BioMart

Data integration in four easy steps

Arek Kasprzyk
European Bioinformatics Institute
22 July 2006

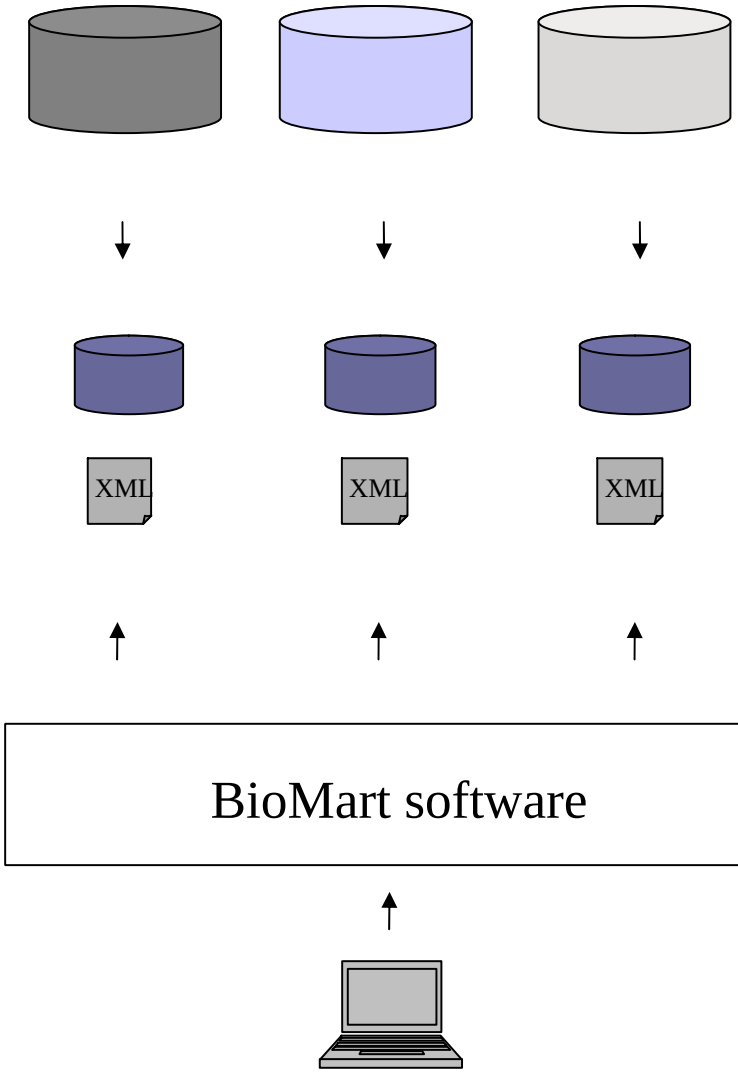
BioMart

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

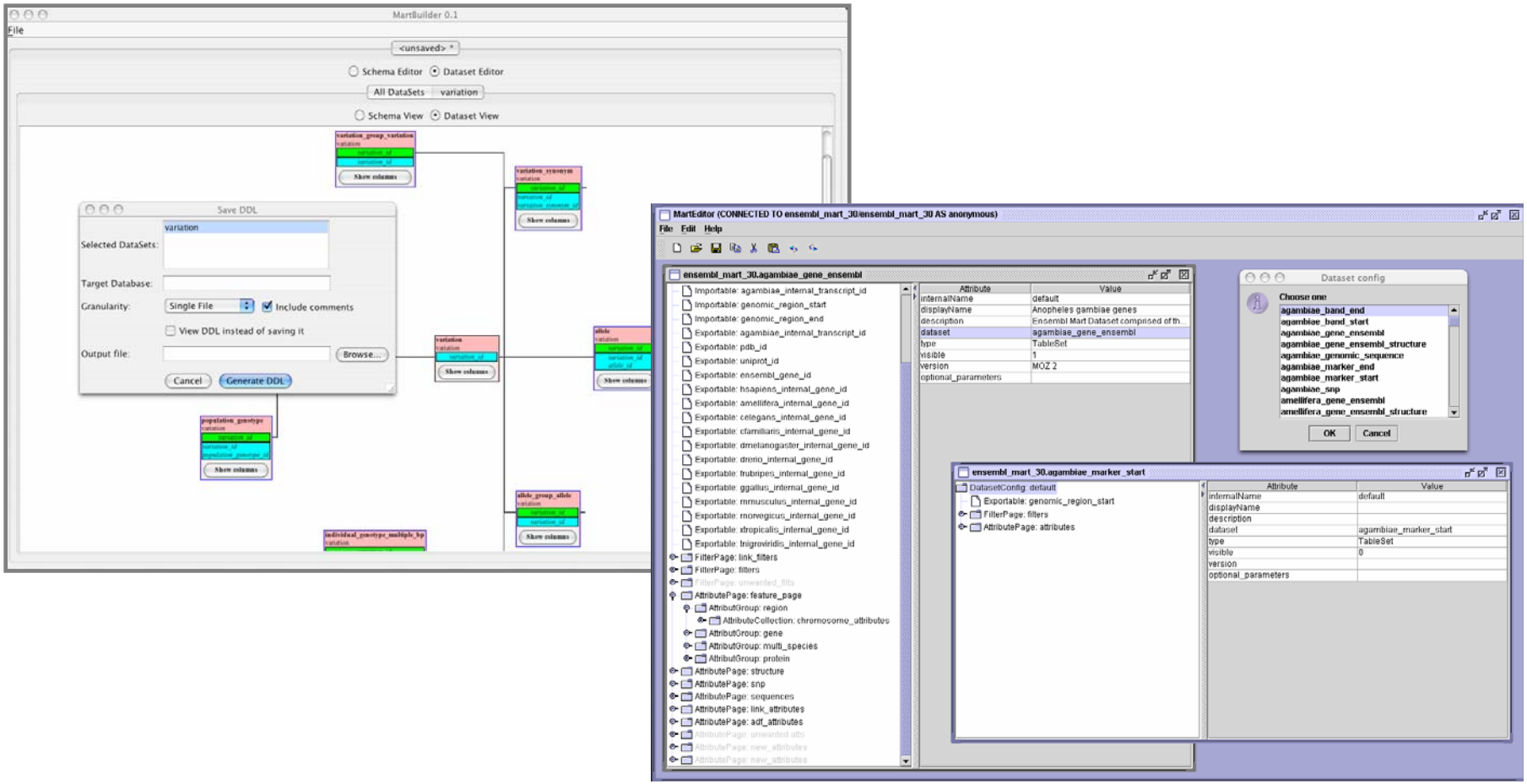
Synopsis

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

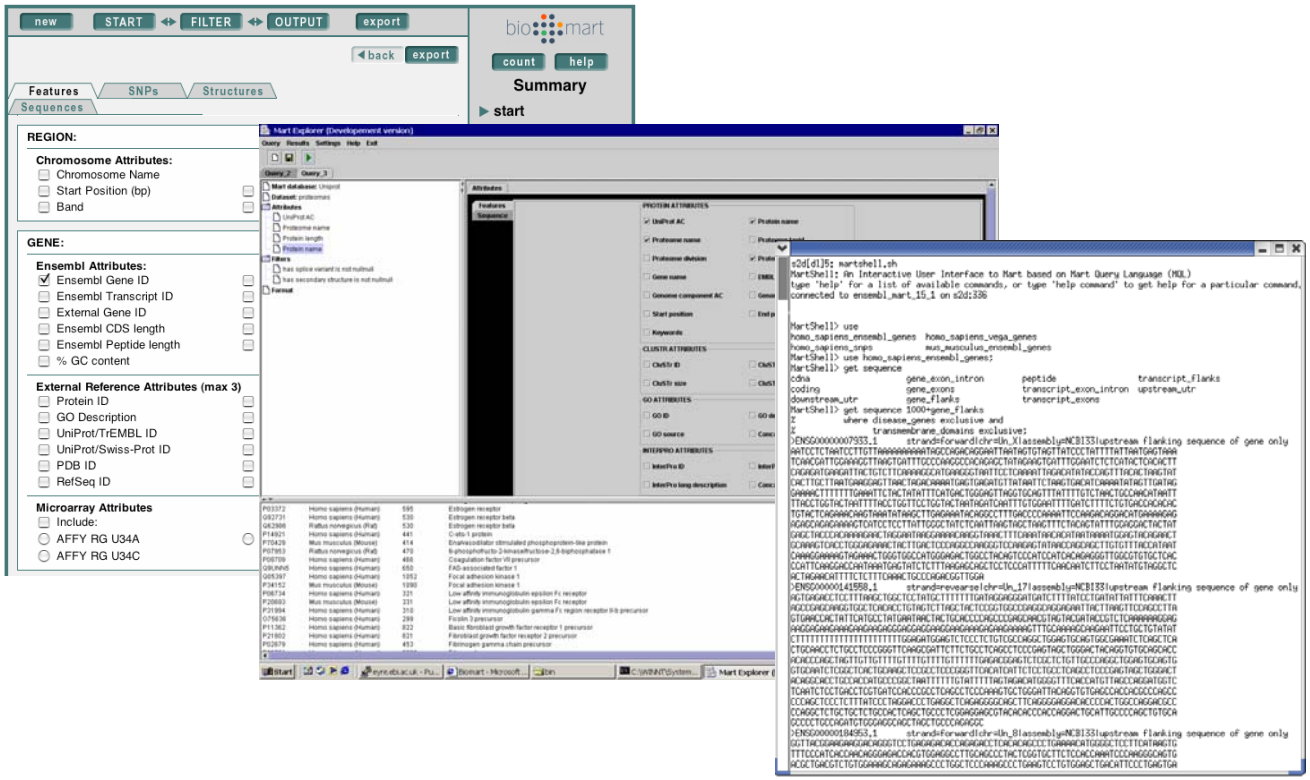
BioMart


Source data
1
Transformation
2
XML
XML
XML
Configuration
3
BioMart software
Querying

Transformation and Configuration Tools



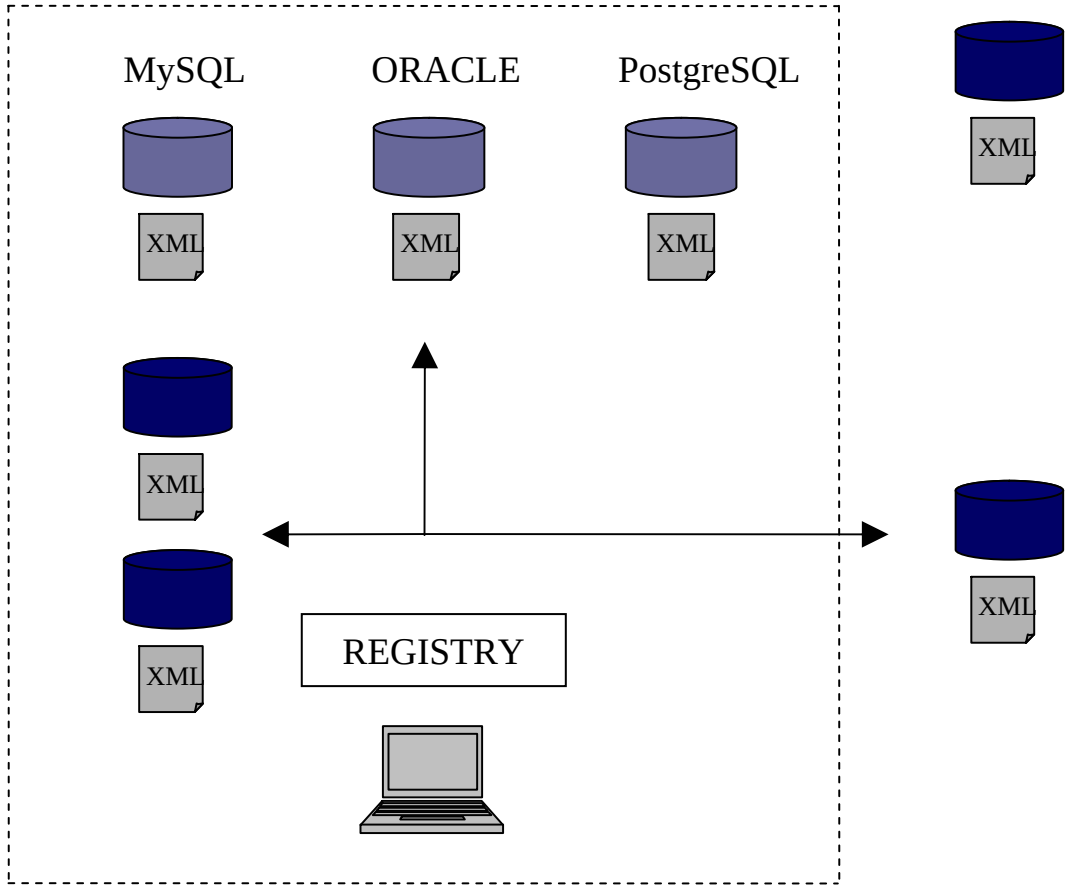
Query interfaces



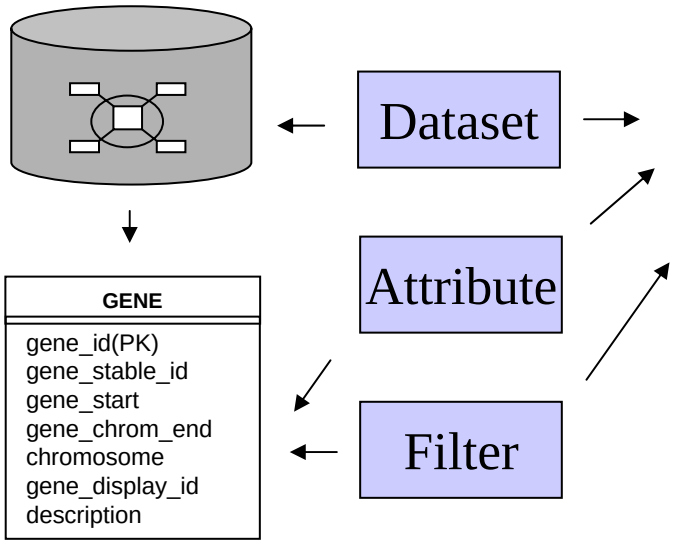
Programmatic access

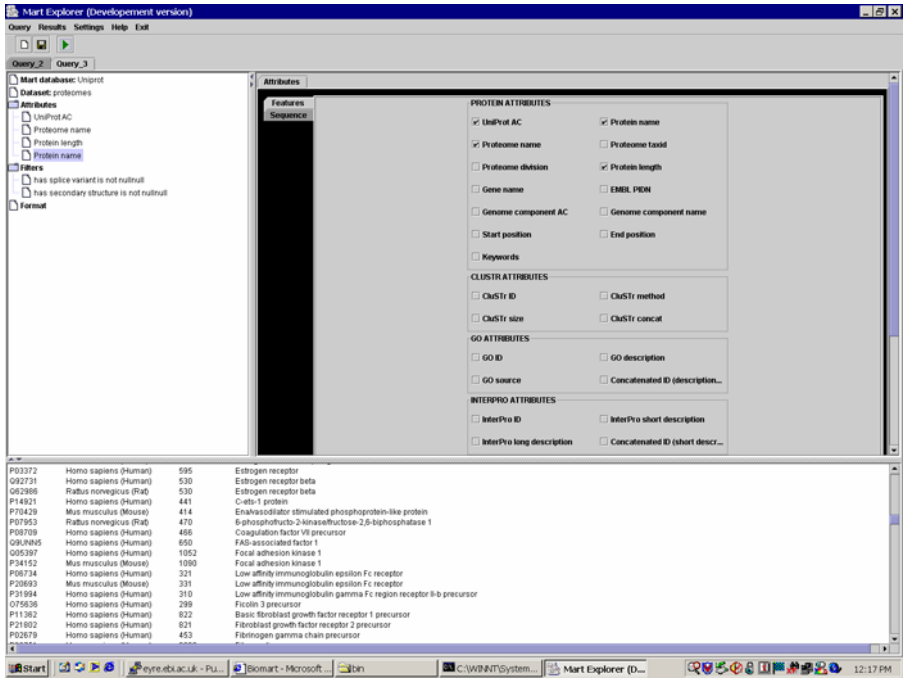
- APIs
  - Perl (biomart-plib)
  - Java (martj)
  - R (biomaRt)
- Web service

Data federation

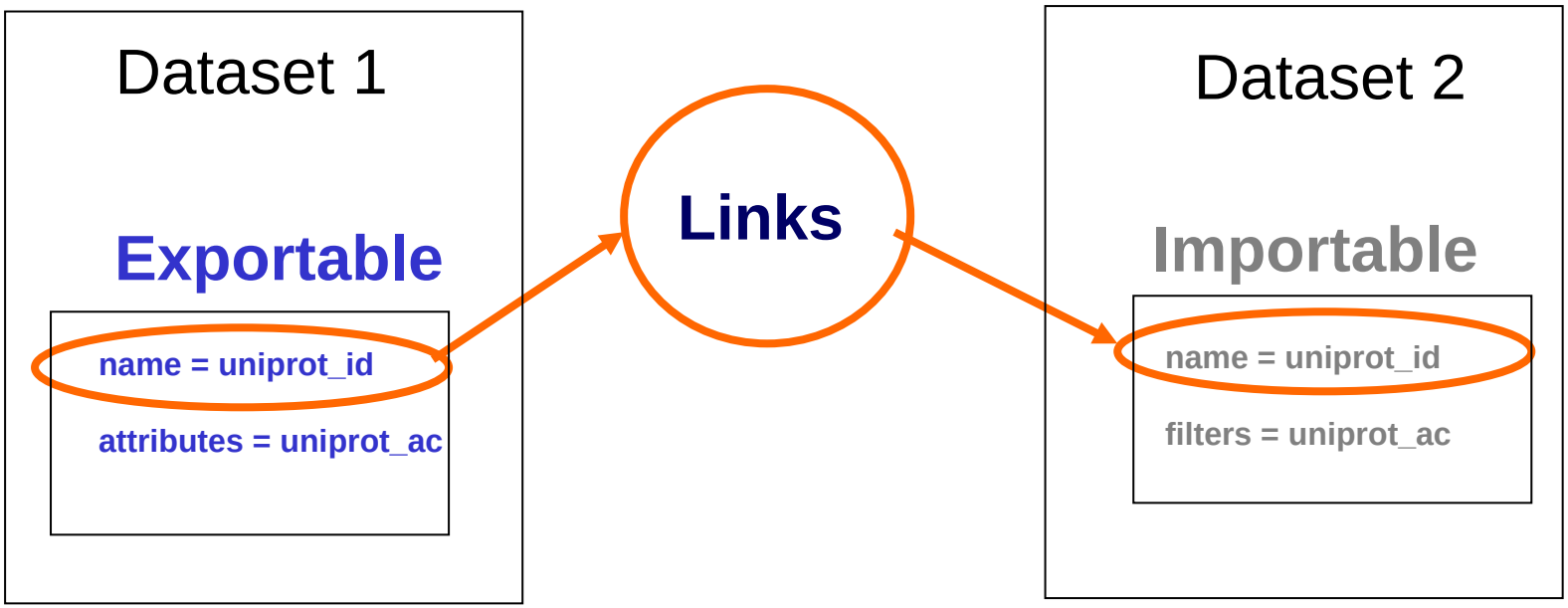


Dataset, Attribute and Filter

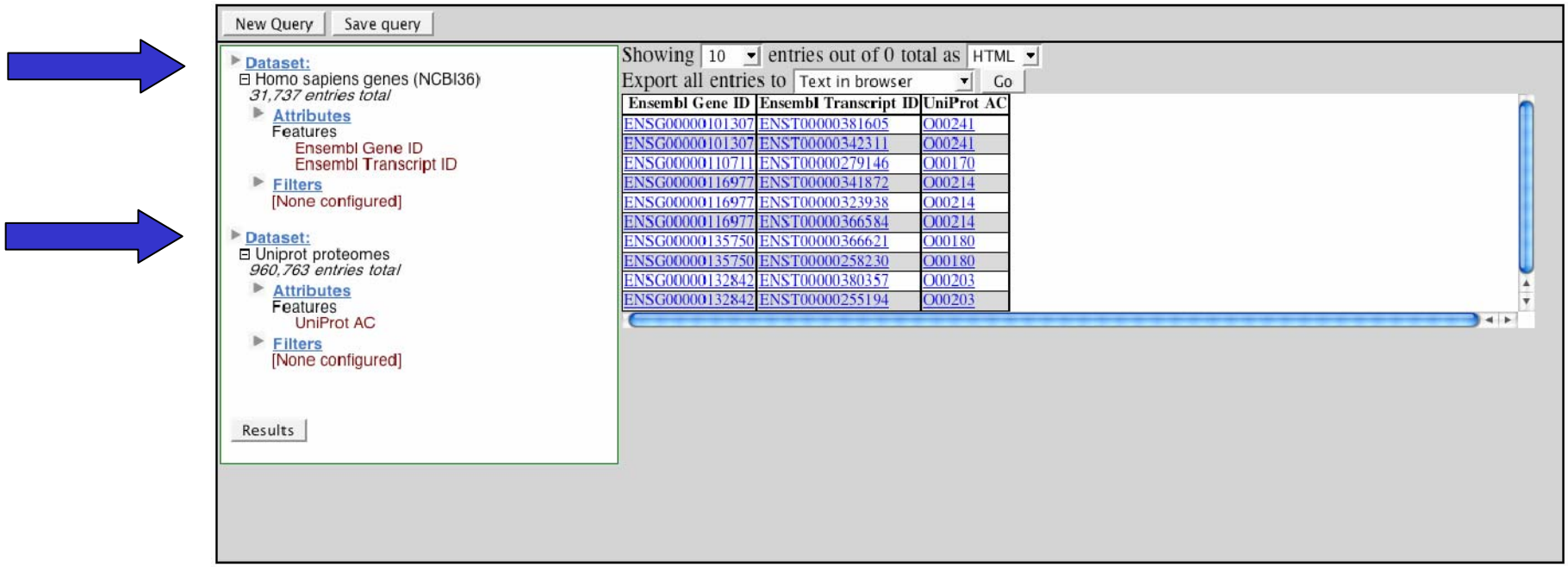




Joining two datasets

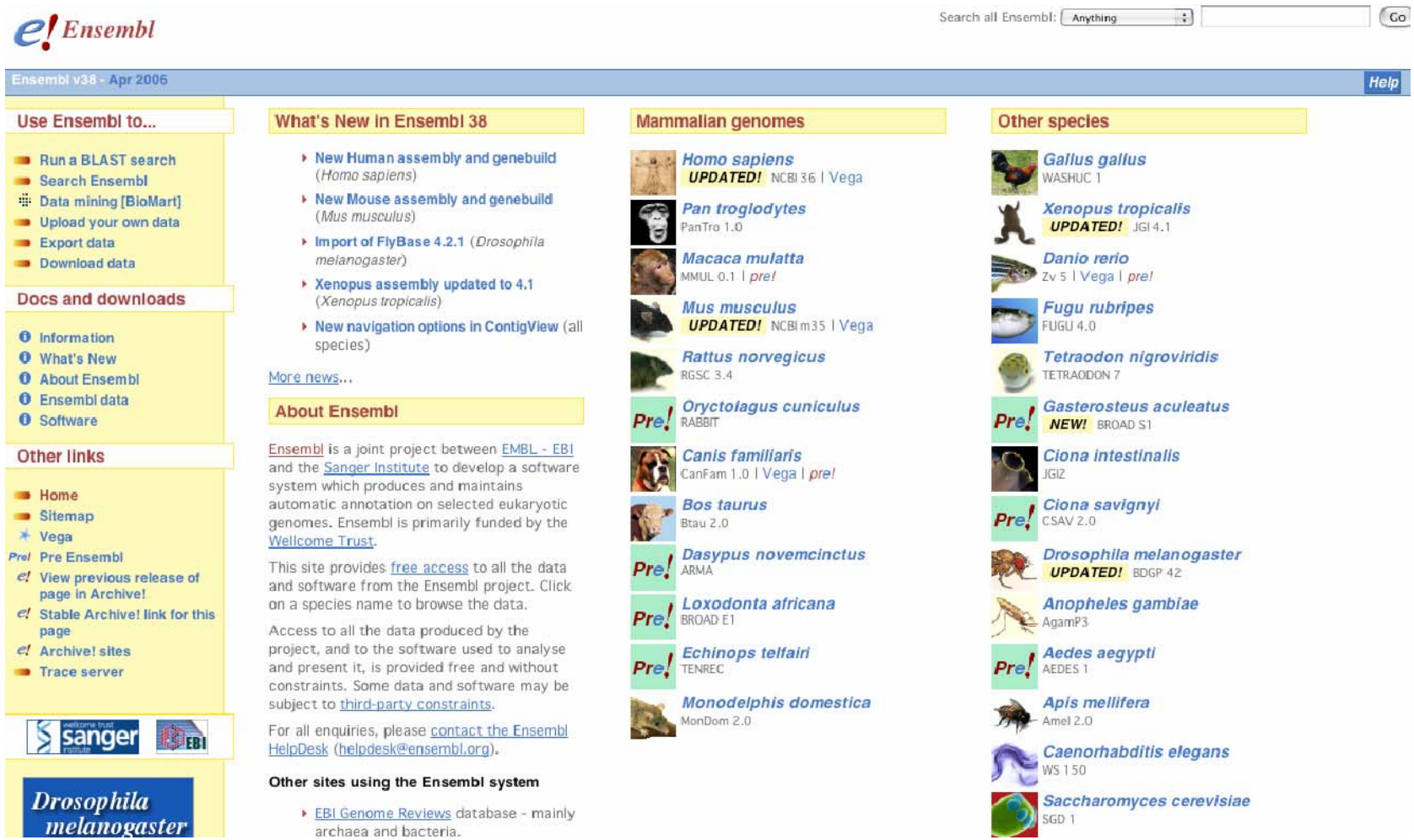


Dataset linking

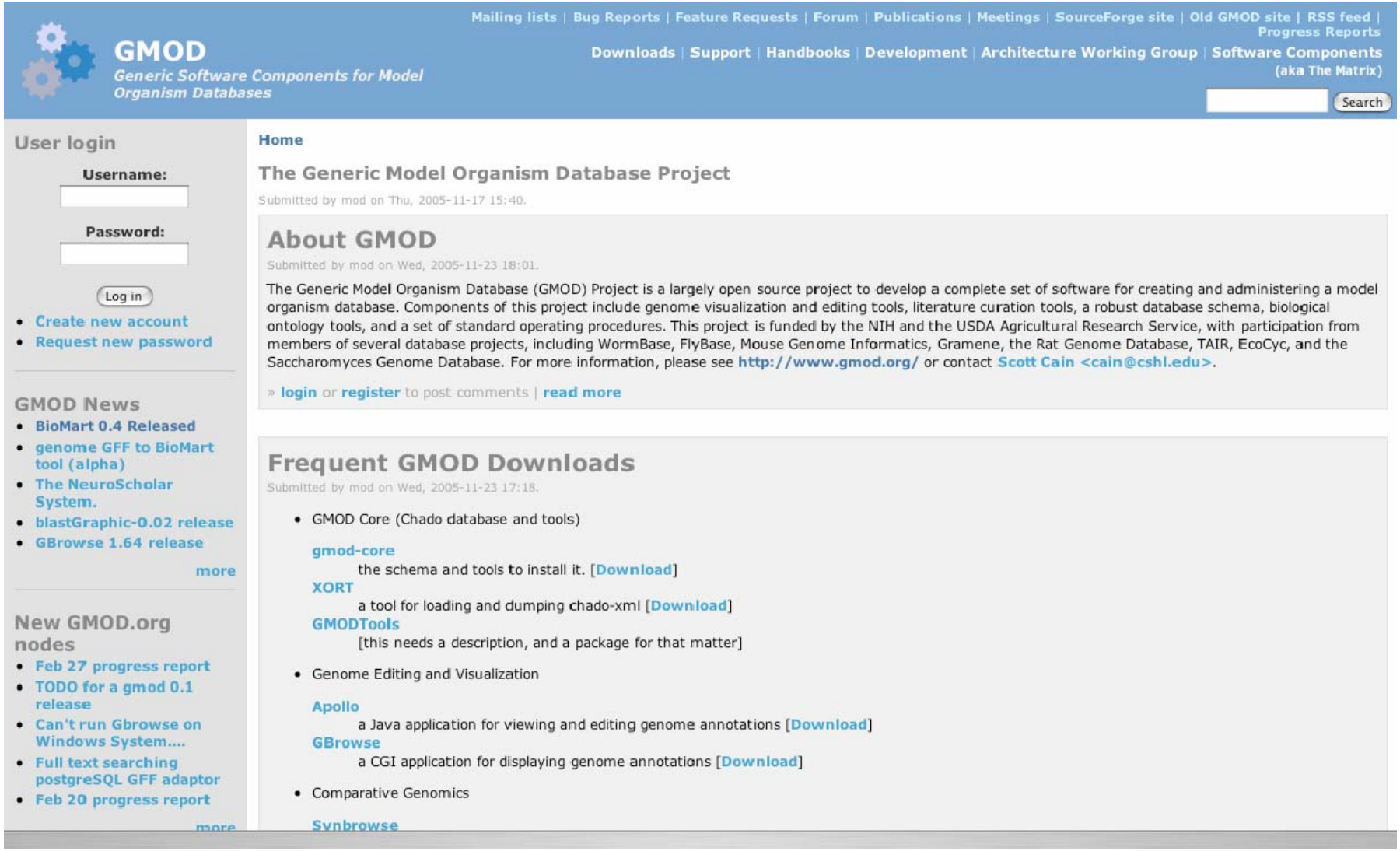


Third party software

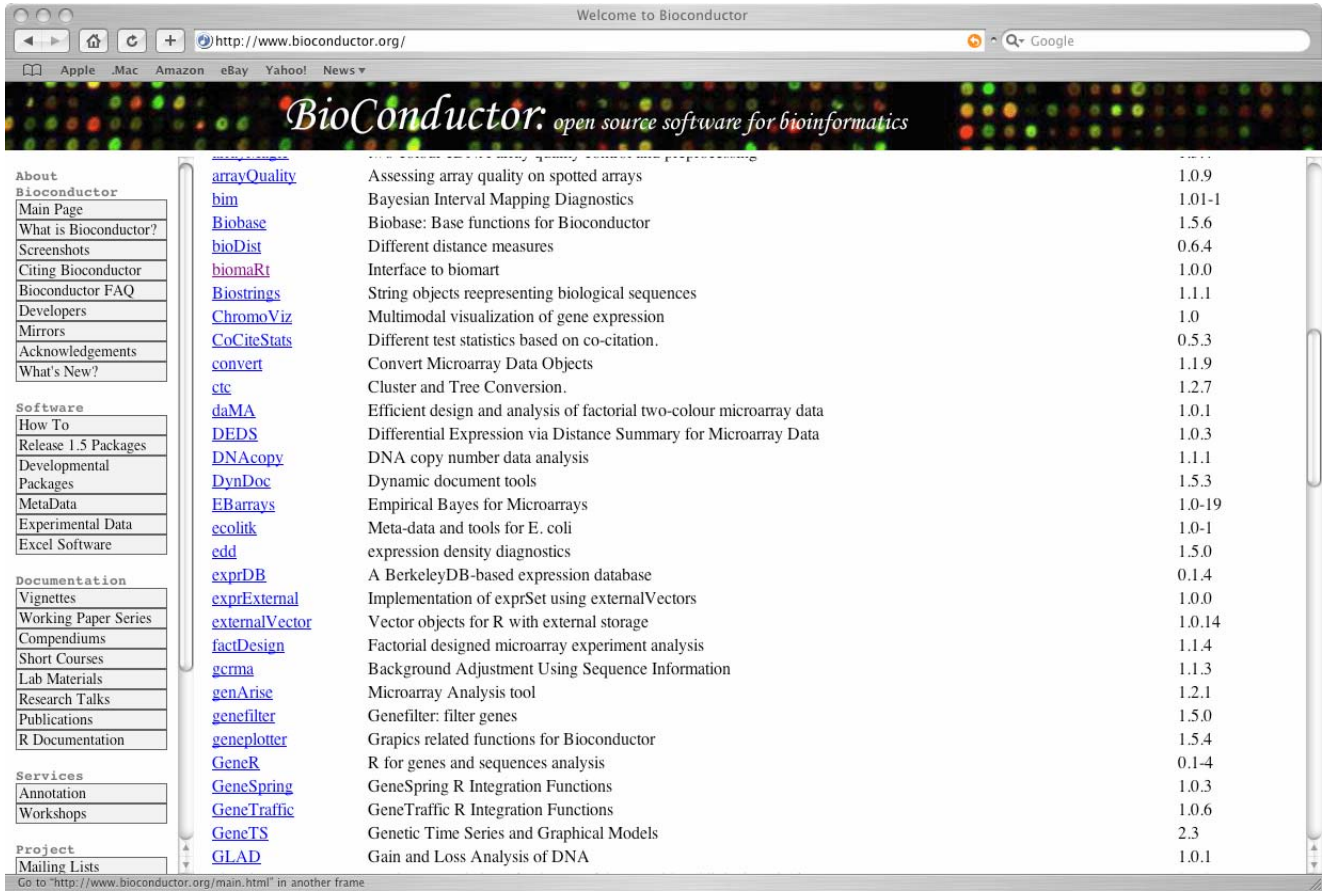
Ensembl


Ensembl
Search all Ensembl: Anything Go
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).
Other sites using the Ensembl system
EBI Genome Reviews database - mainly archaea and bacteria.
Mammalian genomes
Homo sapiens UPDATED! NCBI 36 | 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
Dasypus 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 Zv 5 | 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 Amel 2.0
Caenorhabditis elegans WS 150
Saccharomyces cerevisiae SGD 1

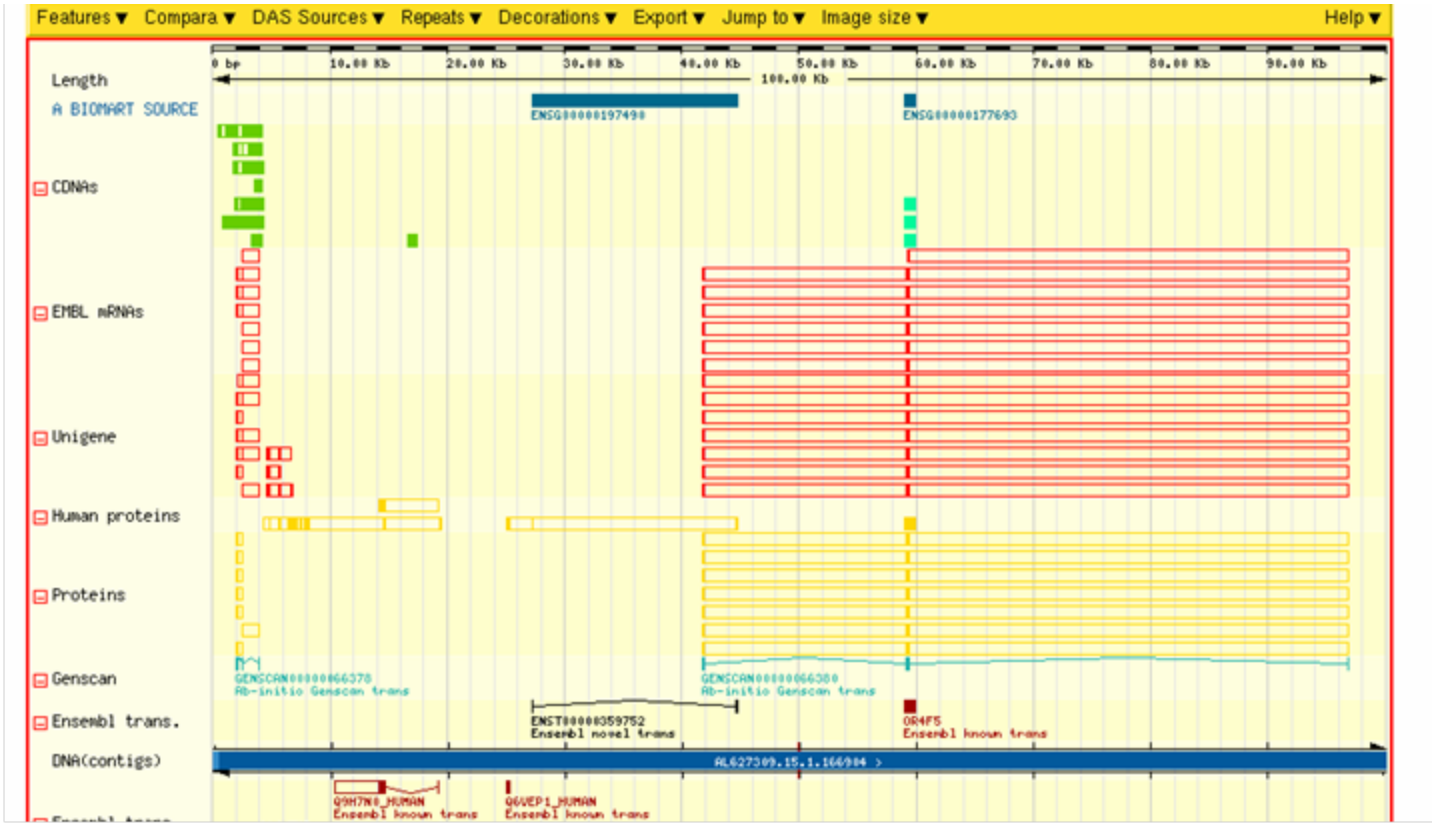
GMOD


Mailing lists | Bug Reports | Feature Requests | Forum | Publications | Meetings | SourceForge site | Old GMOD site | RSS feed | Progress Reports
GMOD
Generic Software Components for Model Organism Databases
Downloads | Support | Handbooks | Development | Architecture Working Group | Software Components (aka The Matrix)
Search
User login
Username:
Password:
Log in
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 postgreSQL 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@cshl.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

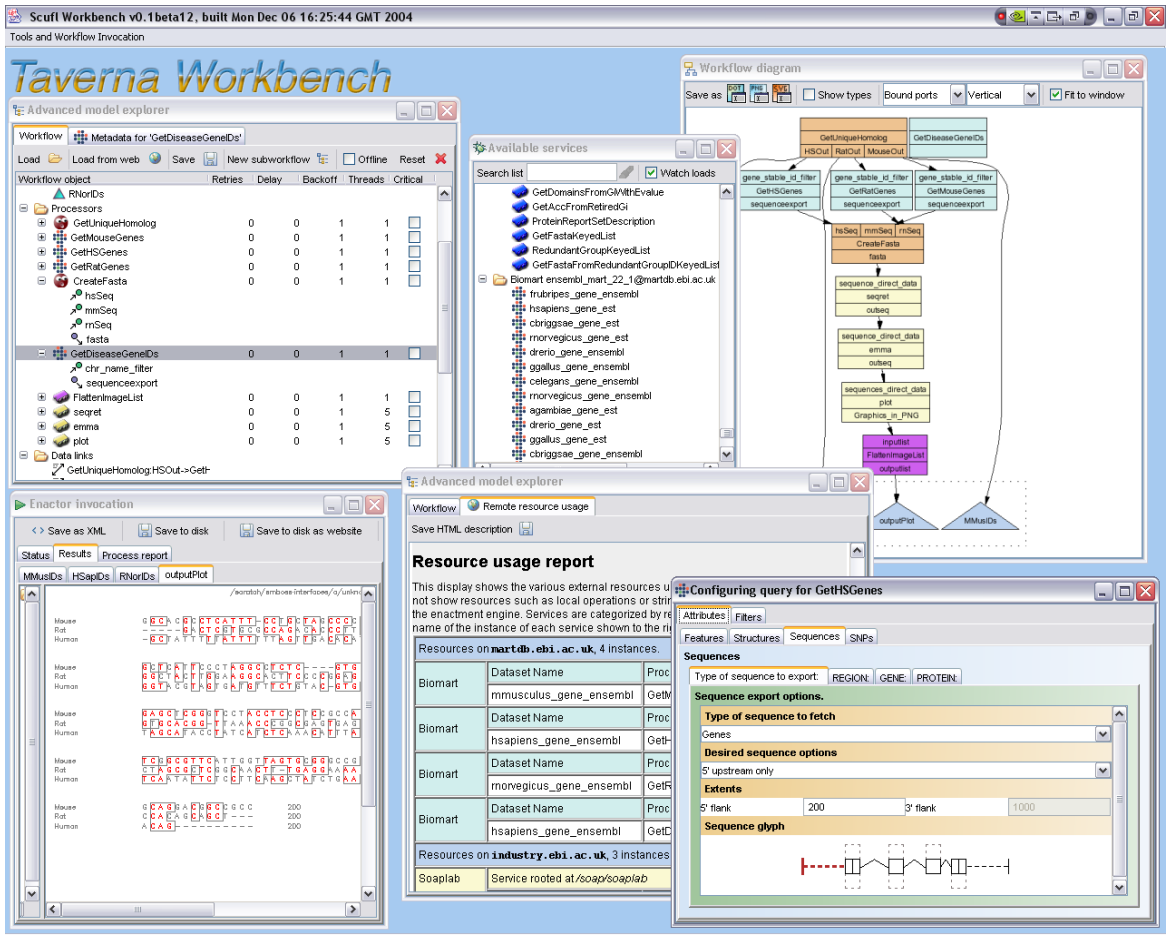
biomaRt


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

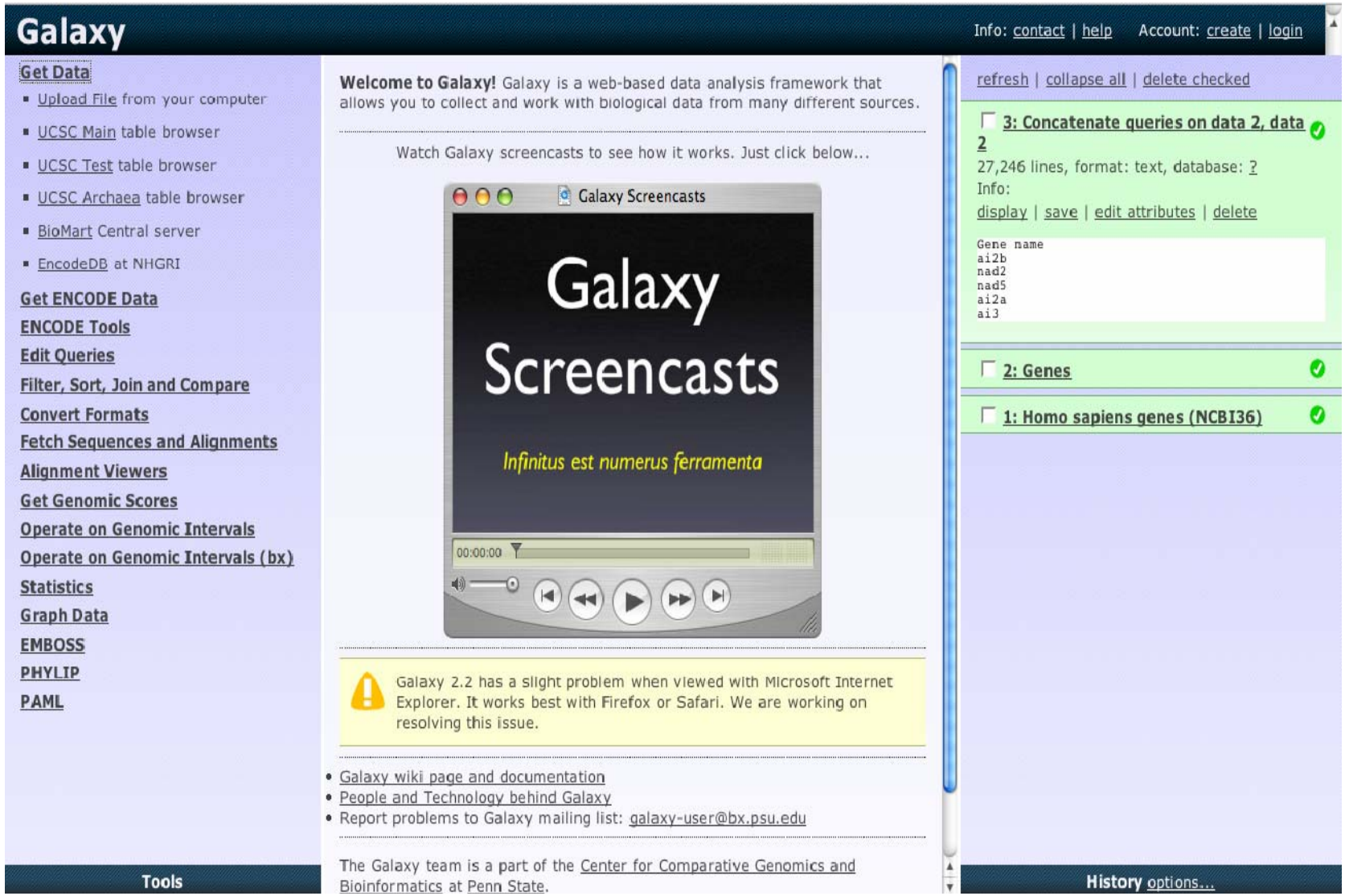
Distributed Annotation System



Taverna

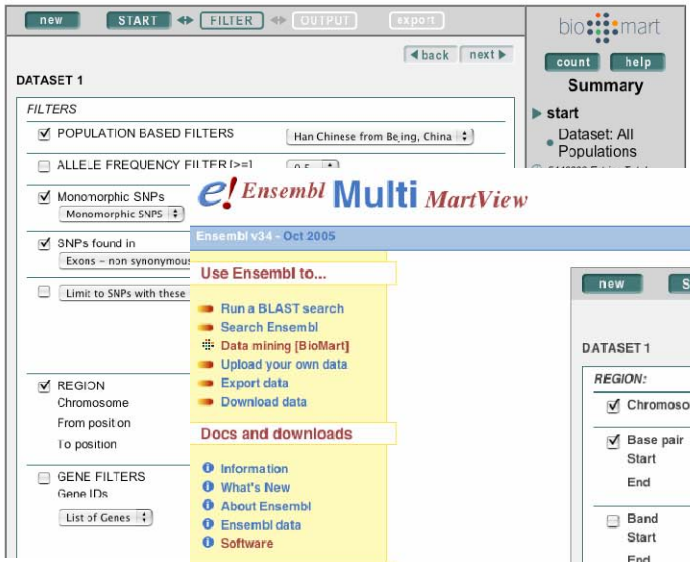

Scufl Workbench v0.1beta12, built Mon Dec 06 16:25:44 GMT 2004
Tools and Workflow Invocation
Taverna Workbench
Advanced model explorer
Workflow
Metadata for 'GetDiseaseGeneIDs'
Load
Load from web
Save
New subworkflow
Offline
Reset
Workflow object
Retries
Delay
Backoff
Threads
Critical
RNorIDs
Processors
GetUniqueHomolog
GetMouseGenes
GetHSGenes
GetRatGenes
CreateFasta
hsSeq
mmSeq
rnSeq
fasta
GetDiseaseGeneIDs
chr_name_filter
sequenceexport
FlattenImageList
seqret
emma
plot
Data links
GetUniqueHomolog:HSOut->GetH
Available services
Search list
Watch loads
GetDomainsFromGMWithEvalue
GetAccFromRetiredGi
ProteinReportSetDescription
GetFastaKeyedList
RedundantGroupKeyedList
GetFastaFromRedundantGroupIDKeyedList
Biomart ensembl_mart_22_1@martdb.ebi.ac.uk
frubripes_gene_ensembl
hsapiens_gene_est
cbriggsae_gene_est
rnorvegicus_gene_est
drerio_gene_ensembl
ggallus_gene_ensembl
celegans_gene_ensembl
rnorvegicus_gene_ensembl
agambiae_gene_est
drerio_gene_est
ggallus_gene_est
cbriggsae_gene_ensembl
Workflow diagram
Save as
Show types
Bound ports
Vertical
Fit to window
GetUniqueHomolog
HSOut
RatOut
MouseOut
GetDiseaseGeneIDs
gene_stable_id_filter
GetHSGenes
GetRatGenes
GetMouseGenes
sequenceexport
hsSeq
mmSeq
rnSeq
CreateFasta
fasta
sequence_direct_data
seqret
outseq
emma
sequences_direct_data
plot
Graphics_in_PNG
inputlist
FlattenImageList
outputlist
outputPlot
MMusIDs
Enactor invocation
Save as XML
Save to disk
Save to disk as website
Status
Results
Process report
MMusIDs
HSapIDs
RNorIDs
outputPlot
Advanced model explorer
Workflow
Remote resource usage
Save HTML description
Resource usage report
This display shows the various external resources u
not show resources such as local operations or strin
the enactment engine. Services are categorized by re
name of the instance of each service shown to the ri
Resources on martdb.ebi.ac.uk, 4 instances.
Biomart
Dataset Name
mmusculus_gene_ensembl
hsapiens_gene_ensembl
morvegicus_gene_ensembl
Resources on industry.ebi.ac.uk, 3 instances
Soaplab
Service rooted at /soap/soaplab
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
Tools
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 Screencasts
Galaxy Screencasts
Infinitus est numerus ferramenta
00:00:00
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
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)
History options...

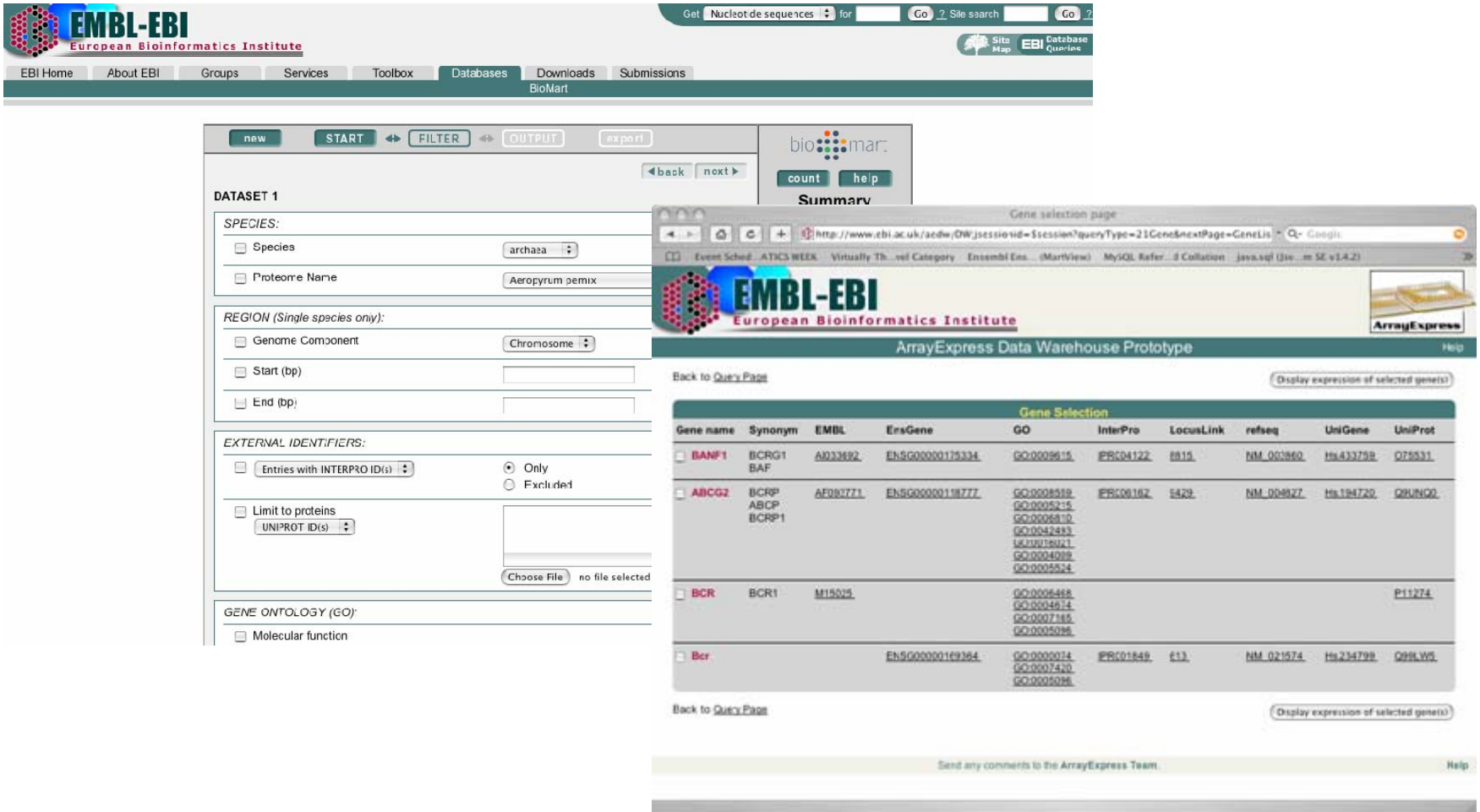
Examples

Genomic data

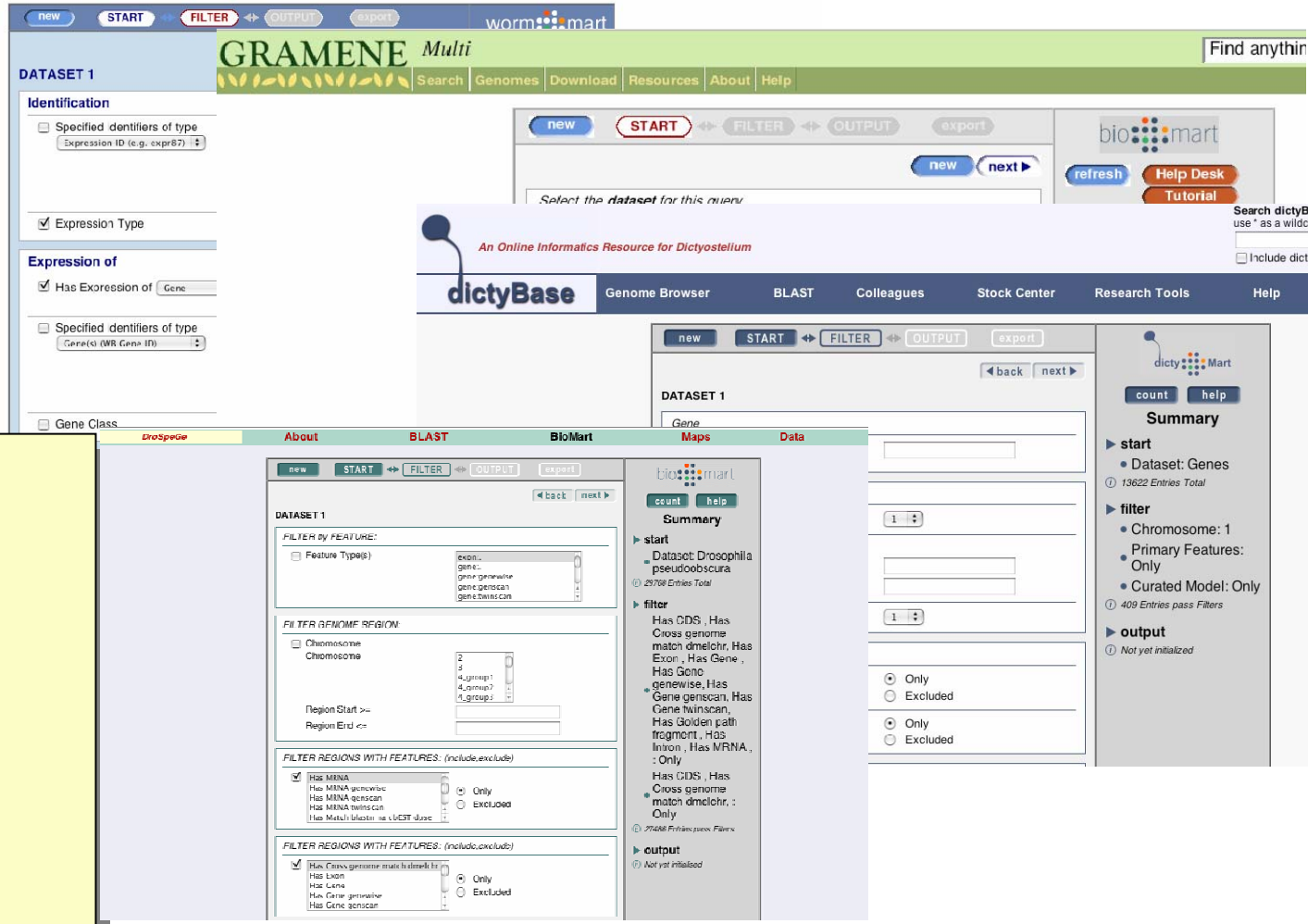


Uniprot, MSD, ArrayExpress

Proteomic, structure, expression

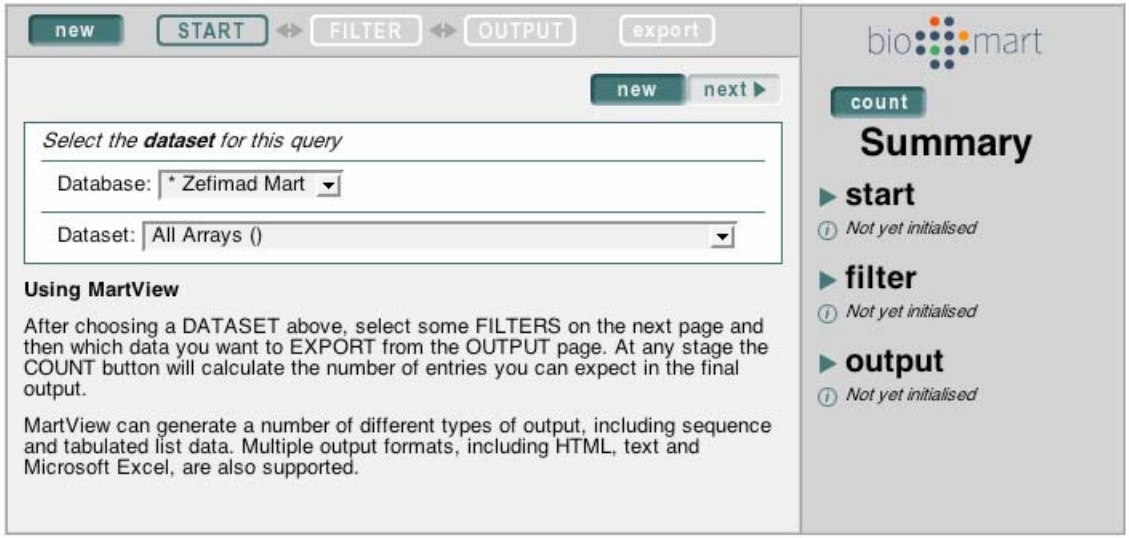


Model organism databases



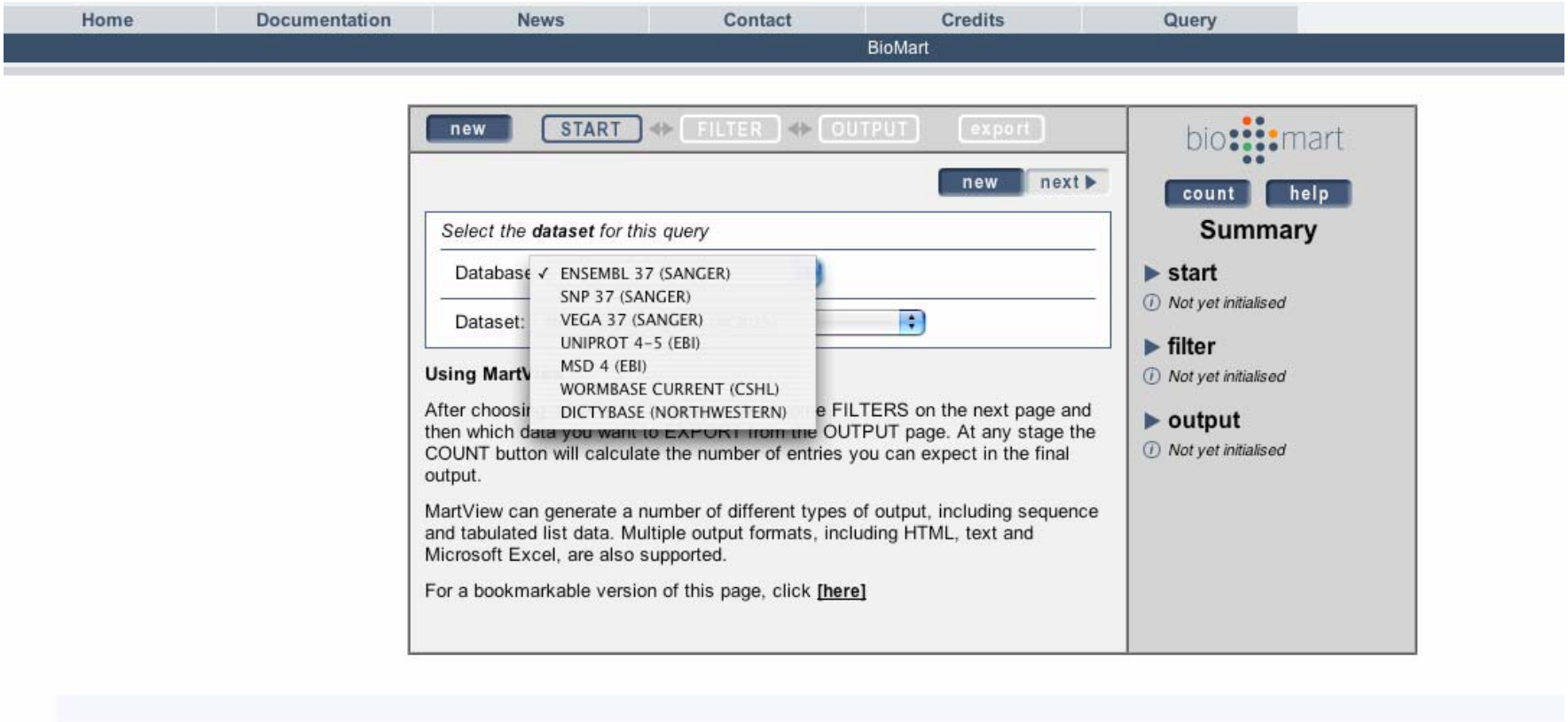
- **Genes**
- **Expression**
- **Phenotypes**
- **Variations**
- Literature
- Ontologies
- Sequence

Zebra Fish models for human development and disease



"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



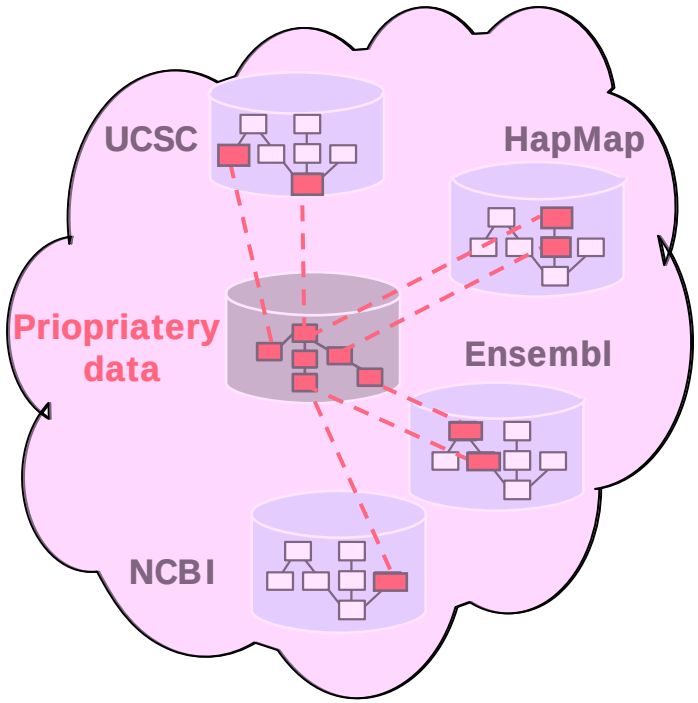
Behind closed doors ;)

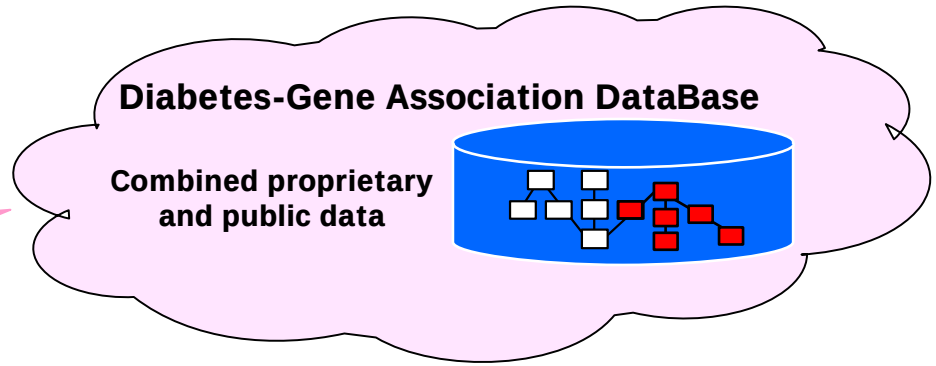
Target SNP selection for the study of one autoimmune disease, type 1 diabetes (T1D), and infectious diseases, malaria and dengue

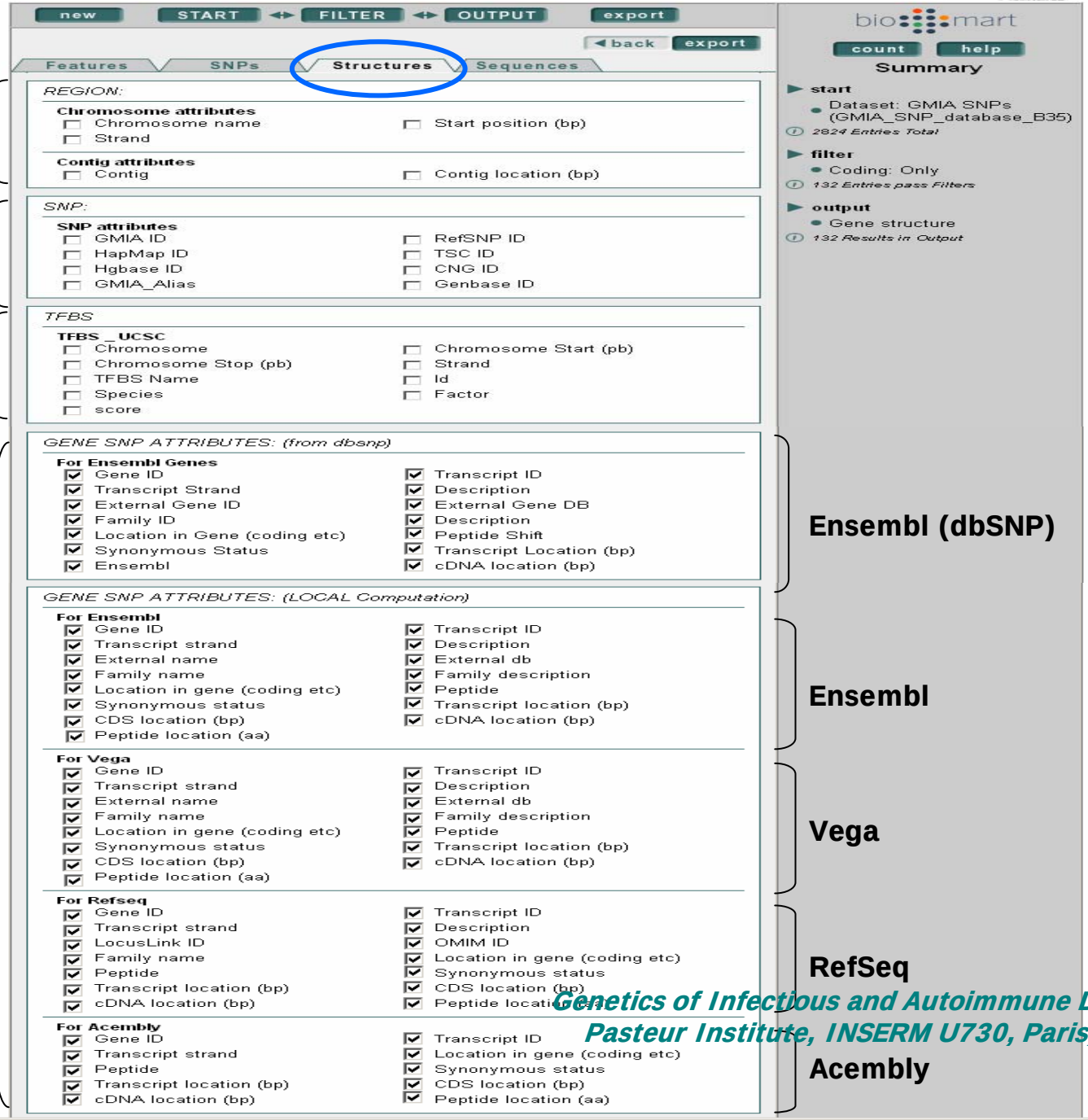
Laboratory of Genetics of Infectious and Autoimmune Diseases

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







Genome Location

Links to databases

Overlaps with TFBS

Location + predicted functional role

Ensembl (dbSNP)

Ensembl

Vega

RefSeq

Acembly

Output format :

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

Using the Molecular Integration Database to Answer CAPRISA's Questions

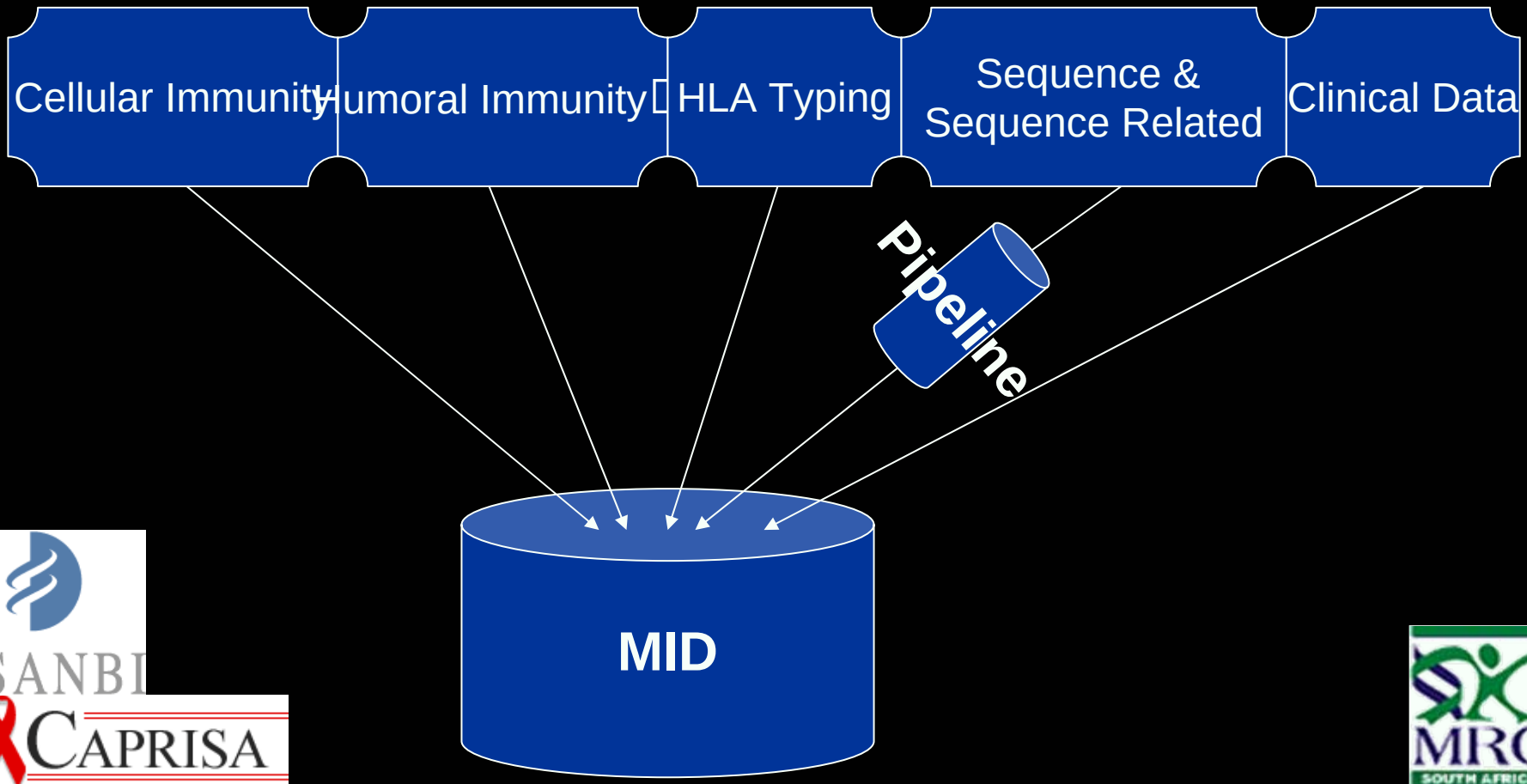
Research that contributes to understanding HIV pathogenesis and epidemiology as well as HIV/AIDS treatment and prevention

SANBI

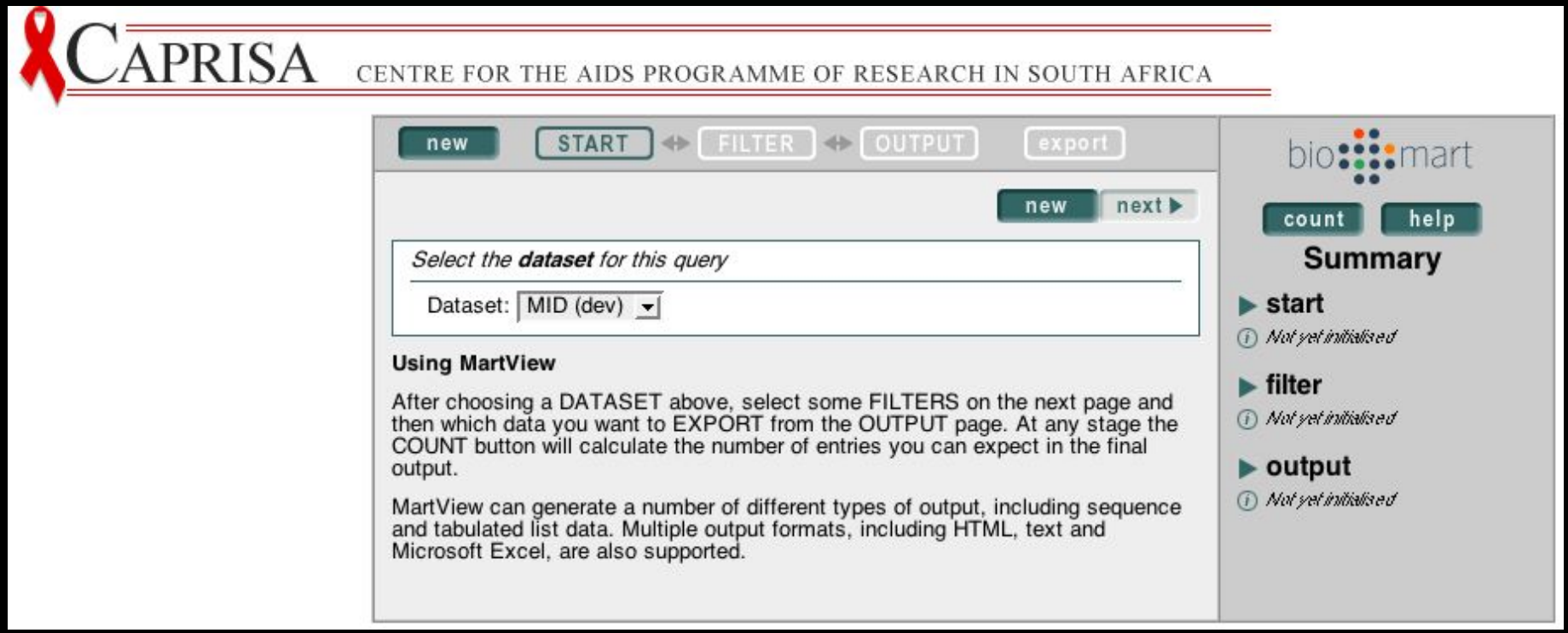
CAPRISA CENTRE FOR THE AIDS PROGRAMME OF RESEARCH IN SOUTH AFRICA

MRC SOUTH AFRICA

How is the MID populated?



Caprisa



SANBI

CAPRISA

MRC SOUTH AFRICA

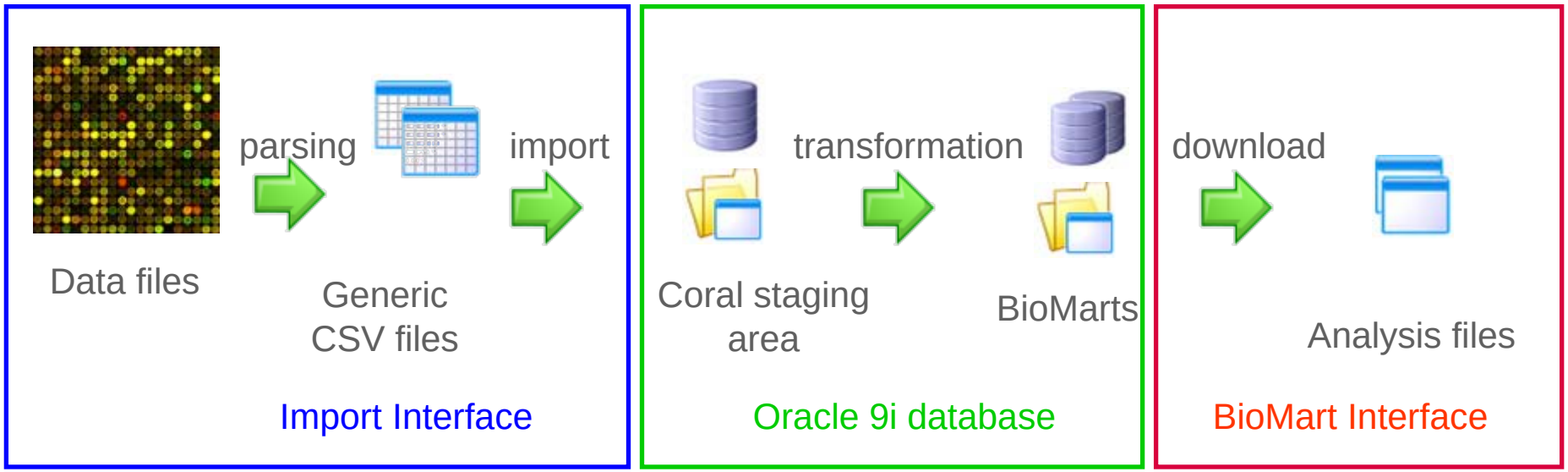
What role for ‘Omics’ ?

- **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

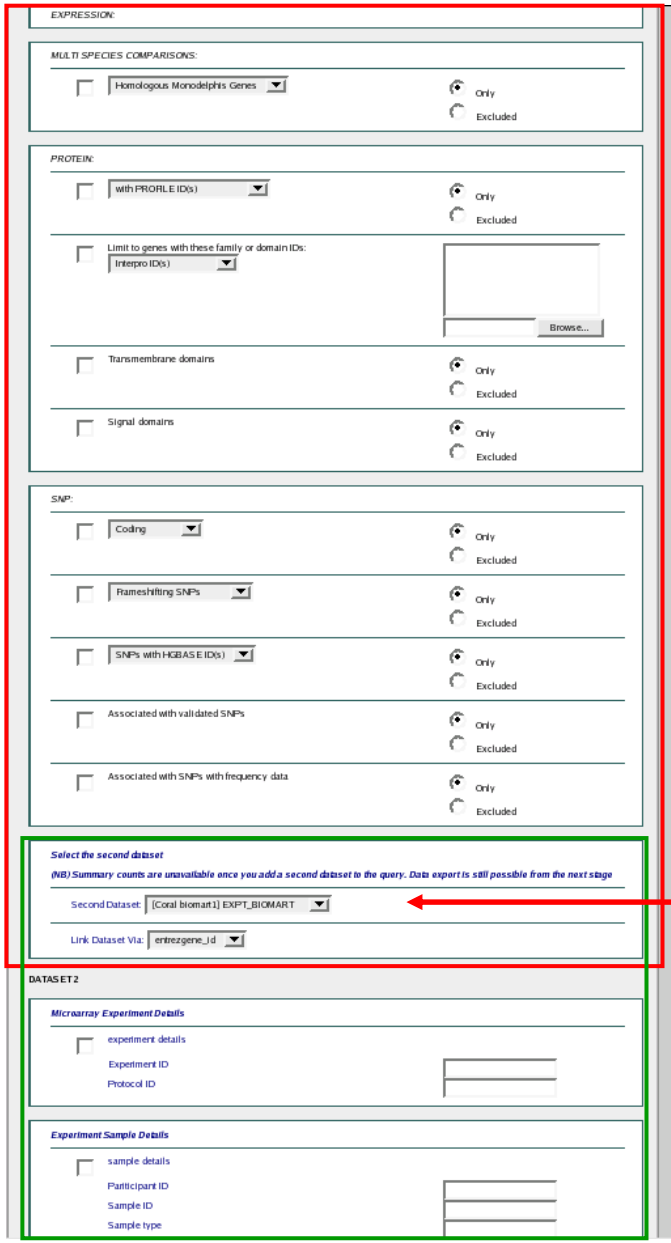


- Requires an extensible file and metadata management system for omics data

- **Oracle 9i** database used for staging area and BioMarts
- **Database indexes files** on a separate file system

Unilever

Adding Annotation



Ensembl Mart

Link on Entrez gene id

Microarray 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

Unilever

Four easy(?) steps

Step 1

Transformation

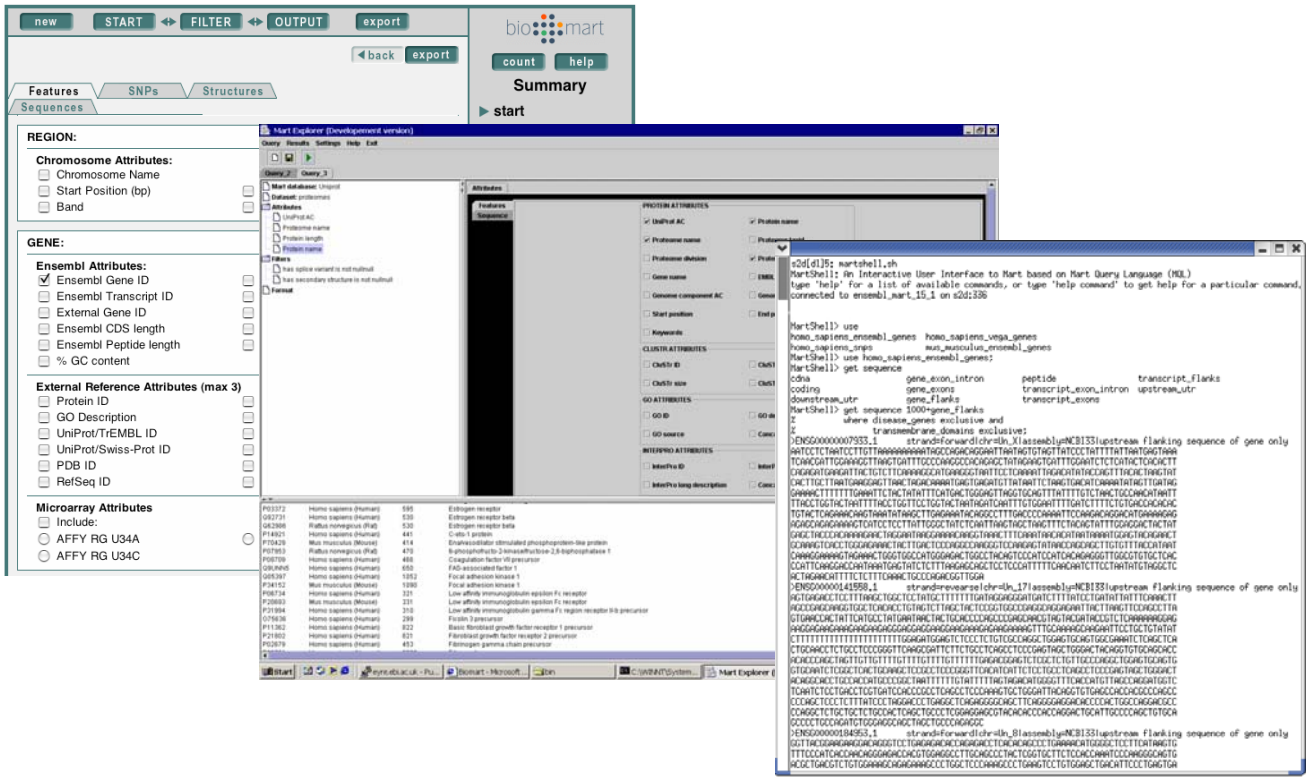
Step 2

Configuration

Step 3

Query

User interfaces



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)