

# BioMart

## Data integration in four easy steps

Arek Kasprzyk  
European Bioinformatics Institute  
22 July 2006

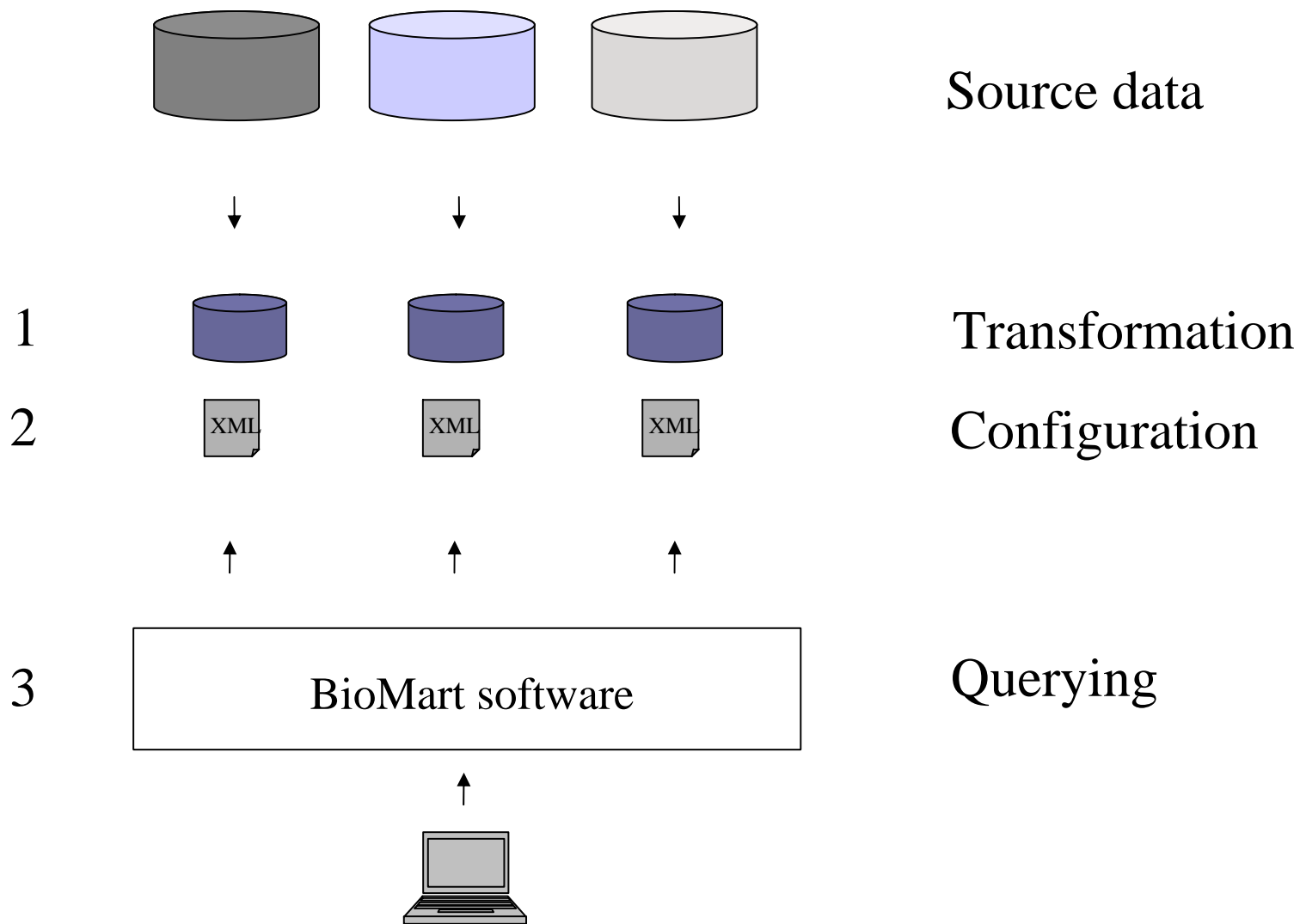
# BioMart

- A joint project
  - European Bioinformatics Institute (EBI)
  - Cold Spring Harbor Laboratory (CSHL)
- Funding
  - Wellcome Trust
  - European Commission
  - NIH

# Synopsis

- Higher level data management system
  - Data mining type access to descriptive data
  - Query optimization
  - Data federation
  - Meta data support

# BioMart



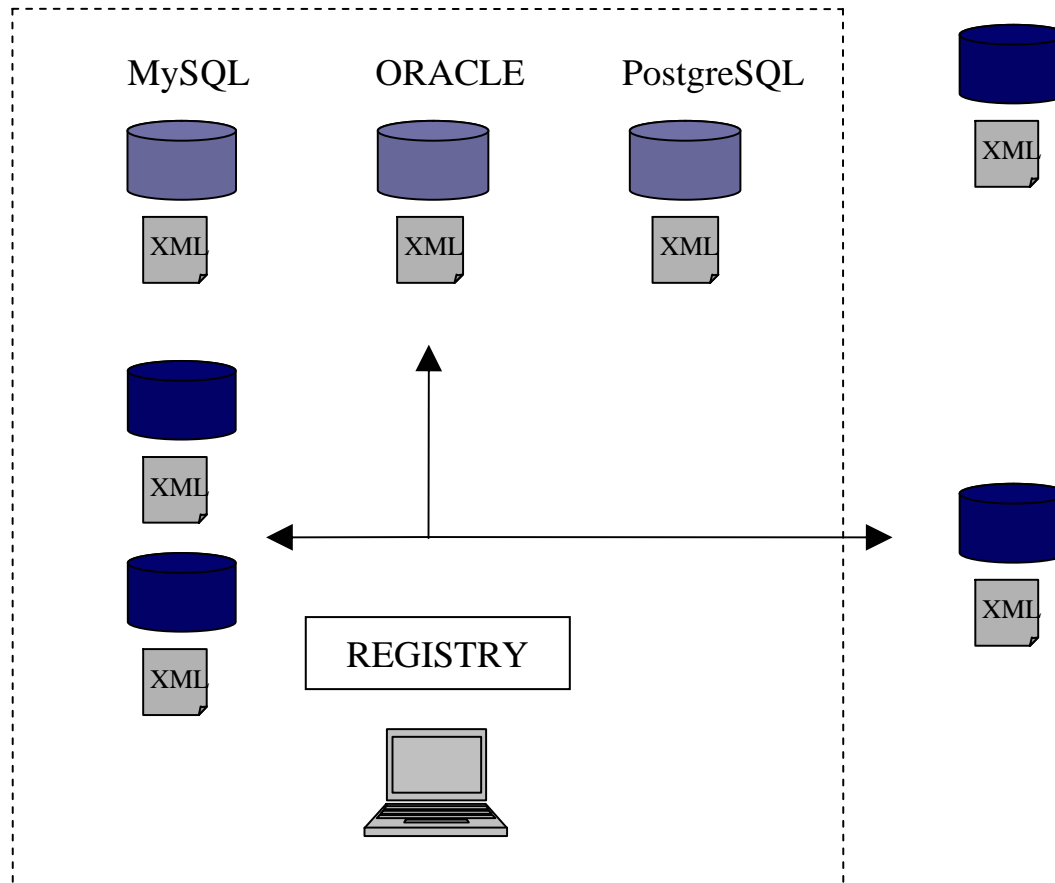




# Programmatic access

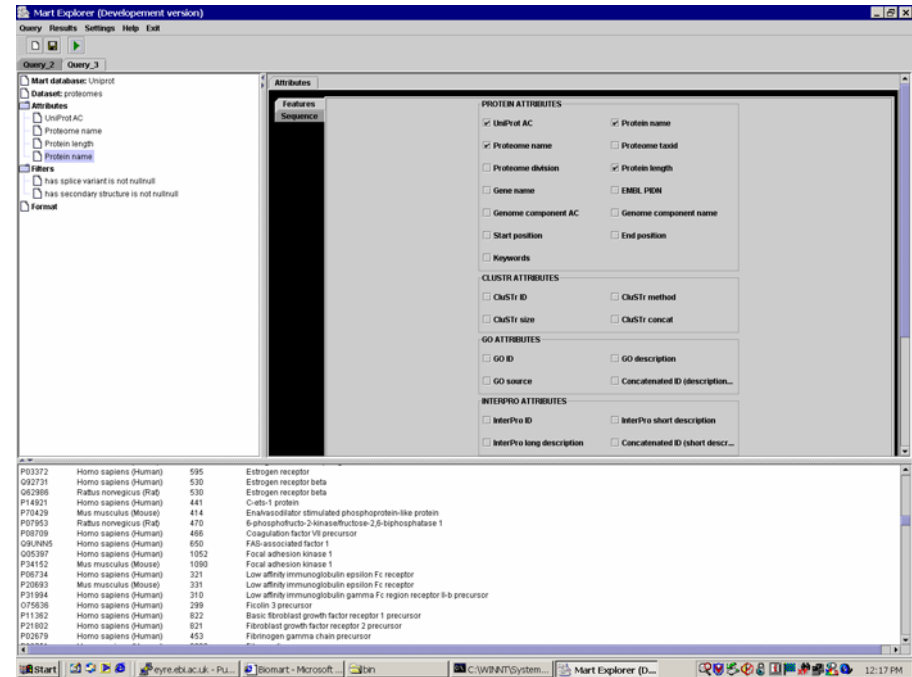
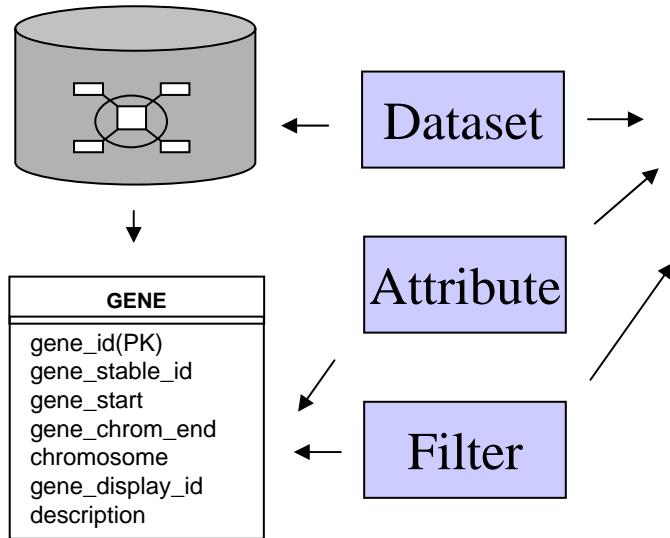
- APIs
  - Perl (biomart-plib)
  - Java (martj)
  - R (biomaRt)
- Web service

# Data federation

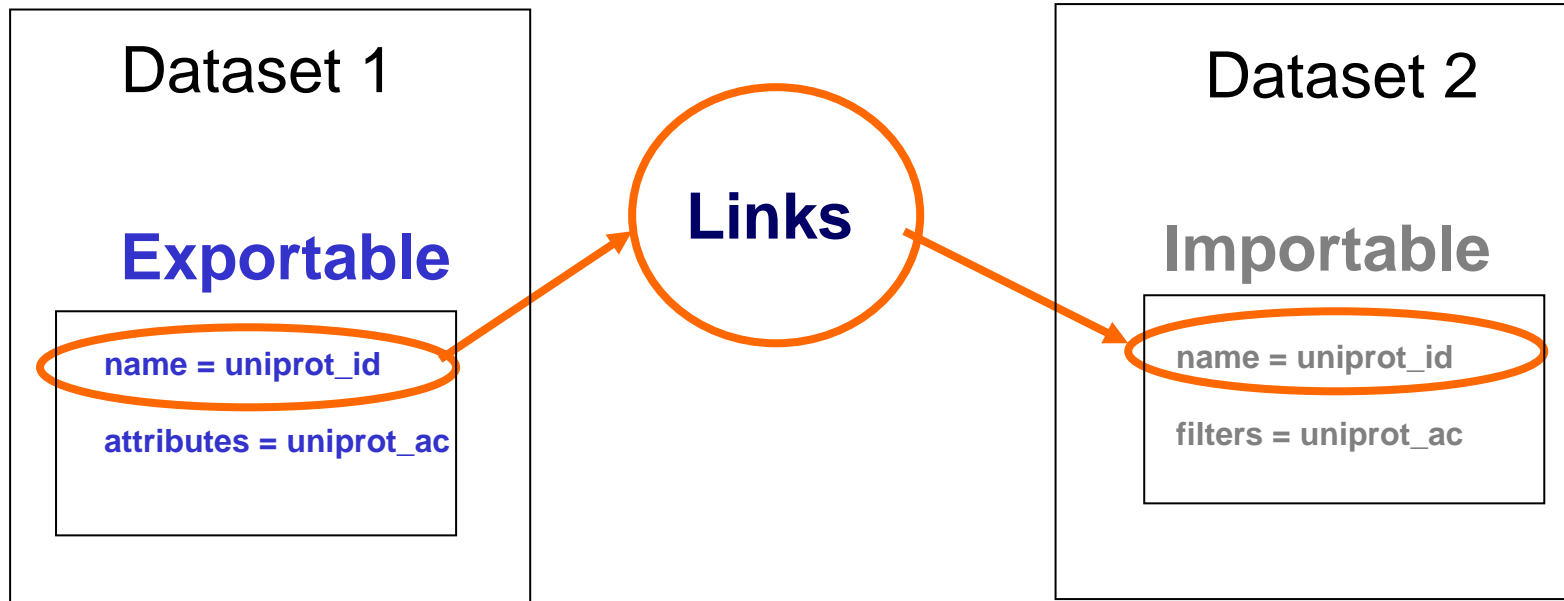




# Dataset, Attribute and Filter



# Joining two datasets



# Dataset linking



New Query | Save query

Showing 10 entries out of 0 total as HTML

Export all entries to Text in browser Go

Ensembl Gene ID	Ensembl Transcript ID	UniProt AC
<a href="#">ENSG00000101307</a>	<a href="#">ENST00000381605</a>	<a href="#">O00241</a>
<a href="#">ENSG00000101307</a>	<a href="#">ENST00000342311</a>	<a href="#">O00241</a>
<a href="#">ENSG00000110711</a>	<a href="#">ENST00000279146</a>	<a href="#">O00170</a>
<a href="#">ENSG00000116977</a>	<a href="#">ENST00000341872</a>	<a href="#">O00214</a>
<a href="#">ENSG00000116977</a>	<a href="#">ENST00000323938</a>	<a href="#">O00214</a>
<a href="#">ENSG00000116977</a>	<a href="#">ENST00000366584</a>	<a href="#">O00214</a>
<a href="#">ENSG00000135750</a>	<a href="#">ENST00000366621</a>	<a href="#">O00180</a>
<a href="#">ENSG00000135750</a>	<a href="#">ENST00000258230</a>	<a href="#">O00180</a>
<a href="#">ENSG00000132842</a>	<a href="#">ENST00000380357</a>	<a href="#">O00203</a>
<a href="#">ENSG00000132842</a>	<a href="#">ENST00000255194</a>	<a href="#">O00203</a>

Results

**Dataset:**  
 Homo sapiens genes (NCBI36)  
*31,737 entries total*

- Attributes**
- Features
  - Ensembl Gene ID
  - Ensembl Transcript ID
- Filters**  
[None configured]

**Dataset:**  
 Uniprot proteomes  
*960,763 entries total*

- Attributes**
- Features
  - UniProt AC
- Filters**  
[None configured]

# Third party software

# Ensembl



Search all Ensembl:

Ensembl v38 - Apr 2006

[Help](#)

## Use Ensembl to...

- Run a BLAST search
- Search Ensembl
- Data mining [BioMart]
- Upload your own data
- Export data
- Download data

## Docs and downloads

- Information
- What's New
- About Ensembl
- Ensembl data
- Software

## Other links

- Home
- Sitemap
- Vega
- Pre Ensembl
- View previous release of page in Archive!
- Stable Archive! link for this page
- Archive! sites
- Trace server



***Drosophila melanogaster***

## What's New in Ensembl 38

- ▶ **New Human assembly and genebuild** (*Homo sapiens*)
- ▶ **New Mouse assembly and genebuild** (*Mus musculus*)
- ▶ **Import of FlyBase 4.2.1** (*Drosophila melanogaster*)
- ▶ **Xenopus assembly updated to 4.1** (*Xenopus tropicalis*)
- ▶ **New navigation options in ContigView** (all species)

[More news...](#)

## About Ensembl

Ensembl is a joint project between [EMBL - EBI](#) and the [Sanger Institute](#) to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes. Ensembl is primarily funded by the [Wellcome Trust](#).

This site provides [free access](#) to all the data and software from the Ensembl project. Click on a species name to browse the data.

Access to all the data produced by the project, and to the software used to analyse and present it, is provided free and without constraints. Some data and software may be subject to [third-party constraints](#).

For all enquiries, please [contact the Ensembl HelpDesk](#) ([helpdesk@ensembl.org](mailto:helpdesk@ensembl.org)).

## Other sites using the Ensembl system

- ▶ [EBI Genome Reviews](#) database - mainly archaea and bacteria.

## Mammalian genomes

- Homo sapiens**  
**UPDATED!** NCBI 3.6 | Vega
- Pan troglodytes**  
PanTro 1.0
- Macaca mulatta**  
MMUL 0.1 | *pre!*
- Mus musculus**  
**UPDATED!** NCBI m35 | Vega
- Rattus norvegicus**  
RGSC 3.4
- Oryctolagus cuniculus**  
RABBIT
- Canis familiaris**  
CanFam 1.0 | Vega | *pre!*
- Bos taurus**  
Btau 2.0
- Dasyus novemcinctus**  
ARMA
- Loxodonta africana**  
BROAD E1
- Echinops telfairi**  
TENREC
- Monodelphis domestica**  
MonDom 2.0

## Other species

- Gallus gallus**  
WASHUC 1
- Xenopus tropicalis**  
**UPDATED!** JGI 4.1
- Danio rerio**  
Zv5 | Vega | *pre!*
- Fugu rubripes**  
FUGU 4.0
- Tetraodon nigroviridis**  
TETRAODON 7
- Gasterosteus aculeatus**  
**NEW!** BROAD S1
- Ciona intestinalis**  
JGI2
- Ciona savignyi**  
CSAV 2.0
- Drosophila melanogaster**  
**UPDATED!** BDGP 42
- Anopheles gambiae**  
AgamP3
- Aedes aegypti**  
AEDES 1
- Apis mellifera**  
Ame1 2.0
- Caenorhabditis elegans**  
WS 1.50
- Saccharomyces cerevisiae**  
SGD 1

# GMOD

[Mailing lists](#) | [Bug Reports](#) | [Feature Requests](#) | [Forum](#) | [Publications](#) | [Meetings](#) | [SourceForge site](#) | [Old GMOD site](#) | [RSS feed](#) | [Progress Reports](#)

[Downloads](#) | [Support](#) | [Handbooks](#) | [Development](#) | [Architecture Working Group](#) | [Software Components \(aka The Matrix\)](#)



## GMOD

Generic Software Components for Model Organism Databases

### User login

Username:

Password:

- [Create new account](#)
- [Request new password](#)

### GMOD News

- [BioMart 0.4 Released](#)
- [genome GFF to BioMart tool \(alpha\)](#)
- [The NeuroScholar System.](#)
- [blastGraphic-0.02 release](#)
- [GBrowse 1.64 release](#)

[more](#)

### New GMOD.org nodes

- [Feb 27 progress report](#)
- [TODO for a gmod 0.1 release](#)
- [Can't run Gbrowse on Windows System....](#)
- [Full text searching postgresSQL GFF adaptor](#)
- [Feb 20 progress report](#)

[more](#)

### Home

## The Generic Model Organism Database Project

Submitted by mod on Thu, 2005-11-17 15:40.

### About GMOD

Submitted by mod on Wed, 2005-11-23 18:01.

The Generic Model Organism Database (GMOD) Project is a largely open source project to develop a complete set of software for creating and administering a model organism database. Components of this project include genome visualization and editing tools, literature curation tools, a robust database schema, biological ontology tools, and a set of standard operating procedures. This project is funded by the NIH and the USDA Agricultural Research Service, with participation from members of several database projects, including WormBase, FlyBase, Mouse Genome Informatics, Gramene, the Rat Genome Database, TAIR, EcoCyc, and the Saccharomyces Genome Database. For more information, please see <http://www.gmod.org/> or contact [Scott Cain <cain@csih.edu>](mailto:Scott.Cain@csih.edu).

» [login](#) or [register](#) to post comments | [read more](#)

### Frequent GMOD Downloads

Submitted by mod on Wed, 2005-11-23 17:18.

- GMOD Core (Chado database and tools)
  - [gmod-core](#)  
the schema and tools to install it. [[Download](#)]
  - [XORT](#)  
a tool for loading and dumping chado-xml [[Download](#)]
  - [GMODTools](#)  
[this needs a description, and a package for that matter]
- Genome Editing and Visualization
  - [Apollo](#)  
a Java application for viewing and editing genome annotations [[Download](#)]
  - [GBrowse](#)  
a CGI application for displaying genome annotations [[Download](#)]
- Comparative Genomics
  - [Synbrowse](#)

Welcome to Bioconductor

<http://www.bioconductor.org/> Google

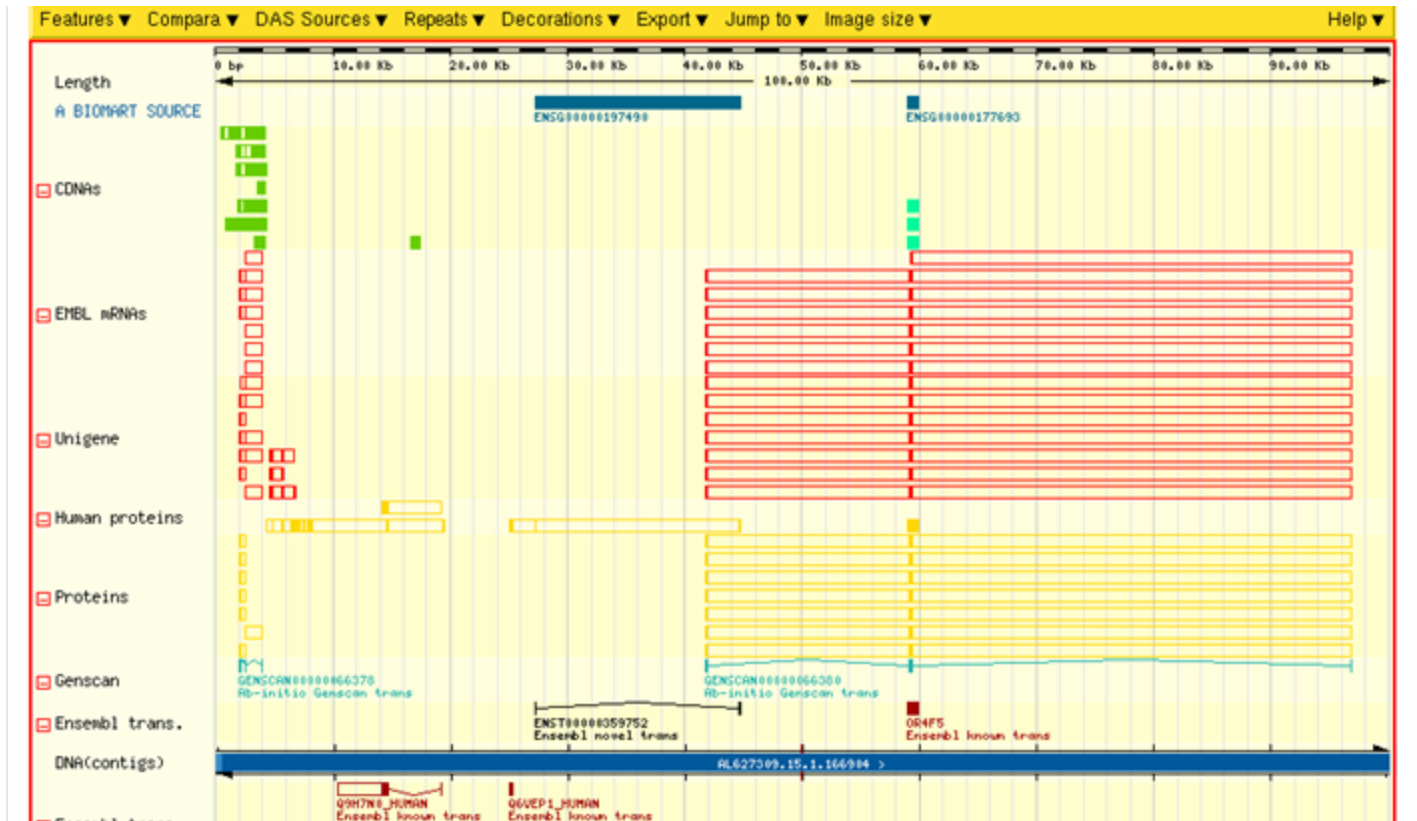
Apple .Mac Amazon eBay Yahoo! News

**BioConductor**: open source software for bioinformatics

<b>About</b>	<a href="#">arrayQuality</a>	Assessing array quality on spotted arrays	1.0.9
<a href="#">Bioconductor</a>	<a href="#">bjm</a>	Bayesian Interval Mapping Diagnostics	1.01-1
<a href="#">Main Page</a>	<a href="#">Biobase</a>	Biobase: Base functions for Bioconductor	1.5.6
<a href="#">What is Bioconductor?</a>	<a href="#">bioDist</a>	Different distance measures	0.6.4
<a href="#">Screenshots</a>	<a href="#">biomaRt</a>	Interface to biomart	1.0.0
<a href="#">Citing Bioconductor</a>	<a href="#">Biostrings</a>	String objects rerepresenting biological sequences	1.1.1
<a href="#">Bioconductor FAQ</a>	<a href="#">ChromoViz</a>	Multimodal visualization of gene expression	1.0
<a href="#">Developers</a>	<a href="#">CoCiteStats</a>	Different test statistics based on co-citation.	0.5.3
<a href="#">Mirrors</a>	<a href="#">convert</a>	Convert Microarray Data Objects	1.1.9
<a href="#">Acknowledgements</a>	<a href="#">ctc</a>	Cluster and Tree Conversion.	1.2.7
<a href="#">What's New?</a>	<a href="#">daMA</a>	Efficient design and analysis of factorial two-colour microarray data	1.0.1
<b>Software</b>	<a href="#">DEDS</a>	Differential Expression via Distance Summary for Microarray Data	1.0.3
<a href="#">How To</a>	<a href="#">DNAcopy</a>	DNA copy number data analysis	1.1.1
<a href="#">Release 1.5 Packages</a>	<a href="#">DynDoc</a>	Dynamic document tools	1.5.3
<a href="#">Developmental Packages</a>	<a href="#">EBarrays</a>	Empirical Bayes for Microarrays	1.0-19
<a href="#">MetaData</a>	<a href="#">ecolink</a>	Meta-data and tools for E. coli	1.0-1
<a href="#">Experimental Data</a>	<a href="#">edd</a>	expression density diagnostics	1.5.0
<a href="#">Excel Software</a>	<a href="#">exprDB</a>	A BerkeleyDB-based expression database	0.1.4
<b>Documentation</b>	<a href="#">exprExternal</a>	Implementation of exprSet using externalVectors	1.0.0
<a href="#">Vignettes</a>	<a href="#">externalVector</a>	Vector objects for R with external storage	1.0.14
<a href="#">Working Paper Series</a>	<a href="#">factDesign</a>	Factorial designed microarray experiment analysis	1.1.4
<a href="#">Compendiums</a>	<a href="#">germa</a>	Background Adjustment Using Sequence Information	1.1.3
<a href="#">Short Courses</a>	<a href="#">genArise</a>	Microarray Analysis tool	1.2.1
<a href="#">Lab Materials</a>	<a href="#">genefilter</a>	Genefilter: filter genes	1.5.0
<a href="#">Research Talks</a>	<a href="#">geneplotter</a>	Graphics related functions for Bioconductor	1.5.4
<a href="#">Publications</a>	<a href="#">GeneR</a>	R for genes and sequences analysis	0.1-4
<a href="#">R Documentation</a>	<a href="#">GeneSpring</a>	GeneSpring R Integration Functions	1.0.3
<b>Services</b>	<a href="#">GeneTraffic</a>	GeneTraffic R Integration Functions	1.0.6
<a href="#">Annotation</a>	<a href="#">GeneTS</a>	Genetic Time Series and Graphical Models	2.3
<a href="#">Workshops</a>	<a href="#">GLAD</a>	Gain and Loss Analysis of DNA	1.0.1
<b>Project</b>			
<a href="#">Mailing Lists</a>			

Go to "<http://www.bioconductor.org/main.html>" in another frame

# Distributed Annotation System





Scufi Workbench v0.1beta12, built Mon Dec 06 16:25:44 GMT 2004

Tools and Workflow Invocation

## Taverna Workbench

Advanced model explorer

Workflow: Metadata for 'GetDiseaseGeneIDs'

Workflow object: Retries, Delay, Backoff, Threads, Critical

Processors:

- GetUniqueHomolog: 0 0 1 1
- GetMouseGenes: 0 0 1 1
- GetHSGenes: 0 0 1 1
- GetRatGenes: 0 0 1 1
- CreateFasta: 0 0 1 1
- hsSeq: 0 0 1 1
- mmSeq: 0 0 1 1
- rmSeq: 0 0 1 1
- fasta: 0 0 1 1
- GetDiseaseGeneIDs: 0 0 1 1
- chr\_name\_filter: 0 0 1 1
- sequenceexport: 0 0 1 1
- FlattenImageList: 0 0 1 5
- secret: 0 0 1 5
- emma: 0 0 1 5
- plot: 0 0 1 5
- Data links: 0 0 1 5
- GetUniqueHomolog HSOOut->Get-

Available services

Search list: Watch loads

- GetDomainsFromGMMEvaluate
- GetAccFromRetiredGi
- ProteinReportSetDescription
- GetFastaKeyedList
- RedundantGroupKeyedList
- GetFastaFromRedundantGroupIDKeyedList
- Biomart ensemble\_mart\_22\_1@martdb.ebi.ac.uk
- frutripes\_gene\_ensembl
- hsapiens\_gene\_ensembl
- chriggsae\_gene\_ensembl
- mmvegicus\_gene\_ensembl
- dreio\_gene\_ensembl
- ggallus\_gene\_ensembl
- celegans\_gene\_ensembl
- mmorvegicus\_gene\_ensembl
- agambiae\_gene\_ensembl
- dmegm\_gene\_ensembl
- ggallus\_gene\_ensembl
- chriggsae\_gene\_ensembl

Workflow diagram

Save as: Show types, Bound ports, Vertical, Fit to window

Enactor invocation

Save as XML, Save to disk, Save to disk as website

Status: Results, Process report

MMusIDs, HSapiDs, RNorIDs, outputPlot

Resources on martdb.ebi.ac.uk, 4 instances

Biomart	Dataset Name	Proc
Biomart	mmusculus_gene_ensembl	GetM
Biomart	hsapiens_gene_ensembl	GetH
Biomart	mmorvegicus_gene_ensembl	GetF
Biomart	hsapiens_gene_ensembl	GetC

Resources on industry.ebi.ac.uk, 3 instances

Soaplab	Service rooted at
Soaplab	Service rooted at/soap/soaplab

Resource usage report

This display shows the various external resources used by the enactment engine. Services are categorized by the name of the instance of each service shown to the right.

Configuring query for GetHSGenes

Attributes, Filters

Features: Structures, Sequences, SNPs

Sequences

Type of sequence to export: REGION, GENE, PROTEIN

Sequence export options

Type of sequence to fetch: Genes

Desired sequence options

5' upstream only

Extents

5' flank: 200, 3' flank: 1000

Sequence glyph

# Galaxy

Galaxy
Info: [contact](#) | [help](#) Account: [create](#) | [login](#)

**Get Data**

- [Upload File](#) from your computer
- [UCSC Main](#) table browser
- [UCSC Test](#) table browser
- [UCSC Archaea](#) table browser
- [BioMart](#) Central server
- [EncodeDB](#) at NHGRI

**Get ENCODE Data**

**ENCODE Tools**

**Edit Queries**

[Filter, Sort, Join and Compare](#)

[Convert Formats](#)

[Fetch Sequences and Alignments](#)

[Alignment Viewers](#)

[Get Genomic Scores](#)

[Operate on Genomic Intervals](#)

[Operate on Genomic Intervals \(bx\)](#)

[Statistics](#)

[Graph Data](#)

**EMBOSS**

**PHYLIP**

**PAML**

**Welcome to Galaxy!** Galaxy is a web-based data analysis framework that allows you to collect and work with biological data from many different sources.

---

Watch Galaxy screencasts to see how it works. Just click below...

Galaxy 2.2 has a slight problem when viewed with Microsoft Internet Explorer. It works best with Firefox or Safari. We are working on resolving this issue.

- [Galaxy wiki page and documentation](#)
- [People and Technology behind Galaxy](#)
- Report problems to Galaxy mailing list: [galaxy-user@bx.psu.edu](mailto:galaxy-user@bx.psu.edu)

---

The Galaxy team is a part of the [Center for Comparative Genomics and Bioinformatics](#) at Penn State.

refresh | collapse all | delete checked

**3: Concatenate queries on data 2, data 2** ✓

27,246 lines, format: text, database: ?

Info:

[display](#) | [save](#) | [edit attributes](#) | [delete](#)

Gene name

```
ai2b
nad2
nad5
ai2a
ai3
```

**2: Genes** ✓

**1: Homo sapiens genes (NCBI36)** ✓

Tools
History options...

# Examples



new START FILTER OUTPUT export

bioMart

count help

Summary

start

- Dataset: All Populations

DATA SET 1

FILTERS

POPULATION BASED FILTERS Han Chinese from Beijing, China

ALLELE FREQUENCY FILTER  $\geq 1$

Monomorphic SNPs Monomorphic SNPs

SNPs found in Ensembl v34 - Oct 2005

Limit to SNPs with these Exons - non synonymous

REGION

Chromosome

From position

To position

GENE FILTERS

Gene IDs

List of Genes

**e!Ensembl Multi MartView**

**Use Ensembl to...**

- Run a BLAST search
- Search Ensembl
- Data mining [BioMart]
- Upload your own data
- Export data
- Download data

**Docs and downloads**

- Information
- What's New
- About Ensembl
- Ensembl data
- Software

**Other links**

- Home
- Sitemap
- Vega
- Pre Ensembl
- View previous release of page in Archive!
- Stable Archive! link for this page
- Archive! sites
- Trace server

Search all Ensembl:

new START FILTER OUTPUT export

bioMart

count help

Summary

start

- Dataset: Homo sapiens genes
- 34294 Entries Total

filter

- Chromosome name: 1
- Start: 1
- End: 100000000
- Disease genes: Only
- Type: protein\_coding
- Transmembrane domains: Only
- Coding: Only
- 10 Entries pass Filters

output

- Not yet initialised

DATA SET 1

REGION:

Chromosome 1

Base pair

Start 1

End 100000000

Band

Start p36.33

End p36.33

Marker

Start

End

Encode type Manual Picks

Encode region 11:115962315:116462315

In encode region

Only

Excluded

GENE:

Disease genes

Only

Excluded

ID list limit

Ensembl Transcript ID(s)



# Uniprot, MSD, ArrayExpress Proteomic, structure, expression

EMBL-EBI European Bioinformatics Institute

Get Nucleotide sequences for [ ] Go Site search [ ] Go

EBI Database Query

EBI Home About EBI Groups Services Toolbox Databases Downloads Submissions BioMart

new START FILTER OUTPUT export

bio·mart count help Summary

DATASET 1

SPECIES:

Species archaza

Protein Name Aeropyrum pernix

REGION (Single species only):

Genome Component Chromosome

Start (bp)

End (bp)

EXTERNAL IDENTIFIERS:

Entries with INTERPRO ID(s) Only Excluded

Limit to proteins UNIPROT ID(s) Choose File no file selected

GENE ONTOLOGY (GO)

Molecular function

Gene selection page

http://www.ebi.ac.uk/acidw/OWJsessionid=\$session?queryType=21Gene&nextPage=GeneList

EMBL-EBI European Bioinformatics Institute

ArrayExpress Data Warehouse Prototype

Back to Query Page Display expression of selected gene(s)

Gene name	Synonym	EMBL	ErsGene	GO	InterPro	LocusLink	refseq	UniGene	UniProt
<input type="checkbox"/> BANF1	BCRG1 BAF	A033692	ENSG00000175334	GO:0005615	IPR04122	8819	NM_003860	Hs433738	Q78531
<input type="checkbox"/> ABCG2	BCRP ABCP BCRP1	AF083771	ENSG00000198777	GO:0008519 GO:0005215 GO:0006810 GO:0042493 GO:0019621 GO:0045989 GO:005524	IPR06192	5429	NM_004827	Hs194720	Q8UNQ0
<input type="checkbox"/> BCR	BCR1	M15015		GO:0004668 GO:0004674 GO:007169 GO:0050396					P11274
<input type="checkbox"/> Bcr			ENSG00000169364	GO:0000074 GO:0007420 GO:0005036	IPR01848	113	NM_021574	Hs234799	Q9BLW5

Back to Query Page Display expression of selected gene(s)

Send any comments to the ArrayExpress Team Help

# Model organism databases

[Home](#)
[Genome](#)
[Blast / Blat](#)
[WormMart](#)
[Batch Sequences](#)
[Markers](#)
[Genetic Maps](#)
[Submit](#)
[Searches](#)
[Site Map](#)

Find:  Any Gene

[new](#)
[START](#)
[FILTER](#)
[OUTPUT](#)
[export](#)

**GRAMENE Multi** Find anything

[Search](#)
[Genomes](#)
[Download](#)
[Resources](#)
[About](#)
[Help](#)

[new](#)
[START](#)
[FILTER](#)
[OUTPUT](#)
[export](#)

bio:mart

[new](#)
[next](#)

refresh [Help Desk](#) [Tutorial](#)

Select the dataset for this query

An Online Informatics Resource for Dictyostelium

dictyBase

[Genome Browser](#)
[BLAST](#)
[Colleagues](#)
[Stock Center](#)
[Research Tools](#)
[Help](#)
[Links](#)

Search dictyBase:  use \* as a wildcard character

Include dicty Newsletter in Search

[new](#)
[START](#)
[FILTER](#)
[OUTPUT](#)
[export](#)

dictyBase

[count](#)
[help](#)

**Summary**

[start](#)

- Dataset: Genes
- 13622 Entries Total

[filter](#)

- Chromosome: 1
- Primary Features: Only
- Curated Model: Only
- 409 Entries pass Filters

[output](#)

- Not yet initialized

[new](#)
[START](#)
[FILTER](#)
[OUTPUT](#)
[export](#)

bio:mart

[count](#)
[help](#)

**Summary**

[start](#)

- Dataset: Drosophila pseudoobscura
- 29708 Entries Total

[filter](#)

- Has CDS, Has Cross genome match dmclchr, Has Exon, Has Gene, Has Gene, Has Gene genewise, Has Gene genscan, Has Gene twinscan, Has Golden path region, Has Intron, Has tRNA, : Only
- Has CDS, Has Cross genome match dmclchr, : Only
- 37488 Entries pass Filters

[output](#)

- Not yet initialized

[Drosophila](#)
[About](#)
[BLAST](#)
[BioMart](#)
[Maps](#)
[Data](#)

**DATASET 1**

**FILTER BY FEATURE:**

Feature Type(s)

exon, gene, gene\_genewise, gene\_genscan, gene\_twinscan

**FILTER GENOME REGION:**

Chromosome

Chromosome: 2, 3, 4, group1, 4, group?, 4, group, 1

Region Start >=, Region End <=

**FILTER REGIONS WITH FEATURES: (include, exclude)**

Has tRNA

Has tRNA genewise, Has tRNA genscan, Has tRNA twinscan, Has Match blastn vs tLDBT\_dust

Only  Excluded

**FILTER REGIONS WITH FEATURES: (include, exclude)**

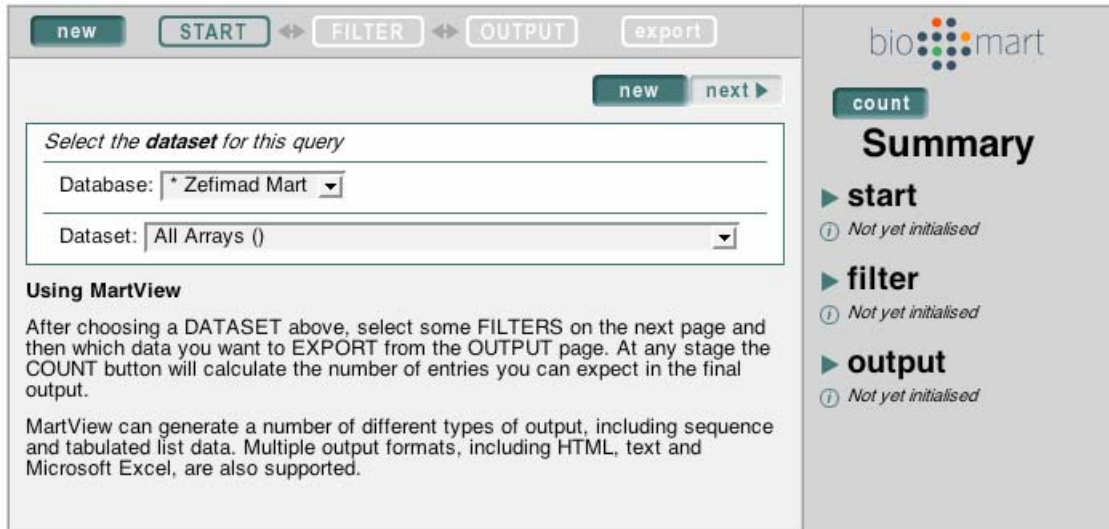
Has Cross genome match dmclchr

Has CDS, Has Exon, Has Gene genewise, Has Gene genscan

Only  Excluded

- Genes
- Expression
- Phenotypes
- Variations
- Literature
- Ontologies
- Sequence

# Zebra Fish models for human development and disease



new START FILTER OUTPUT export

new next

Select the **dataset** for this query

Database: \* Zefimad Mart

Dataset: All Arrays ()

**Using MartView**

After choosing a DATASET above, select some FILTERS on the next page and then which data you want to EXPORT from the OUTPUT page. At any stage the COUNT button will calculate the number of entries you can expect in the final output.

MartView can generate a number of different types of output, including sequence and tabulated list data. Multiple output formats, including HTML, text and Microsoft Excel, are also supported.

bio:mart

count

**Summary**

- ▶ start  
ⓘ Not yet initialised
- ▶ filter  
ⓘ Not yet initialised
- ▶ output  
ⓘ Not yet initialised



"ZF-MODELS - Zebrafish Models for Human Development and Disease" is an Integrated Project funded by the European Commission as part of its 6th Framework Programme.

# Central Server

[Home](#)
[Documentation](#)
[News](#)
[Contact](#)
[Credits](#)
[Query](#)

BioMart

new
START
FILTER
OUTPUT
export

new
next ▶

Select the **dataset** for this query

Database:  ENSEMBL 37 (SANGER)  
                    SNP 37 (SANGER)

Dataset:

**Using MartView**  
 After choosing a dataset, you can apply FILTERS on the next page and then which data you want to EXPORT from the OUTPUT page. At any stage the COUNT button will calculate the number of entries you can expect in the final output.

MartView can generate a number of different types of output, including sequence and tabulated list data. Multiple output formats, including HTML, text and Microsoft Excel, are also supported.  
  
 For a bookmarkable version of this page, click [here](#)

bioMart
count
help

### Summary

- ▶ **start**  
ⓘ Not yet initialised
- ▶ **filter**  
ⓘ Not yet initialised
- ▶ **output**  
ⓘ Not yet initialised



Behind closed doors ;)

# Target SNP selection for the study of one autoimmune disease, type 1 diabetes (T1D), and infectious diseases, malaria and dengue

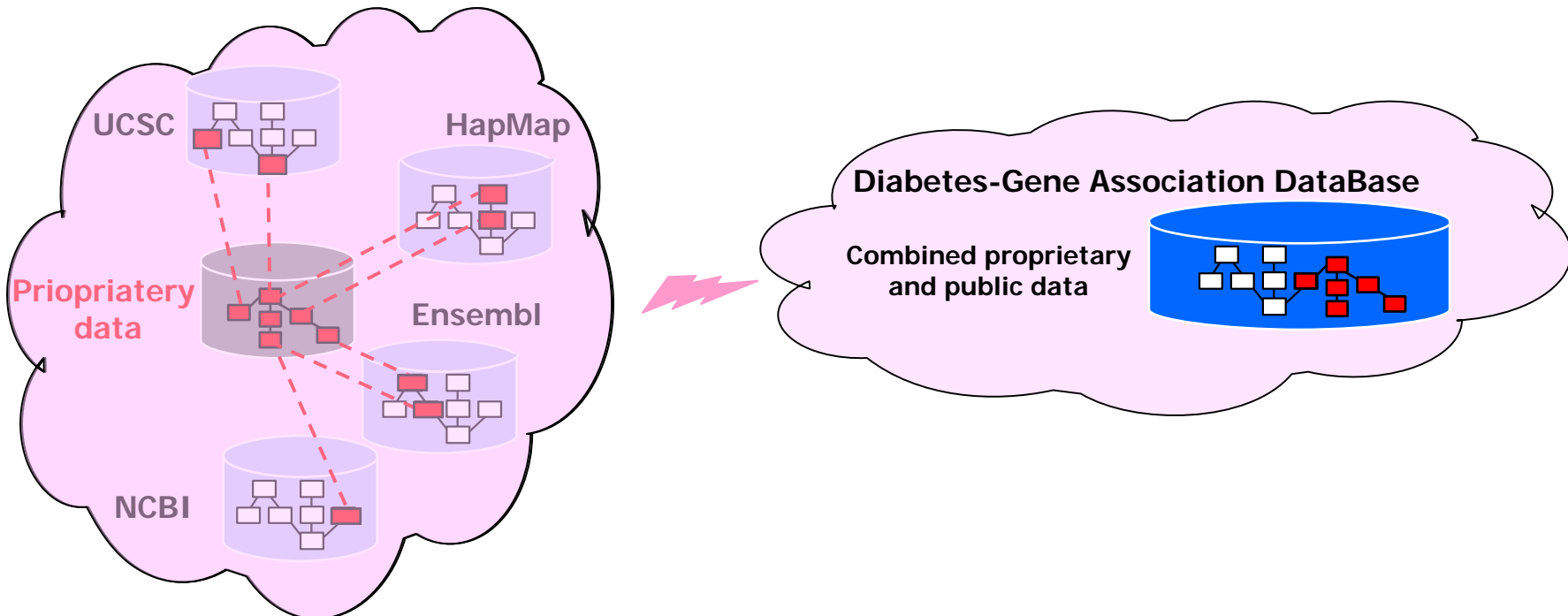
Laboratory of Genetics of Infectious and Autoimmune Diseases



Name	FragmentPosition	Alleles	strand
SNP1	AL1392581659852	T/A	1
SNP2	NT_25698	2569873	C/T -1
SNP3	chr13	1125698	C/G 1



**Data conversion and integration**



new    START    FILTER    OUTPUT    export

Features    SNPs    **Structures**    Sequences

REGION:

**Chromosome attributes**

Chromosome name     Start position (bp)

Strand

**Contig attributes**

Contig     Contig location (bp)

SNP:

**SNP attributes**

GMIA ID     RefSNP ID

HapMap ID     TSC ID

Hgbase ID     CNG ID

GMIA\_Alias     Genbase ID

TFBS

**TFBS\_UCSC**

Chromosome     Chromosome Start (pb)

Chromosome Stop (pb)     Strand

TFBS Name     Id

Species     Factor

score

GENE SNP ATTRIBUTES: (from dbsnp)

**For Ensembl Genes**

Gene ID     Transcript ID

Transcript Strand     Description

External Gene ID     External Gene DB

Family ID     Description

Location in Gene (coding etc)     Peptide Shift

Synonymous Status     Transcript Location (bp)

Ensembl     cDNA location (bp)

GENE SNP ATTRIBUTES: (LOCAL Computation)

**For Ensembl**

Gene ID     Transcript ID

Transcript strand     Description

External name     External db

Family name     Family description

Location in gene (coding etc)     Peptide

Synonymous status     Transcript location (bp)

CDS location (bp)     cDNA location (bp)

Peptide location (aa)

**For Vega**

Gene ID     Transcript ID

Transcript strand     Description

External name     External db

Family name     Family description

Location in gene (coding etc)     Peptide

Synonymous status     Transcript location (bp)

CDS location (bp)     cDNA location (bp)

Peptide location (aa)

**For Refseq**

Gene ID     Transcript ID

Transcript strand     Description

LocusLink ID     OMIM ID

Family name     Location in gene (coding etc)

Peptide     Synonymous status

Transcript location (bp)     CDS location (bp)

cDNA location (bp)     Peptide location (aa)

**For Acembly**

Gene ID     Transcript ID

Transcript strand     Location in gene (coding etc)

Peptide     Synonymous status

Transcript location (bp)     CDS location (bp)

cDNA location (bp)     Peptide location (aa)

Summary

- ▶ start
  - Dataset: GMIA SNPs (GMIA\_SNP\_database\_B35)
  - ① 2824 Entries Total
- ▶ filter
  - Coding: Only
  - ① 132 Entries pass Filters
- ▶ output
  - Gene structure
  - ① 132 Results in Output

Ensembl (dbSNP)

Ensembl

Vega

RefSeq

Acembly

Genetics of Infectious and Autoimmune Diseases,  
 Pasteur Institute, INSERM U730, Paris, France.

Genome Location

Links to databases

Overlaps with TFBS

Location  
 +  
 predicted functional  
 role

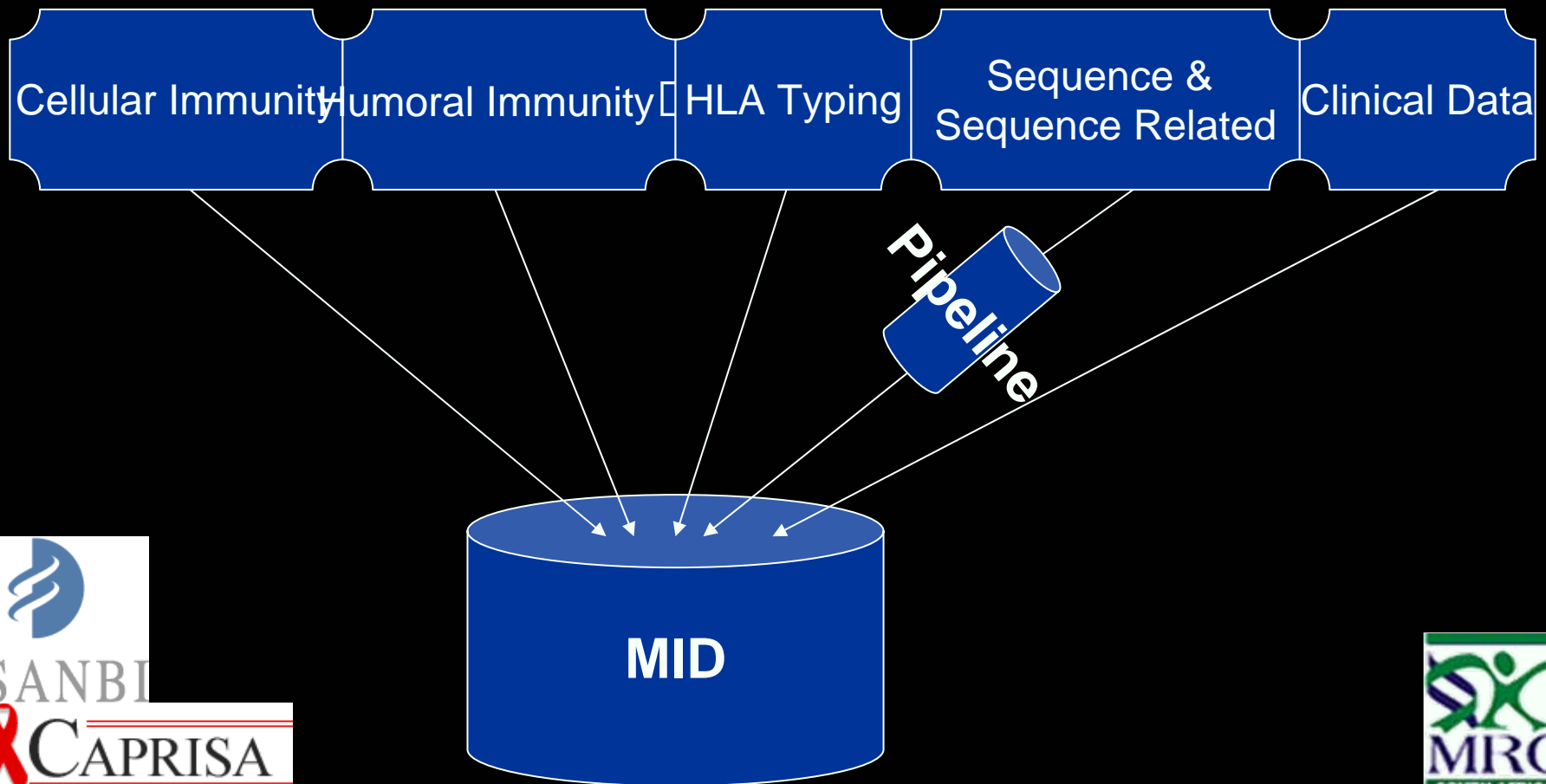
Output format :

# Using the Molecular Integration Database to Answer CAPRISA's Questions

Research that contributes to  
understanding HIV pathogenesis  
and epidemiology as well as  
HIV/AIDS treatment and  
prevention



# How is the MID populated?



# Caprisa



CENTRE FOR THE AIDS PROGRAMME OF RESEARCH IN SOUTH AFRICA

[new](#) [START](#) ↔ [FILTER](#) ↔ [OUTPUT](#) [export](#)

[new](#) [next ▶](#)


Select the **dataset** for this query

Dataset:  ▼

**Using MartView**

After choosing a DATASET above, select some FILTERS on the next page and then which data you want to EXPORT from the OUTPUT page. At any stage the COUNT button will calculate the number of entries you can expect in the final output.

MartView can generate a number of different types of output, including sequence and tabulated list data. Multiple output formats, including HTML, text and Microsoft Excel, are also supported.

bio  mart

[count](#) [help](#)

**Summary**

- ▶ **start**  
① *Not yet initialised*
- ▶ **filter**  
① *Not yet initialised*
- ▶ **output**  
① *Not yet initialised*



# What role for 'Omics' ?

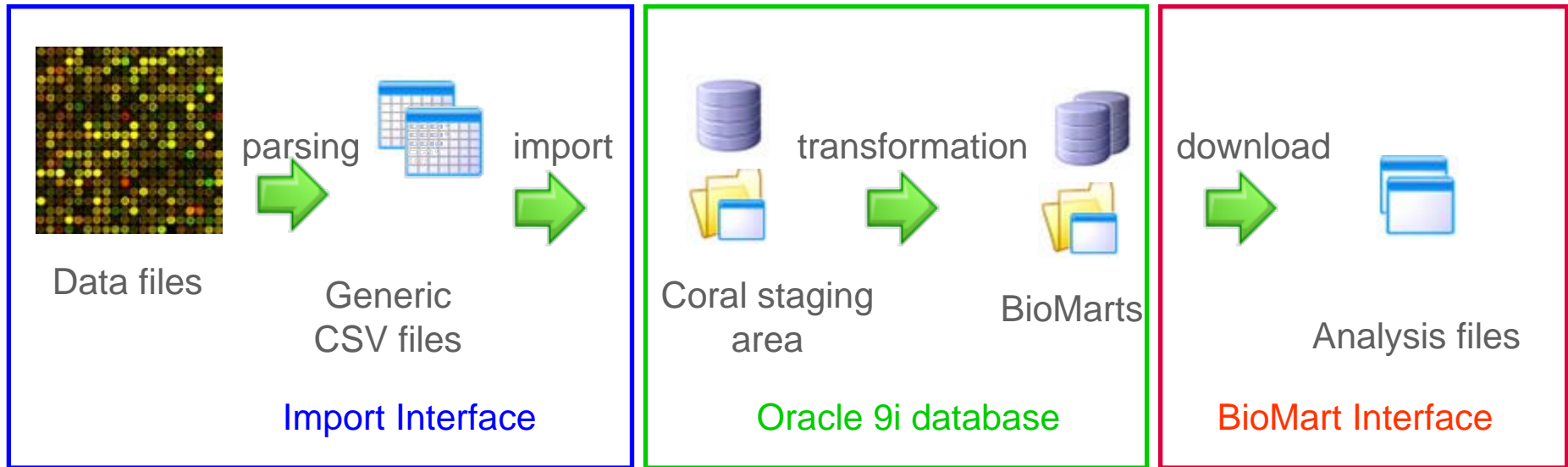
- **Human study to evaluate Omics** in assessing safety indicators
- **Study of skin inflammation** in response to detergent
- Skin samples taken and analyzed with multiple Omics techniques.
  - Blood
  - Skin biopsy
  - Microdialysis



Unilever



# System Data Flow



- Requires an extensible file and metadata management system for omics data
- **Oracle 9i** database used for staging area and BioMarts
- **Database indexes files** on a separate file system

# Adding Annotation

EXPRESSION:

MULTI SPECIES COMPARISONS:

Homologous Monodelphs Genes  Only  Excluded

PROTEIN:

with PROBLE IDs  Only  Excluded

Limit to genes with these family or domain IDs:  
Interpro IDs

Transmembrane domains  Only  Excluded

Signal domains  Only  Excluded

SNP:

Coding  Only  Excluded

Frameshifting SNPs  Only  Excluded

SNPs with HGBASE IDs  Only  Excluded

Associated with validated SNPs  Only  Excluded

Associated with SNPs with frequency data  Only  Excluded

Select the second dataset  
(NB) Summary counts are unavailable once you add a second dataset to the query. Data export is still possible from the next page

Second Dataset:

Link Datasets Via:

DATASET2

Microarray Experiment Details

experiment details

Experiment ID

Protocol ID

Experiment Sample Details

sample details

Participant ID

Sample ID

Sample type

Ensembl Mart

- **Query Ensembl** for details of genes measured or identified in experiments e.g. GeneSpring Annotation
- **For example**, we can link to Ensembl from Microarray Experiments by Gene ID

Link on Entrez gene id

Microarray Mart

# Four easy(?) steps

# Step 1

## Transformation

# Step 2

## Configuration

# Step 3

Query



# Web service

```

<Query virtualSchemaName = "default" count = "0" >
  <Dataset name = "hsapiens_gene_ensembl">
    <Attribute name = "gene_stable_id" />
    <Filter name = "chr_name" value = "22"/>
  </Dataset>
  <Dataset name = "uniprot">
    <Attribute name = "accession" />
    <Filter name = "pfam" value = "only"/>
  </Dataset>
</Query>

```



# API

```
my $initializer = BioMart::Initializer->new('registryFile'=>$confFile);  
my $registry = $initializer->getRegistry();  
$registry->configure();
```

```
$query->addAttribute('hsapiens_gene_ensembl','ensembl_gene_id');  
$query->addFilter('hsapiens_gene_ensembl','chromosome_name',['1']);  
$query->addAttribute('uniprot','accession',);  
$query->addFilter('uniprot','chromosome_name',['1']);  
$query->formatter('HTML');
```

```
my $runner = BioMart::QueryRunner->new();  
$runner->execute($query);  
$runner->printResults();
```

# Step 4

Ask for a pay rise :)

# Summary

- A generic data management system
- Provides building blocks for designing your own 'tailor-made' data management
  - A set of easily configurable user interfaces
  - Distributed Data federation
  - Query optimization
- Easy to install and manage
  - A project for bioinformatics students
- Open source software.
  - No restrictions for academics or commercial users

# Credits

- BioMart
  - Syed Haider
  - Richard Holland
  - Damian Smedley
  - Gudmundur Thorisson
- Contributors
  - Steffen Durinck (NCI, NIH)
  - Eric Just (Northwestern University)
  - Don Gilbert (Indiana University)
  - Darin London (Duke University)
  - Will Spooner (CSHL)
  - Benoit Ballester (Universite de la Mediterranee)
  - James Smith (Ensembl)
  - Arne Stabenau (Ensembl)
  - Andreas Kahari (Ensembl)
  - Craig Melsopp (Ensembl)
  - Katerina Tzouvara (EBI)
  - Paul Donlon (Unilever)