>
<td>(21)</td>
</tr>
<tr>
<td>i-Pharm&lt;sup&gt;a&lt;/sup&gt;</td>
<td>PharmDB-K is an integrated bio-pharmacological network databases for TKM (Traditional Korean Medicine)</td>
<td>Information Center for Bio-pharmacological Network (i-Pharm), South Korea</td>
<td>NA</td>
</tr>
<tr>
<td>InterPro</td>
<td>Integrated database of predictive protein ‘signatures’ used for the classification and automatic annotation of proteins and genomes</td>
<td>European Bioinformatics Institute (EBI), UK</td>
<td>(44)</td>
</tr>
<tr>
<td>KazusaMart</td>
<td>Cyanobase, rhizobia, and plant genome databases</td>
<td>Kazusa DNA Research Institute (Kazusa), Japan</td>
<td>NA</td>
</tr>
<tr>
<td>MGI</td>
<td>Mouse genome features, locations, alleles, and orthologs</td>
<td>Jackson Laboratory, US</td>
<td>(23)</td>
</tr>
<tr>
<td>Pancreatic Expression Database</td>
<td>Results from published literature Paramecium genome database</td>
<td>Barts Cancer Institute UK</td>
<td>(24)</td>
</tr>
<tr>
<td>ParameciumDB</td>
<td>Paramecium genome database</td>
<td>Centre National de la Recherche Scientifique (CNRS), France</td>
<td>(25)</td>
</tr>
<tr>
<td>Phytozome</td>
<td>Comparative genomics of green plants</td>
<td>Joint Genome Institute (JGI)/Center for Integrative Genomics (CIG), US</td>
<td>(26)</td>
</tr>
</tbody>
</table>
Table 1. Continued

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

<sup>a</sup>Denotes new databases that have become available since last publication (9).
<sup>b</sup>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 datasets, 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).
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

<table>
<thead>
<tr>
<th>Database and dataset</th>
<th>Filters</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensembl 78 Short Variations (WTSI, UK)</td>
<td>Ensembl Gene ID(s): ENSG0000006468</td>
<td>Chromosome name (WTSI, UK)</td>
</tr>
<tr>
<td>Homo sapiens Short Variation (SNPs and indels) (GRCh38)</td>
<td>SIFT Prediction: deleterious</td>
<td>Chromosome position start (bp) (GRCh38)</td>
</tr>
<tr>
<td></td>
<td>PolyPhen Prediction: probably damaging</td>
<td>Chromosome position end (bp) (GRCh38)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Strand (GRCh38)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Variant Alleles (GRCh38)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ensembl Gene ID (GRCh38)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Consequence to transcript (GRCh38)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Associated variation names (GRCh38)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Study External Reference (GRCh38)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Source name (GRCh38)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Associated gene with phenotype (GRCh38)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phenotype description (GRCh38)</td>
</tr>
</tbody>
</table>

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 powerful link between biological databases and microarray data analysis.

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 [09TB-BAR to BCCTB bioinformatics portal]; Office of Science of the U.S. Department of Energy [DE-AC02–05CH11231 to Phystozone]; 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 ciDB]; 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 Yoshhide 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.).

For funding for open access charge: King Abdulaziz University.

Conflict of interest statement. None declared.

REFERENCES


