Original article

Ensembl BioMarts: a hub for data retrieval across taxonomic space

Rhoda J. Kinsella^{1,*}, Andreas Kähäri¹, Syed Haider², Jorge Zamora¹, Glenn Proctor¹, Giulietta Spudich¹, Jeff Almeida-King¹, Daniel Staines¹, Paul Derwent¹, Arnaud Kerhornou¹, Paul Kersey¹ and Paul Flicek^{1,*}

¹European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD and ²Department of Computer Science and Technology, Computer Laboratory, University of Cambridge, 15 JJ Thomson Avenue, Cambridge CB3 0FD, UK

*Corresponding author. Rhoda J. Kinsella. Tel: +44 (0)1223 492608; Fax: +44 (0)1223 494484; Email: rhoda@ebi.ac.uk, helpdesk@ensembl.org

Correspondence may also be addressed to Paul Flicek. Tel: +44 (0)1223 429581; Fax: +44 (0)1223 494484; Email: flicek@ebi.ac.uk Present address: Jorge Zamora, Structural Computational Biology Group, Spanish National Cancer Research Centre, C/ Melchor Fernández Almagro, 3, 28029, Madrid, Spain

Submitted 20 April 2011; Revised 12 June 2011; Accepted 16 June 2011

For a number of years the BioMart data warehousing system has proven to be a valuable resource for scientists seeking a fast and versatile means of accessing the growing volume of genomic data provided by the Ensembl project. The launch of the Ensembl Genomes project in 2009 complemented the Ensembl project by utilizing the same visualization, interactive and programming tools to provide users with a means for accessing genome data from a further five domains: protists, bacteria, metazoa, plants and fungi. The Ensembl and Ensembl Genomes BioMarts provide a point of access to the high-quality gene annotation, variation data, functional and regulatory annotation and evolutionary relationships from genomes spanning the taxonomic space. This article aims to give a comprehensive overview of the Ensembl and Ensembl Genomes BioMarts as well as some useful examples and a description of current data content and future objectives. **Database URLs:** http://www.ensembl.org/biomart/martview/; http://protists.ensembl.org/biomart/martview/; http://fungi.ensembl.org/biomart/martview/

Project description

The Ensembl project (http://www.ensembl.org) was launched in 2000 and is a joint effort by the European Bioinformatics Institute (EBI) and the Wellcome Trust Sanger Institute (WTSI). Ensembl aims to provide high-quality genomic resources including gene annotations, multiple sequence alignments, whole-genome variation data and other information valuable for reuse by the community in a wide variety of research contexts (1).

As of release 61 (February 2011), 56 species are supported in Ensembl. The project focuses its support on chordate species and particularly on human genome resources and those of key model organisms such as mouse, rat and zebrafish. Ensembl also includes three non-chordate species because of their historical use as models for basic biological process. Four of the 56 supported species are in a pre-release state and can be viewed at http://pre.ensembl. org. The remaining 52 species all include comprehensive, evidence-based gene annotations and assignments of gene homology relationships. A smaller number of species include additional genomic data resources, largely chosen as a result of data availability and collaboration with species-specific or targeted resources. For example, Ensembl variation data resources include those in dbSNP (2) as well as variation data created by the project in the context of genome analysis (3). Close collaboration with other projects at the EBI including InterPro (4), the Database of Genomic Variants archive (DGVa) (5) and HGNC (6) ensures that Ensembl resources are integrated and available through other important bioinformatics

© The Author(s) 2011. Published by Oxford University Press.

This is Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http:// creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited. Page 1 of 9

resources. Recently somatic mutation data from the Catalogue of Somatic Mutations in Cancer (COSMIC) (7) has been incorporated into the Ensembl variation database.

The Ensembl Genomes project (http://www.ensemblgen omes.org) is comprised of separate websites for five distinct domains of life: bacteria, fungi, protists, plants and invertebrate metazoa (8). This project utilizes the Ensembl tools to provide genome-centric resources for species spanning the taxonomic space. Since the project launch in 2009, this portal has increased the number of genomes it represents from 122 species (bacteria, metazoa and protists) to 313 species (Ensembl Genomes release 8) of non-vertebrate genomes. For many species, the annotation is produced through collaborative efforts with scientific communities specializing in a particular domain, supplemented by the import of other publicly available information, while data from other important species is imported from various public repositories.

Ensembl and Ensembl Genomes are totally open projects and encourage others to incorporate the Ensembl code into their projects as well as provide specific tools for comprehensive data analysis and mining of the Ensembl data resources. In addition to long standing data resources such as the Ensembl gene sets (9) and gene trees (10), Ensembl provides other resources such as up-to-date microarray annotations (11). Widely used tools include the Variant Effect Predictor (VEP) (12) and the Ensembl API (13). The Ensembl genome browser at http://www.ensembl.org (14) provides a comprehensive visualization for accessing and using Ensembl data. The Ensembl BioMart (15,24) provides a final method for data access and querying data. Since the formative years of the Ensembl project, the BioMart data management system has played an important part in

providing access for the scientific community to the growing volume of genome data. Each of the five Ensembl Genomes portals also contains a BioMart for optimized querying of the data.

Data content

The Ensembl BioMarts are created using the database schemas and data generated by the various components of the Ensembl project. The Ensembl BioMarts are comprised of seven databases (three hidden and four visible). The four visible databases on the BioMart interface are: Ensembl Genes, Ensembl Variation, Ensembl Regulation and Vega. The three hidden BioMart databases contain supporting information for the visible databases including sequence data, ontology data and miscellaneous genomic features such as Encyclopedia of DNA Elements (ENCODE) (16) and karyotype data. The data in these three databases are accessed via the visible BioMart databases on the interface. Additional databases are integrated from the PRIDE (17) and Reactome (18,22) projects using the BioMart database federation technology. The gene-centric Ensembl Genes database as of Ensembl release 61 contains 52 fully supported species, the Ensembl Variation database contains variation-centric data for 18 species, the Ensembl Regulation feature-set-centric database contains data for three species and the Vega database contains manually annotated gene-centric data for three species (Table 1).

The Ensembl Genomes BioMarts are created using the BioMart database schemas generated by the Ensembl project and these are adapted to suit the specific requirements for each of the domains. A gene-centric database is

Data set	Description of data content
Ensembl Genes 61	Genes from 52 species with annotated external references, protein domains, multi species comparison (orthologs, possible orthologs and paralogs), variation (germline and somatic), regulation (probe set mapping for microarray platforms), gene ontology, expression (GNF/Atlas) and transcript splicing event data
Ensembl Variation 61	Variation data for 18 species including human somatic mutation data from COSMIC (7), human structural variation, human phenotype, Genome Wide Association Studies (GWAS) and variation set data. Strain specific data is available for certain other species.
Ensembl Regulation 61	Regulation data for human, mouse and <i>Drosophila melanogaster</i> (annotated, regulatory and external features)
Vega 41	Manually curated genes for human, mouse and zebrafish by the HAVANA group at WTSI and displayed in the VEGA database (21)
Reactome	Manually curated and peer-reviewed pathways from the BioMart (22) at http://www.reactome.org/ cgi-bin/mart
PRIDE (EBI UK)	Proteomics data from the PRIDE PRoteomics IDEntifications (17) BioMart database at http://www.ebi.ac.uk/pride/prideMart.do

Table 1. Summary of data available at the Ensembl BioMart as of Ensembl release 61

Data set	Description of data content
Ensembl Bacteria 8	249 genomes across 10 different clades (Gene database)
Ensembl Protists 8	11 species including <i>Plasmodium falciparum, Plasmodium knowlesi, Plasmodium vivax</i> and three oomycete genomes (Gene database for all species and Variation database for one species)
Ensembl Fungi 8	13 species, including eight Aspergillus species, Neosartorya fischeri, Puccinia graminis f. sp. Tritici, Saccharomyces cerevisiae, Schizosaccharomyces pombe (Gene database for all species and Variation database for one species)
Ensembl Metazoa 8	30 species, including 12 Drosphila, five Caenorhabditis, <i>Aedes aegypti</i> and <i>Apis mellifera</i> (Gene data- base for all species and Variation database for two species)
Ensembl Plants 8	10 species, including Arabidopsis lyrata, Arabidopsis thaliana, Brachypodium distachyon, Oryza sativa, Oryza sativa indica group and Zea mays (Gene database for all species and Variation database for four species)

Table 2. Summary of data available at the Ensembl Genomes BioMarts as of Ensembl Genomes release 8

Table	3.	Summary	of	sources	of	help	and	documentation	at	Ensembl
Table		Juinnary	01	Jources	01	ncip	anu	uocumentation	uι	LIISCHIDI

Information resource	URL or Email address
Ensembl frequently asked questions	http://www.ensembl.org/Help/Faq
BioMart frequently asked questions	http://www.biomart.org/faqs.html
Tutorials	http://www.ensembl.org/info/website/tutorials
YouTube videos	http://www.youtube.com/user/EnsemblHelpdesk
Ensembl news containing information about updates to mart databases	http://www.ensembl.org/info/website/news
Ensembl Blog	http://www.ensembl.info
Ensembl archives containing archived BioMart databases	http://www.ensembl.org/info/website/archives
Ensembl helpdesk mailing list	helpdesk@ensembl.org
Ensembl Genomes helpdesk mailing list	helpdesk@ensemblgenomes.org
Ensembl Genomes portal website containing project information	http://www.ensemblgenomes.org

available for each of the five domains and a variationcentric database is available for Protists, Fungi, Metazoa and Plants (Table 2).

The Ensembl BioMart tables are made available for download from the FTP site (ftp://ftp.ensembl.org/pub) for each release (e.g. Ensembl Genes 61 BioMart database available from ftp://ftp.ensembl.org/pub/release-61/mysgl/ ensembl_mart_61). Users can access the BioMarts by web interface, BioMart API, biomaRt package from bioconductor (19), SOAP based and RESTful webservices and by publicly available MySQL server offering direct access to the BioMart databases (http://www.ensembl.org/info/data/ mysql.html). Help and documentation details are summarized in Table 3. The Ensembl and Ensembl Genomes BioMarts are also displayed on the main BioMart central portal http://www.biomart.org. Three Ensembl mirrors have been created to improve the website performance for users around the globe. These mirrors, located on the west and east coasts of the USA (http://uswest.ensembl.org, http://useast.ensembl.org) and in Asia (http://asia.ensembl. org) also contain the Ensembl BioMarts to facilitate more effective data access.

Query examples

To demonstrate the utility of the Ensembl and Ensembl Genomes BioMarts we present several biologically relevant queries that can be performed using available tools and interfaces.

Query #1: The G-protein coupled receptor domain (GPCR) has the InterPro ID of IPR000276. Find the human protein-coding genes in Ensembl that code for this domain, and investigate whether any of them are detectable with the Affy HuGene 1_0 st v1 array.

Database: Data sets	Attributes	
Ensembl Genes 61: Homo sapiens	Gene type: protein_coding	Ensembl Gene ID
genes (GRCh37.p2)	Limit to genes with these family or domain IDs: IPR000276	Associated Gene Name
		Affy HuGene 1_0 st v1

The GPCR genes make up a large protein family that covers a wide range of functions. A scientist may already

C <mark>l</mark> Ensembl blast/	BLAT BioMart Tools Downloads Help &	Documentation Blog	Mirrors
🧿 New 📓 Count 📓 Results			🖕 URL 🛐 XML 🛃 Peri 🛞 Help
Dataset 777 / 53630 Genes	Export all results to	File	: TSV : Unique results only og Go
Homo sapiens genes (GRCh37.p2)	Email notification to		
Filters	The second se	and the second second second	
Gene type : protein coding	View	20 ÷ rows as	HTML 🗧 🗹 Unique results only
Interpro ID(s) [e.g.	Ensembl Gene ID Associated Gene Name	Affy HuGene 1 0 et v1	
IPR007087]: [ID-list specified]	ENSG00000127530 OR7C1	8034890	
Attributes	ENSG00000188269 OR7A5	8034892	
Freesel Ones ID	ENSG00000127515 OR7A10	8034897	
Ensembl Gene ID	ENSG00000172148 OR7A2P	8034899	
Associated Gene Name	ENSG00000185385 OR7A17	8034901	
Affy HuGene 1_0 st v1	ENSG00000127529 OR7C2	8026388	
Bibli Caller	ENSG0000094661 OR111	8026405	
	ENSG00000171942 OR10H2	8026483	
Dataset	ENSG00000171936 OR10H3	8026486	
[Nono Selected]	ENSG00000172519 OR10H5	8026488	
[None Selected]	ENSG00000186723 OR10H1	8035078	
	ENSG00000176231 OR10H4	8026494	
	ENSG00000127533 F2RL3	8026631	
	ENSG00000157219 HTR5A	8137517	
	ENSG00000181773 GPR3	7899343	
	ENSG00000169403 PTAFR	7914184	
	ENSG00000116329 OPRD1	7899528	
	ENSG00000230178 OR4F3	8148962	
	ENSG00000230178 OR4F3	7896744	
	ENSG00000230178 OR4F3	8110672	

Figure 1. There are 777 Ensembl protein coding genes that code for the GPCR domain with InterPro ID (IPR000276) and that are detectable with the Affy HuGene 1_0 st v1 array 25.

Chromosome	Sequence region start	Sequence region end	Structural Variation	Structural Variation Description	Source
Name	(bp)	(bp)	Name		Name
12	16265092	16446378	esv263	Redon 2006 "Global variation in copy number in the human genome" PMID:17122850 [remapped from build NCBI35]	DGVa:estd1

Figure 2. The esv263 structural variation from DGVa occurs between 16265092 and 16446378 bp on chromosome 12.

know the InterPro ID of the GPCR rhodopsin-like domain and wish to investigate how many Ensembl gene IDs code for this GPCR and whether these were detected using the Affy HuGene 1_0 st v1 array. To do this query, the user must select the protein_coding filter from the GENE filter section and filter with the known InterPro ID in the PROTEIN DOMAINS filter section. Attributes are selected from Features:GENE and Features:EXTERNAL sections (Figure 1).

Query #2: esv263 is the DGVa accession number of a structural variation from Redon *et al.* (20). What genomic region does this copy number variation span?

Attributes

Recent studies such as Redon *et al.* (20) have mapped copy number variations (CNV) in the human population. Redon *et al.* (20) studied 270 individuals from four populations whose DNA was screened for CNVs. Having read the article, a user may be interested in finding out more about a particular structural variation, such as the size of the genomic region that a particular structural variation spans (Figure 2). To do this query, the user must filter on the Structural Variation Name in the GENERAL STRUCTURAL VARIATION FILTERS and the attributes can be selected from the STRUCTURAL VARIATION attribute section.

Que	ry #	43: Ar	e there	any	genes	in	Ensem	bl	that	contain	som-
atic	mut	tation	s assoc	ated	with [.]	tum	nors in	th	e eye	e?	

			alle mataliene associated man tantors in the eyer					
Ensembl Variation 61: Homo sapiens	Limit to variants with these IDs:	Chromosome Name	Database: Data sets	Filters	Attributes			
Structural Variation	esv263	Sequence region start (bp) Ensembl Variat 61: Homo sap	Ensembl Variation 61: Homo sapiens	Phenotype: COSMIC: tumor_site:eye	Variation ID			
	start (pp) Sequence region end (bp) Structural Variation Name Structural Variation Description Source Name	Sequence region end (bp) Structural Variation Name Structural Variation Description Source Name	Somatic Variation (COSMIC 50)		Chromosome name Position on Chromosome (bp) Allele Phenotype description Associated gene Ensembl Gene ID			

Database: Data sets Filters

	Export all results to	~				
lataset 100 / 22586 SNPs	Export an results to	Fil	e	: TSV : O	Unique	results only
omo sapiens Somatic Variation COSMIC 50)	Email notification to					
Filters	a second and a second s	0.040		N. 19 10 10		
Phonotypo :	View	10	rows as HTML ↓ ✓	Unique results only		
COSMIC:tumour_site:eve	Variation ID Associated av	Chromosomo nomo	Besition on Chromosome (hp)	Dhapatupa description	Allala	Encombl Cone ID
Attributes	COSM476 BBAF	7	140453136	COSMIC tumour site over	Anele A/T	ENSG0000015776
Attributes	COSM26445 EGER	7	55249002	COSMIC:tumour_site:eve	СЛ	ENSG0000224057
Variation ID	COSM26445 EGER	7	55249002	COSMIC:tumour_site:eve	СЛ	ENSG0000146648
Associated gene	COSM28770 GNAO	9	80409487	COSMIC:tumour_site:eve	T/A	ENSG00000156052
Chromosome name	COSM28761 GNAO	9	80409487	COSMIC:tumour_site:eve	TTG/ATA	ENSG00000156052
Position on Chromosome (bp)	COSM28758 GNAQ	9	80409488	COSMIC:tumour site:eve	T/G	ENSG00000156052
Phenotype description	COSM28757 GNAQ	9	80409488	COSMIC:tumour site:eve	T/A	ENSG00000156052
Allele	COSM28760 GNAQ	9	80409488	COSMIC:tumour site:eye	T/C	ENSG00000156052
Ensembl Gene ID	COSM520 KRAS	12	25398284	COSMIC:tumour_site:eye	C/A	ENSG00000133703
	COSM521 KRAS	12	25398284	COSMIC:tumour_site:eye	С/Т	ENSG00000133703

Figure 3. Shows that there are 100 single nucleotide polymorphisms in the human somatic variation data set associated with tumors in the eye and the list of Ensembl gene IDs containing these variations can be downloaded for further study or one can click on an entry in the Ensembl Gene ID column on the interface which links to the main Ensembl website.

The COSMIC project focuses on somatic mutations relating to human cancers. A somatic variation data set has been incorporated into the Ensembl Variation BioMart database to give users access to this data. A scientist can select from a list of COSMIC phenotypes from the GENERAL VARIATION FILTERS filter section, choose a selection of useful attributes from the Variation:SEQUENCE VARIATION and Variation: GENE ASSOCIATED INFORMATION attribute sections and export their results in a selection of file formats (Figure 3).

Query #4: Find the HGNC symbols for a list of human variations.

Database: Data sets	Filters	Attributes
Ensembl Variation 61: <i>Homo sapiens</i> variation (dbSNP 132;ENSEMBL)	Limit to variants with these IDs dbSNP rs IDs: rs348, rs362, rs364, rs565, rs645	Variation ID Chromosome name Position on chromo- some (bp) Ensembl Gene ID
Ensembl Genes 61: Homo sapiens genes (GRCh37.p2)		HGNC ID HGNC symbol

This query requires that the user selects filters and attributes from the human data set in the Variation BioMart database as well as selecting attributes from the human data set in the Ensembl Genes BioMart database. The linking of two data sets is a useful feature of the BioMart technology and allows for complex cross database queries to be constructed. In this query the user may have a list of dbSNP IDs and would like to obtain a list of Ensembl gene IDs and their corresponding HGNC IDs that contain these variations (Figure 4). The user must first upload their list of dbSNP IDs to the GENERAL VARIATION FILTERS section and then select the required attributes from the Variation: SEQUENCE VARIATION and Variation:GENE ASSOCIATED INFORMATION attribute sections. Then select the second data set [*Homo sapiens* genes (GRCh37.p2) from Ensembl Genes mart] from the left sidebar on the screen. Then select the HGNC ID and HGNC symbol from the features: EXTERNAL attribute section.

Query #5: Find the genes from *Escherichia coli* strain K12 that are found within the region '360473–365601' and discover whether there are any orthologs in the related strains *E. coli O157:H7 EC4115* and *E. coli DH10B.*

Database: Data sets	Filters	Attributes
Ensembl Bacteria Bacterial Mart	Gene start (bp): 360473	Ensembl Gene ID
(Release 8):	Gene end (bp):	Ensembl Transcript ID
Escherichia coli	365601	Associated Gene Name
K12 genes		Escherichia coli DH10B Ensembl Gene ID
		Escherichia coli DH10B
		Chromosome Start (bp)
		Escherichia coli DH10B
		Chromosome End (bp)
		Escherichia coli O157:H7
		EC4115 Ensembl Gene ID
		Escherichia coli O157:H7
		EC4115 Chromosome
		Start (bp)
		Escherichia coli O157:H7
		EC4115 Chromosome
		End (bp)

C Ensembl blas	T/BLAT BioMar	t Tools Download	ds Help & Documentation Blo	og Mirrors 🛛 🛃	•	Login	· Register
New Count Results			술 URL 🛛 🔊 XML 🔄 Pe	rl 🕐 Help			
Dataset	Export all re	esults to	ile	‡) (TSV 🛟 🗆 U	Inique resu	its only 📀 Go	
(dbSNP 132:ENSEMBL)	Email notific	cation to					
Filters							
dbSNP rsID(s): [ID-list	View	1	o 🛟 rows as [html 🛟 🗹 l	Unique results only	у		
specified]	Variation ID	Chromosome name	Position on Chromosome (bp)	Ensembl Gene ID	HGNC ID(s)	HGNC symbol	
Attributes	rs348	13	32449504	ENSG00000229715	30486	EEF1DP3	
Variation ID	rs362	13	32477206	ENSG00000229715	30486	EEF1DP3	
Chromosome name	<u>rs364</u>	13	32479297	ENSG00000229715	30486	EEF1DP3	
Position on Chromosome	<u>rs565</u>	15	31231190	ENSG00000166912	25999	MTMR10	
(bp)	<u>rs565</u>	<u>15</u>	31231190	ENSG00000198690	<u>29170</u>	FAN1	
Ensembl Gene ID							
Dataset							
Homo sapiens genes (GRCh37.p2)							
Filters							
[None selected]							
Attributes							
HGNC ID(s) HGNC symbol							

Figure 4. Five dbSNP rs IDs were used to filter the human variation data set and Ensembl gene IDs containing these five variations were selected in the attributes. Then linking to the second data set, human gene data set from Ensembl Genes database, the HGNC ID and symbol were selected in the attribute section to retrieve the corresponding gene names from HGNC. They are FAN1, MTMR10 and EEF1DP3.

This query involves finding what *E. coli* genes lie in the given region and then discovering whether there are any orthologs in two related strains of *E. coli*. This is interesting as it may highlight bacterial genes that may have been acquired by some strains when compared to others and some genes may have been lost relative to other related strains (Figure 5). To do this query, add the gene start and end coordinates in the REGION filter section and then select the attributes from the Homologs:GENE and Homologs: ORTHOLOGS attribute sections.

Query #6: The three-gene APL1 locus encodes essential components of the mosquito immune defense against malaria parasites. Find the variations within the APL1A, APL1B and APL1C genes as well as the strain name, strain genotype, allele and biotype.

Database: Data sets	Filters	Attributes
Ensembl Metazoa Metazoa Variation Mart (release 8): Anopheles gambiae variations (AgamP3)	Ensembl Gene IDs: AGAP007035 AGAP007036 AGAP007033	Variation ID Chromosome name Position on Chromosome (bp) Allele dbSNP rsID Strain Name Strain Genotype Ensembl Gene ID Biotype

The Ensembl Metazoa Variation BioMart database consolidates single nucleotide polymorphisms from high-density, genome-wide mosquito SNP-genotyping array mapping and enables users to retrieve variations from the SNP-array identified through sequencing of two genetically diverged molecular forms of *A. gambiae*, Mopti (M) and Savanna (S) (23). This resource could help to analyze parasite susceptibility alleles from population subgroups. Query 6 shows how a user can obtain variation data for a particular gene or set of genes of interest (Figure 6). To do this query, the user must upload the gene IDs to the GENE ASSOCIATED VARIATION FILTERS section and then select the attributes of interest from the Variation: SEQUENCE VARIATION and Variation:GENE ASSOCIATED INFORMATION sections.

Query #7: Find the coding sequence for all human genes on chromosome 22 along with the gene name and gene start and end.

Database: Data sets	Filters	Attributes
Ensembl Gene 61: <i>Homo sapiens</i> genes (GRCh37.p2)	Chromosome 22	Coding sequence Ensembl Gene ID Associated Gene Name Associated Gene DB Gene Start (bp) Gene End (bp)

Ensembl Gene ID	Ensembl Transcript ID	Escherichia coli O157:H7 EC4115 Ensembl Gene ID	Escherichia coli O157:H7 EC4115 Chromosome Start (bp)	Escherichia coli O157:H7 EC4115 Chromosome End (bp)	Escherichia coli DH10B Ensembl Gene ID	Escherichia coli DH10B Chromosome Start (bp)	Escherichia coli DH10B Chromosome End (bp)	Associated Gene Name
EBESCG0000003288	EBESCT0000004025	EBESCG0000033465	421624	422235	EBESCG00000010829	1380009	1380620	lacA
EBESCG0000003288	EBESCT0000004024	EBESCG0000033465	421624	422235	EBESCG00000010829	1380009	<u>1380620</u>	lacA
EBESCG0000003288	EBESCT0000004023	EBESCG0000033465	421624	422235	EBESCG00000010829	1380009	1380620	lacA
EBESCG0000003288	EBESCT0000004026	EBESCG0000033465	421624	422235	EBESCG0000010829	1380009	1380620	lacA
EBESCG0000003502	EBESCT0000004294	EBESCG0000028996	422301	423554	EBESCG0000009230	1380686	1381939	lacY
EBESCG0000003502	EBESCT0000004295	EBESCG0000028996	422301	423554	EBESCG0000009230	1380686	1381939	lacY
EBESCG0000003502	EBESCT0000004296	EBESCG0000028996	422301	423554	EBESCG0000009230	1380686	1381939	lacY
EBESCG0000003502	EBESCT0000004293	EBESCG0000028996	422301	423554	EBESCG0000009230	1380686	1381939	lacY
EBESCG0000001573	EBESCT00000001917	EBESCG0000033401	423606	426680		3		lacZ
EBESCG0000001573	EBESCT00000001918	EBESCG0000033401	423606	426680				lacZ
EBESCG0000001573	EBESCT00000001919	EBESCG0000033401	423606	426680		8		lacZ
EBESCG0000001573	EBESCT00000001916	EBESCG0000033401	423606	426680				lacZ

Figure 5. The genes in the filtered region were lacA, lacY and lacZ and we can see that there are no orthologs for the lacZ gene in the *E. coli* DH10B strain.

<i>EnsemblMetazoa</i>							M •		e
Home							BLA	ST BloMart FTF	Docs & FAQ
🔊 New 📓 Count 📓 Results							👷 U	JRL 😰 XML 🔄	Perl 💿 Help
Dataset	Export all	results to	File		\$ TSV \$	🛛 🖂 Unique	e results only	@ Go	
Anopheles gambiae variations (AgamP3)	Email notif	lication to							
Filters	10					2			
Ensembl Gene ID(s) : [ID-list	View		20 C rows as HTML	÷ • •	Jnique resul	ts only			
specified]	Variation ID	Chromosome	Position on Chromosome (bp)	Allele	dbSNP rsID	Strain Name	Strain Genotype	Ensembl Gene ID	Biotype
Veriation ID	rs3536462	2L	41260908	T/G				AGAP007033	protein_coding
Variation ID	rs3536463	<u>2L</u>	41263234	G/-				AGAP007033	protein_coding
Chromosome name	rs3536463	<u>2L</u>	41263234	G/-				AGAP007035	protein_coding
Position on Chromosome (bp)	rs5303484	<u>2L</u>	41275027	T/C	ss252529588	pest	TIN	AGAP007036	protein_coding
Allele	rs5303484	<u>2L</u>	41275027	T/C	ss252529588	mopti	CIN	AGAP007036	protein_coding
dDSNP rsiD	rs5303987	<u>2L</u>	41274907	T/C		pest	TIN	AGAP007036	protein_coding
Strain Name	rs5303987	<u>2L</u>	41274907	T/C		mopti	CIN	AGAP007036	protein_coding
Strain Genotype	rs5304006	<u>2L</u>	41274898	C/G		pest	CIN	AGAP007036	protein_coding
Ensembl Gene ID	rs5304006	<u>2L</u>	41274898	C/G		mopti	GIN	AGAP007036	protein_coding
Biotype	rs5304478	<u>2L</u>	41274808	A/G	ss252529587	pest	AIN	AGAP007036	protein_coding
	rs5304478	<u>2L</u>	41274808	A/G	ss252529587	mopti	GIN	AGAP007036	protein_coding
	rs5304948	<u>2L</u>	<u>41274721</u>	G/A	ss252529586	pest	GIN	AGAP007036	protein_coding
Dataset	rs5304948	<u>2L</u>	41274721	G/A	ss252529586	mopti	AIN	AGAP007036	protein_coding
None Selected]	rs5305463	<u>2L</u>	41274606	C/T		pest	CIN	AGAP007036	protein_coding
	rs5305463	<u>2L</u>	41274606	C/T		mopti	TIN	AGAP007036	protein_coding
	rs5305823	<u>2L</u>	41274568	A/C	ss252529585	pest	AIN	AGAP007036	protein_coding
	rs5305823	21	41274568	A/C	<u>ss252529585</u>	mopti	CIN	AGAP007036	protein_coding
	rs5305875	<u>2L</u>	41274541	A/C	ss252529584	pest	AIN	AGAP007036	protein_coding
	rs5305875	<u>2L</u>	41274541	A/C	ss252529584	mopti	CIN	AGAP007036	protein_coding
	rs5359835	2L	41276180	T/C		pest	TIN	AGAP007036	protein coding

Figure 6. Having first retrieved the Ensembl gene IDs for the three APL1 genes, these are used to filter the *A. gambiae* data set. Fifty variations were retrieved that lie within the three genes of the APL1 locus.

The BioMart technology allows for the download of sequence information in a usable format. This is a powerful feature that allows users to retrieve flanking sequence, exon sequence, 3' and 5'-UTR, cDNA sequence, coding sequence and protein sequence. Query 7 illustrates how to retrieve coding sequences for all genes on chromosome 22 as well as obtaining information about the gene name and the location of the gene start and end (Figure 7). To do this query, select the chromosome from the REGION filter section and the attributes of interest from the Sequences: SEQUENCES and Sequence:Header Information attribute sections.

Discussion and future directions

The BioMart interface and querying platform provides the Ensembl and Ensembl Genomes projects with the necessary

tools to design BioMart databases from the various source databases produced by the project. The BioMart databases and accompanying interface provides users with a fast and flexible means of querying the customized sets of biological data using a wide range of querying methods. The BioMart software also allows federation to other databases of scientific interest so that cross querying can be accomplished. It also allows the Ensembl and Ensembl Genomes databases to be incorporated into other portals with ease such as www.biomart.org.

As scientific activity evolves and in an effort to provide the most useful resources for our users, both the Ensembl and Ensembl Genomes projects will incorporate data from additional species and additionally handle new types of data, which will be included in the project BioMarts. In the future, we plan to move both projects to the new BioMart 0.8 code (24) and incorporate the new interface into the main Ensembl website.

Clensembl Blast	/BLAT BioMart Tools Downloads	Help & Documentation Blog Mirrors	2 .
🧿 New 📓 Count 📓 Results		🖕 URL 🗗 XML 🔄 Perl 💿 Help	
Dataset 1225 / 53630 Genes	Export all results to	File 🗧 🗧 FASTA 🔅 🗆 Unique results only	🎯 Go
Homo sapiens genes (GRCh37.p2)	Email notification to		S
Filters	NP		
Chromosome: 22	View	10 ; rows as FASTA ; Unique results only	
Attributes	>ENSC00000099992 TBC1D10	A HCNC Symbol 30687979 30723035	
Ensembl Gene ID Coding sequence Associated Gene Name Associated Gene DB Gene Start (bp) Gene End (bp)	ATGGCGAAGAGCAACGGAGAGAAT GGAACCCGGGAGAGCCTGGCCAAC CTCGGGTTGGACTCGAGGCCCAAC ATCGTGGGCTCGCAGGGCGCCGAG CAGAGGAGTCCAACTGGCGGCGC CACAAAAAGATTCGTCTGCGGTGG TGGCAGTACCTGTCAGGAGCAAC CTGGACATGTCCCCTGGGGACCCC	IGGCCCCCCCCCCCGCGCCGGGAAAGCCTGTCG IGGCCCCGACGCCGCACCGACGAAGTCAGCTCT IGGCTCGCCGACGCCCCATCGACAAGTTCGGCTTC IGGCGCCGCTGGAGGAAGTACCCCTGGAGGGGCCTGAGG IATGCTCAACAACTGGGACAAATGGATGGCCAAGAAG ICAAAAGGCAACCCCTCCTCCTGCGGGGCCGTGCCT IGTGAAGTTACAGCAGAACCCTGGAAAGTTTGACCAG IAAGTGGCTGGACGTGATTGAGCGTGACCTGCACCGG	
Dataset	CAGTTCCCATTCCATGAGATGTTT	GTGTCCCGGGGGGGCCACGGCCAGCAGGACCTATTC	
[None Selected]	CGTOTGCTGAAGGCCTATGCTCATG CCATTGCCGTGAAGGCTTTGCTCATG CAGATCTGTGAGAAGTACCTGCCC GACCGGGAGATCCTTTTCTCGCTC CGAACCTTGCCCTGGAGCTCTGTC AAGATCATCTTCCGGGTGGGCTC GTCAAAGCCTGCCAGGGCCAGTAC ATCATGCAGGAGGCCTTTCTGGTC ATTGATGCAGGAGGCCCAGTCCATTCAC TGCCGCTCCCCGCCCAGGCGCGCG GGAGAGGCACTCGAGAAGCCC GGAGAGGCACTCGAGAAGCCC CGGAGAGGCACCTCCCACAGCAT CCTCAGGATTGGCCCCCAGGCCCC CAAGAGGTGAGAGACACCTCTCACG >ENSG0000063515 GSC2 EG ATGGCGGCAGCGGCTGGGGGGCGC TCCATCGAGCACATCCTCCCAG	TACCOGCCCCAAACCACCCCCCCCCCCCCCCCCCCCCCCC	

Figure 7. The ability to retrieve sequence information for genes of interest is a powerful feature of the BioMart tool. Here a user can download the coding sequence for all genes on chromosome 22 as well as additional information about each gene and this can be exported in a useful format.

Acknowledgements

The authors thank all the users of the Ensembl and Ensembl Genomes projects especially those who have provided us with feedback about the Ensembl BioMarts. The authors would also like to thank the members of the BioMart team at the Ontario Institute for Cancer Research (OICR), especially Dr Arek Kasprzyk, for providing sustained technical support and assistance over the years.

Funding

The Wellcome Trust provide majority funding for the Ensembl project (grant number WT062023) with additional support from the European Commission under SLING, grant agreement number 226073 (Integrating Activity) within Research Infrastructures of the FP7 Capacities Specific Programme; the UK Biotechnology and Biological Sciences Research Council (grant numbers BB/F019793/1, BB/I001077/1); the Bill and Melinda Gates Foundation;

and the European Molecular Biological Laboratory. Funding for open access charge: The Wellcome Trust.

Conflict of interest. None declared.

References

- 1. Flicek,P., Amode,M.R., Barrell,D. et al. (2011) Ensembl 2011. Nucleic Acids Res., **39**, D800–D806.
- Foelo, M.L. and Sherry, S.T (2007) NCBI dbSNP Database: content and searching. In: Weiner, M.P., Gabriel, S.B. and Stephens, J.C (eds), *Genetic Variation: A Laboratory Manual.* Cold Spring Harbour Laboratory Press, Cold Spring Harbour, NY, pp. 41–61.
- Chen,Y., Cunningham,F., Rios,D. et al. (2010) Ensembl variation resources. BMC Genomics, 11, 293.
- Hunter, S., Apweiler, R., Attwood, T.K. et al. (2009) InterPro: the integrative protein signature database. Nucleic Acids Res., 37, D211–D215.
- Church,D.M., Lappalainen,I., Sneddon,T.P. et al. (2010) Public data archives for genomic structural variation. Nat. Genet., 42, 813–814.

- Bruford,E.A., Lush,M.J., Wright,M.W. et al. (2008) The HGNC database in 2008: a resource for the human genome. *Nucleic Acids Res.*, 36, D445–D448.
- Forbes, S.A., Tang, G., Bindal, N. *et al.* (2010) COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. *Nucleic Acids Res.*, 38, D652–D657.
- 8. Kersey, P.J., Lawson, D., Birney, E. *et al.* (2010) Ensembl genomes: extending ensembl across the taxonomic space. *Nucleic Acids Res.*, **38**, D563–D569.
- 9. Curwen, V., Eyras, E., Andrews, T.D. et al. (2004) The Ensembl automatic gene annotation system. Genome Res., 14, 942–950.
- Vilella,A.J., Severin,J., Ureta-Vidal,A. *et al.* (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, **19**, 327–335.
- 11. Ballester, B., Johnson, N., Proctor, G. and Flicek, P. (2010) Consistent annotation of gene expression arrays. *BMC Genomics*, **11**, 294.
- 12. McLaren,W., Pritchard,B., Rios,D. *et al.* (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP effect predictor. *Bioinformatics*, **26**, 2069–2070.
- 13. Stabenau, A., McVicker, G., Melsopp, C. *et al.* (2004) The Ensembl core software libraries. *Genome Res.*, **14**, 929–933.
- 14. Parker, A., Bragin, E., Brent, S. *et al.* (2010) Using caching and optimization techniques to improve performance of the Ensembl website. *BMC Bioinformatics*, **11**, 239.
- 15. Smedley, D., Haider, S., Ballester, B. et al. (2009) BioMart biological queries made easy. BMC Genomics, 10, 22.

- Raney,B.J., Cline,M.S., Rosenbloom,K.R. et al. (2011) ENCODE whole-genome data in the UCSC genome browser (2011 update). Nucleic Acids Res., 39, D871–D875.
- Vizcaíno, J.A., Reisinger, F., Côté, R. and Martens, L (2011) PRIDE and "Database on Demand" as valuable tools for computational proteomics. *Meth. Mol. Biol.*, 696, 93–105.
- Croft, D., O'Kelly, G., Wu, G. et al. (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, 39, D691–D697.
- Durinck,S., Spellman,P.T., Birney,E. and Huber,W. (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.*, 4, 1184–1191.
- 20. Redon,R., Ishikawa,S., Fitch,K.R. *et al.* (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.
- Wilming,L.G., Gilbert,J.G., Howe,K. et al. (2008) The vertebrate genome annotation (Vega) database. Nucleic Acids Res., 36, D753–D760.
- 22. Shepherd, R., Forbes, S.A., Beare, D. *et al.* (2011) The Reactome BioMart. *Database*.
- Neafsey,D.E., Lawniczak,M.K., Park,D.J. *et al.* (2010) SNP genotyping defines complex gene-flow boundaries among African malaria vector mosquitoes. *Science*, **330**, 514–517. Erratum in: Science. 330, 1477.
- 24. Zhang,J., Haider,S., Guberman,J. *et al.* (2011) BioMart: A data federation framework for large collaborative projects. *Database*.