Empowering microarray data analysis with metadata from biological databases through biomaRt

Steffen Durinck[1,2], Wolfgang Huber[2]

sdurinck@ebi.ac.uk

1. SCD-ESAT, KULeuven, Leuven, Belgium
2. EBI, Hinxton-Cambridge, UK

This workshop

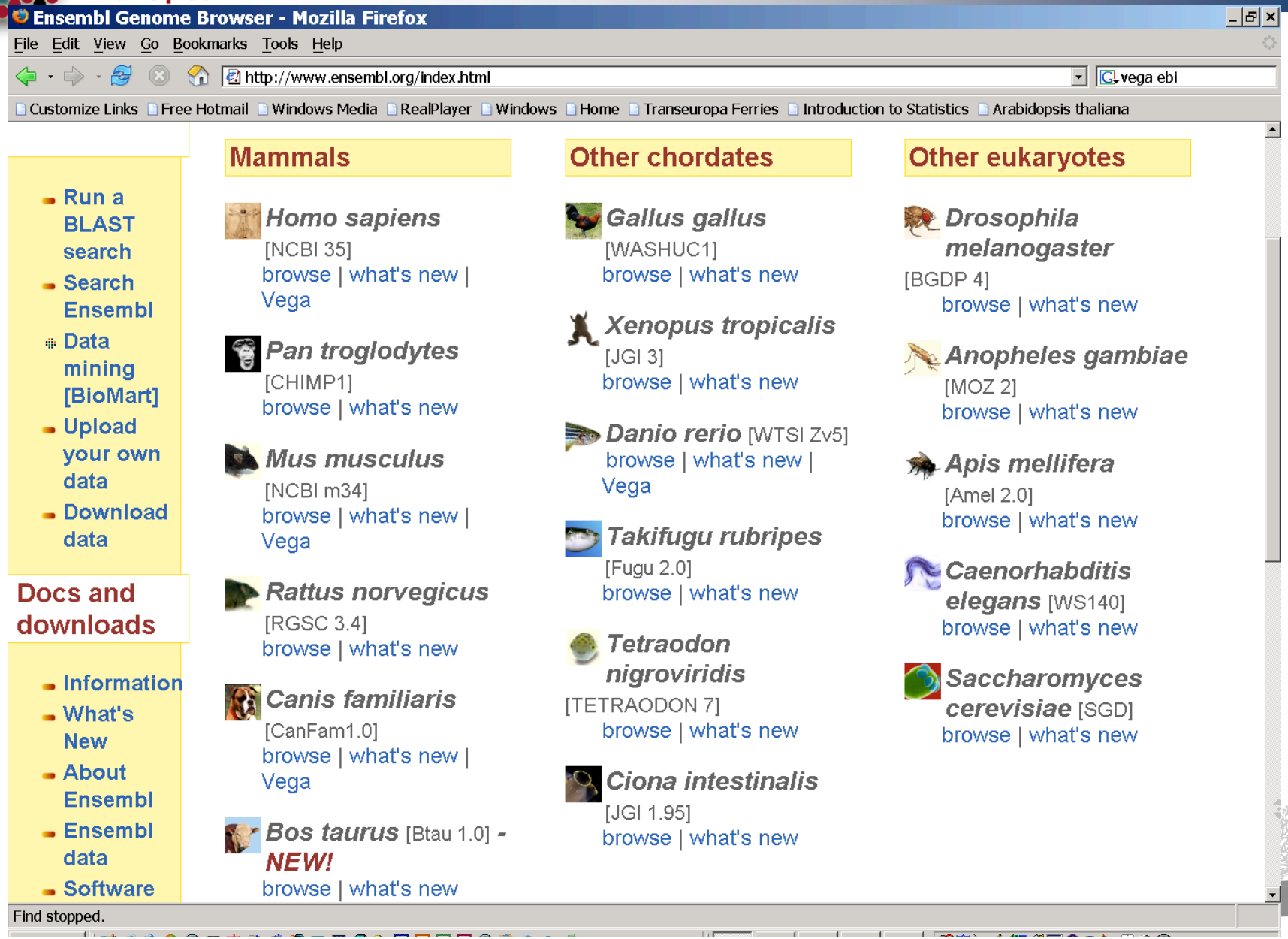
- What is BioMart?
- Overview of biomaRt package for Bioconductor
- Hands-on

BioMart databases

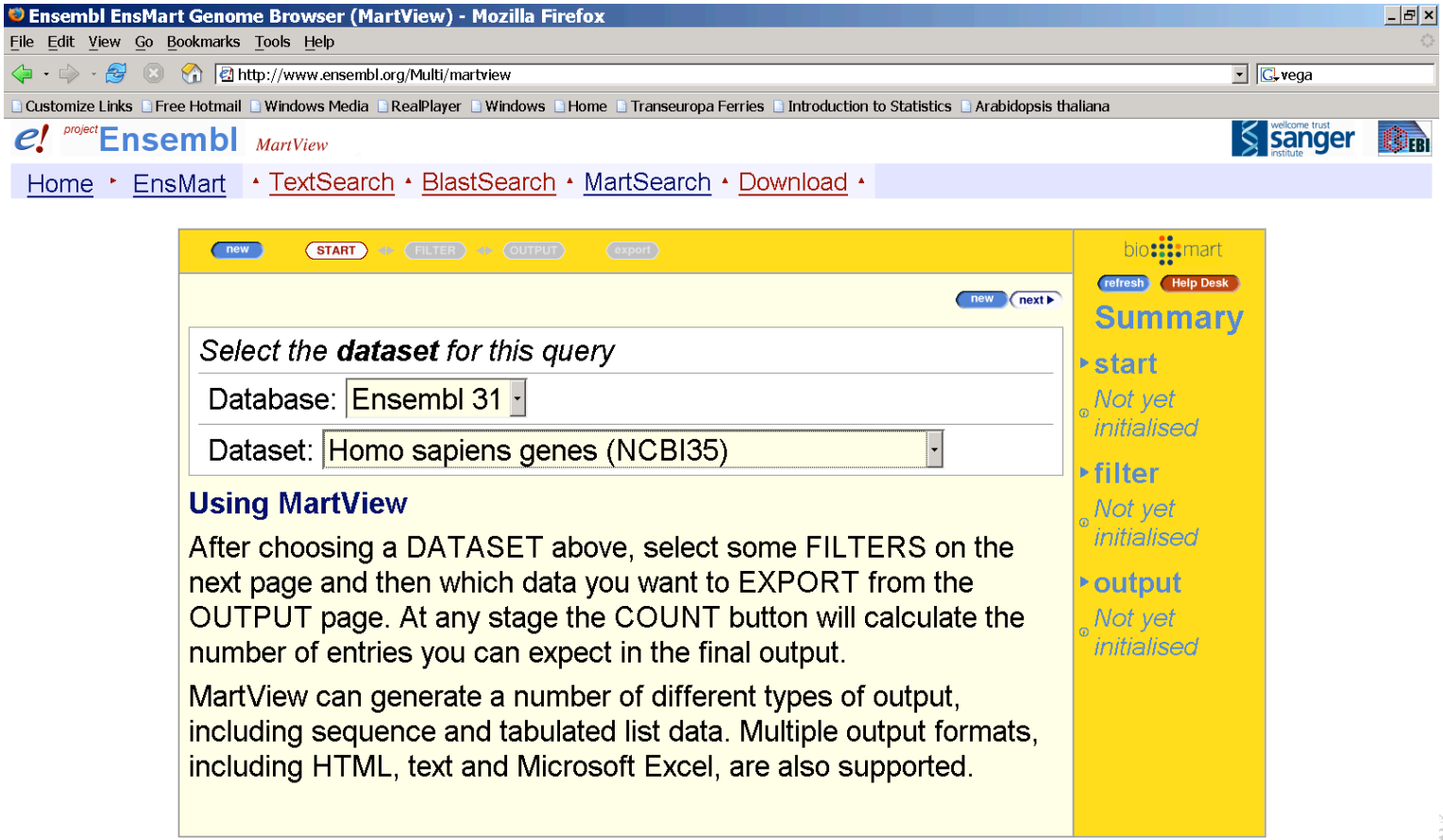
Ensembl

- Joint project between EMBL EBI and the Sanger Institute
- Produces and maintains automatic annotation on selected eukaryotic genomes.

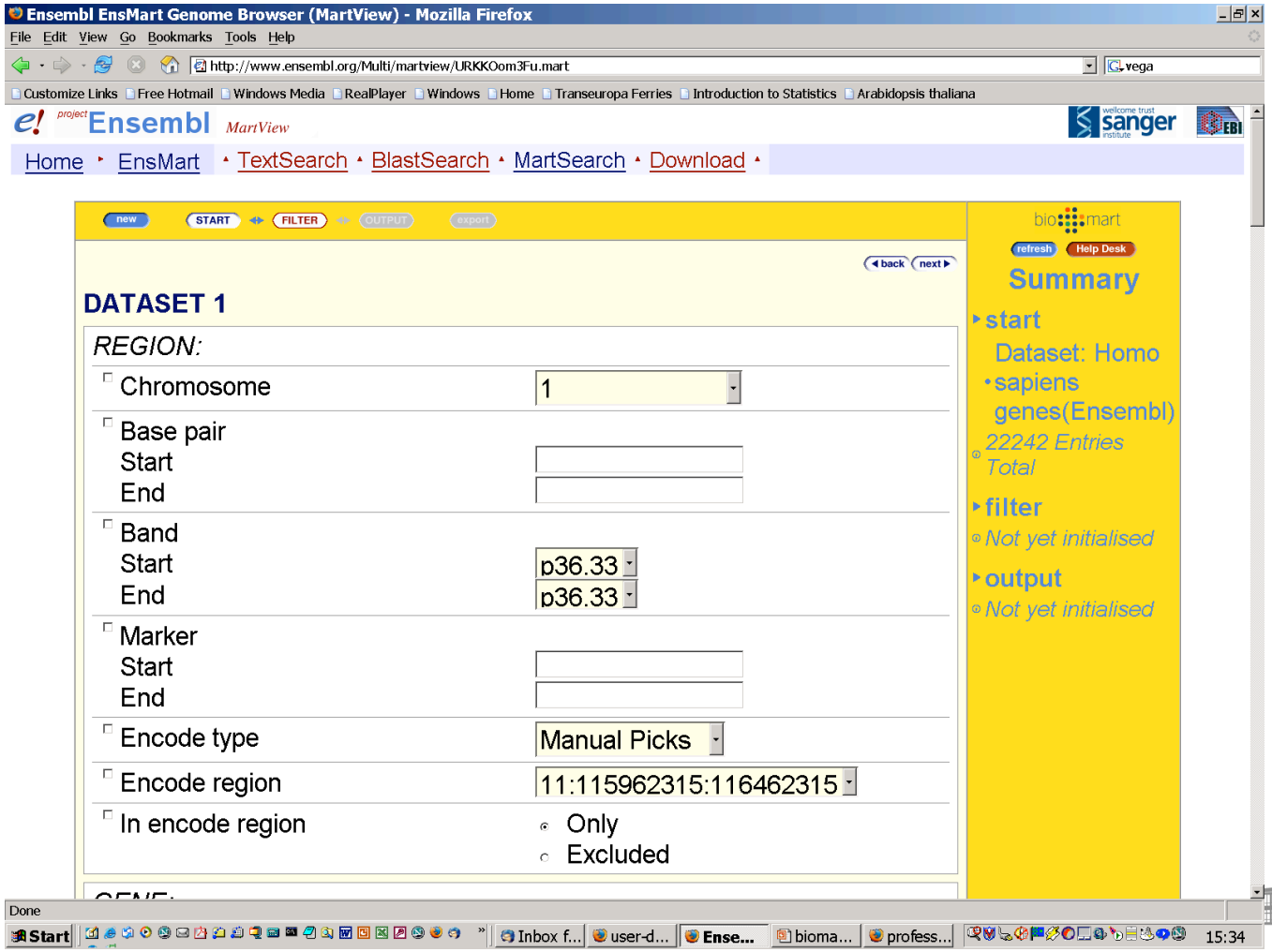
http://www.ensembl.org


Ensembl Genome Browser - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.ensembl.org/index.html
vega ebi
Customize Links Free Hotmail Windows Media RealPlayer Windows Home Transeuropa Ferries Introduction to Statistics Arabidopsis thaliana
Run a BLAST search
Search Ensembl
Data mining [BioMart]
Upload your own data
Download data
Docs and downloads
Information
What's New
About Ensembl
Ensembl data
Software
Mammals
Homo sapiens [NCBI 35] browse | what's new | Vega
Pan troglodytes [CHIMP1] browse | what's new
Mus musculus [NCBI m34] browse | what's new | Vega
Rattus norvegicus [RGSC 3.4] browse | what's new
Canis familiaris [CanFam1.0] browse | what's new | Vega
Bos taurus [Btau 1.0] - NEW! browse | what's new
Other chordates
Gallus gallus [WASHUC1] browse | what's new
Xenopus tropicalis [JGI 3] browse | what's new
Danio rerio [WTSI Zv5] browse | what's new | Vega
Takifugu rubripes [Fugu 2.0] browse | what's new
Tetraodon nigroviridis [TETRAODON 7] browse | what's new
Ciona intestinalis [JGI 1.95] browse | what's new
Other eukaryotes
Drosophila melanogaster [BGDP 4] browse | what's new
Anopheles gambiae [MOZ 2] browse | what's new
Apis mellifera [Amel 2.0] browse | what's new
Caenorhabditis elegans [WS140] browse | what's new
Saccharomyces cerevisiae [SGD] browse | what's new
Find stopped.

Ensembl


Ensembl EnsMart Genome Browser (MartView) - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.ensembl.org/Multi/martview
Customize Links Free Hotmail Windows Media RealPlayer Windows Home Transeuropa Ferries Introduction to Statistics Arabidopsis thaliana
project Ensembl MartView
wellcome trust sanger institute
EBI
Home EnsMart TextSearch BlastSearch MartSearch Download
new START FILTER OUTPUT export
new next
Select the dataset for this query
Database: Ensembl 31
Dataset: Homo sapiens genes (NCBI35)
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.
biomart
refresh Help Desk
Summary
start
Not yet initialised
filter
Not yet initialised
output
Not yet initialised

Ensembl


Ensembl EnsMart Genome Browser (MartView) - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.ensembl.org/Multi/martview/URKKOom3Fu.mart
Customize Links Free Hotmail Windows Media RealPlayer Windows Home Transeuropa Ferries Introduction to Statistics Arabidopsis thaliana
project Ensembl MartView
Home EnsMart TextSearch BlastSearch MartSearch Download
new START FILTER OUTPUT export
back next
DATASET 1
REGION:
Chromosome 1
Base pair
Start
End
Band
Start p36.33
End p36.33
Marker
Start
End
Encode type Manual Picks
Encode region 11:115962315:116462315
In encode region
Only
Excluded
biomart
refresh Help Desk
Summary
start
Dataset: Homo sapiens genes(Ensembl)
22242 Entries Total
filter
Not yet initialised
output
Not yet initialised
Done
Start
Inbox f...
user-d...
Ense...
bioma...
profess...
15:34

VEGA

The Vertebrate Genome Annotation (VEGA) database is a central repository for high quality, frequently updated, manual annotation of vertebrate finished genome sequence.

Current release:
- Human
- Mouse
- Zebrafish
- Dog

http://vega.sanger.ac.uk

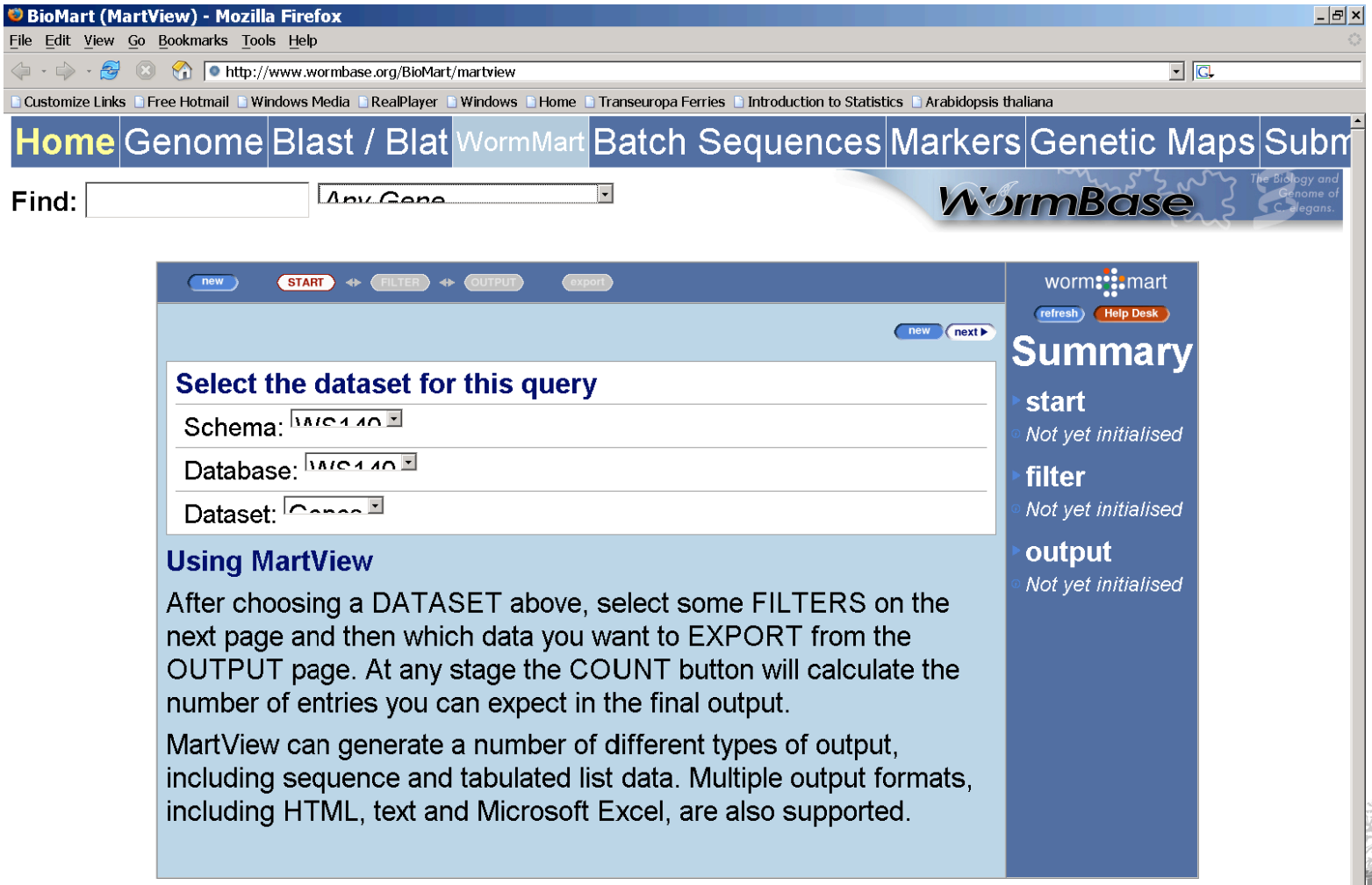
WormBase

WormBase
The Biology and Genome of C. elegans.

WormBase is the repository of mapping, sequencing and phenotypic information for *C. elegans* (and some other nematodes).

http://www.wormbase.org

WormMart


BioMart (MartView) - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.wormbase.org/BioMart/martview
Customize Links Free Hotmail Windows Media RealPlayer Windows Home Transeuropa Ferries Introduction to Statistics Arabidopsis thaliana
Home Genome Blast / Blat WormMart Batch Sequences Markers Genetic Maps Subm
Find:
Any Gene
WormBase
The Biology and Genome of C. elegans
new START FILTER OUTPUT export
new next
Select the dataset for this query
Schema:
Database:
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.
wormmart
refresh Help Desk
Summary
start
Not yet initialised
filter
Not yet initialised
output
Not yet initialised

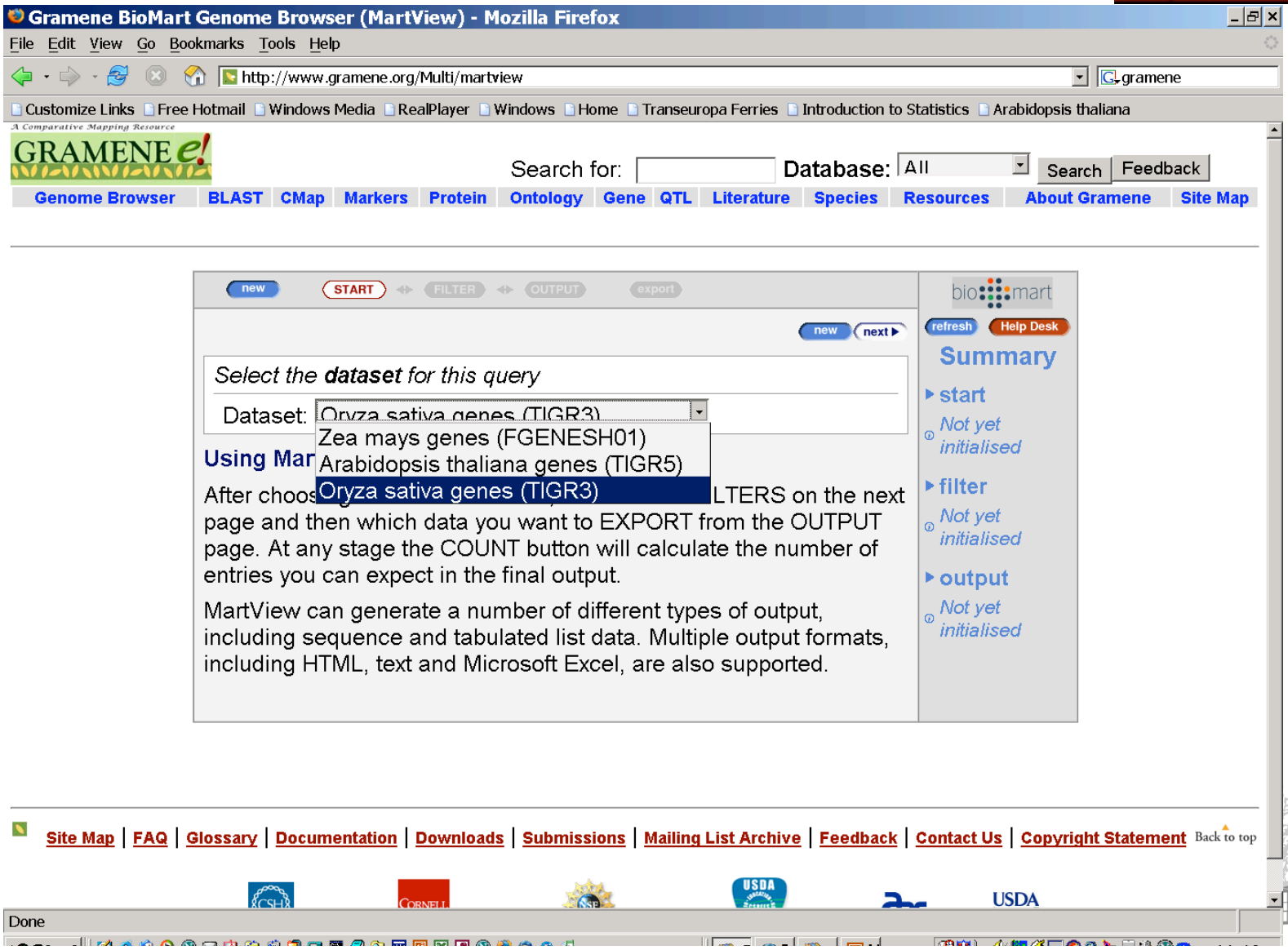
GrameneMart

GRAMENE

Gramene: A Comparative Mapping Resource for Grains

Gramene is a curated, open-source, Web-accessible data resource for comparative genome analysis in the grasses.

http://www.gramene.org


Gramene BioMart Genome Browser (MartView) - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.gramene.org/Multi/martview
Customize Links Free Hotmail Windows Media RealPlayer Windows Home Transeuropa Ferries Introduction to Statistics Arabidopsis thaliana
A Comparative Mapping Resource
GRAMENE e!
Search for:
Database: All
Search
Feedback
Genome Browser BLAST CMap Markers Protein Ontology Gene QTL Literature Species Resources About Gramene Site Map
new START FILTER OUTPUT export
new next
Select the dataset for this query
Dataset: Oryza sativa genes (TIGR3)
Zea mays genes (FGENESH01)
Arabidopsis thaliana genes (TIGR5)
Oryza sativa genes (TIGR3)
Using Mar
After choos LTERS 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.
biomart
refresh Help Desk
Summary
start
Not yet initialised
filter
Not yet initialised
output
Not yet initialised
Site Map | FAQ | Glossary | Documentation | Downloads | Submissions | Mailing List Archive | Feedback | Contact Us | Copyright Statement Back to top
USDA
Done

BioMart databases: other

- dbSNP (via Ensembl)
- Sequence Mart: Ensembl genome sequences

BioMart intro

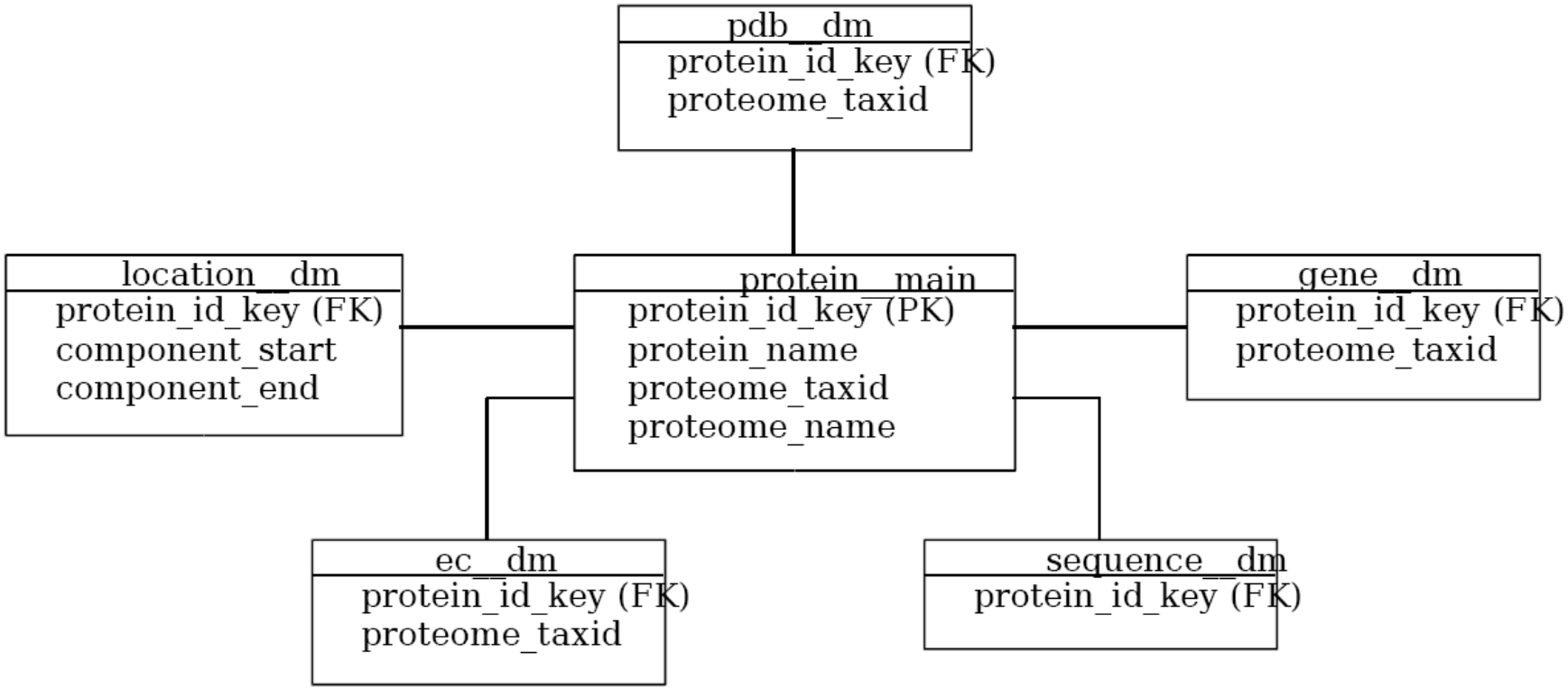
BioMart

- Generic data management system
- Range of advanced query interfaces and administration tools
- Conduct fast and powerful queries using:
  - web
  - graphical or text based applications
  - software libraries written in Perl and Java.

- http://www.ebi.ac.uk/biomart/

BioMart Database Schema's

Simple star-like schema's avoid complex joins and enable fast data retrieval



BioMart user interfaces

MartShell

- MartShell is a command line BioMart user interface based on a structured query language Mart Query Language (MQL)

arek@localhost:~

File Edit View Terminal Go Help

```
[arek@bones bin]$ ./martshell.sh
Starting Interactive MartShell


MartShell: An Interactive User Interface to BioMart databases based on Mart Query Language (MQL)
type 'help' for a list of available commands, or type 'help command' to get help for a particular command.

MartShell> list marts;

ArrayExpress
Ensembl_28
MSD_3
SNP_28
UniProt_13
Vega_28

MartShell> use ArrayExpress.AE1;
MartShell> get experiment_accession, experiment_type ;
E-MEXP-2        compound_treatment_design,time_series_design
E-MEXP-1        time_series_design,compound_treatment_design
E-TOXM-1        compound treatment design,dose response design
E-MEXP-32       disease_state_design
E-MEXP-88       cellular_modification_design
E-MEXP-25       disease_state_design
MartShell>
```

Martview

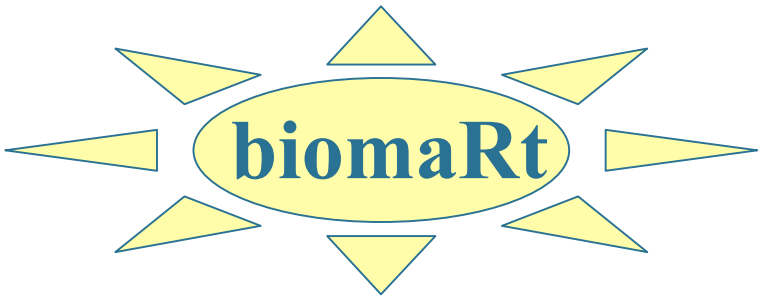
- web based user interface for BioMart.
- Provides functionality for remote users to query all databases hosted by the BioMart server.

Start -> Filter -> output

Other

- MartExplorer
- Perl and Java libraries
- biomaRt interface to R/bioconductor

BioMart and R/Bioconductor



Overview

biomaRt package - BioConductor

- Development started February 2005
- Direct MySQL queries to BioMart systems allow fast data retrieval
- Current BioMarts covered:
  - Ensembl Mart
  - VEGA Mart
  - Sequence Mart
  - SNP Mart

biomaRt - Idea

Integrate public BioMart databases and data analysis in Bioconductor/R

biomaRt - Use

- Annotating genes
- Retrieving GO, OMIM and other information
- Prioritize groups of genes with particular properties
- Data mining

Current biomaRt version

- biomaRt 1.1.8
- http://www.bioconductor.org

  Link: developmental packages

  http://www.ebi.ac.uk/~sdurinck/bioc2005

Installation

- For some users installing biomaRt is a bit difficult as it depends on RMySQL
- Platforms on which biomaRt has been installed:
  - Linux
  - OSX
  - Windows

Installation – Linux, OSX

- Need:
  - DBI
  - RMySQL
  - MySQL client (http://www.mysql.com)

Installation - Windows

- Need:
  - DBI
  - RMySQL (includes libmySQL.dll)
    (http://stat.bell-labs.com/RS-DBI/download/)
  - You have to set the RMySQL/libs directory in your search path or copy dll in your R/bin

MartTable

- Output of most biomaRt functions
  - Slot id: usually contains the query id's
  - Slot table: is a named list containing the information retrieved from the databases

Note: this might be replaced with a common output class for annaffy and biomaRt packages

Connect to BioMart database

- Public BioMart Database

```
> library(biomaRt)
Loading required package:
Biobase Loading required package: tools Welcome to
    Bioconductor Vignettes contain introductory material. To
    view, simply type: openVignette() For details on reading
    vignettes, see the openVignette help page.
 Loading required package: RMySQL Loading required package:
 DBI

> mart <- martConnect()
 - Connected to: ensembl_mart_32 -
```

Local BioMart databases

```
> martConnect( host = "localhost",
                user = "itsme",
                password = "localpasswd",
                local = TRUE)
```

Gene annotation

- biomaRt enables you to get gene annotation for many types of identifiers
- Supported identifiers are:
  - Affy, RefSeq, Entrez-Gene, EMBL, HUGO and Ensembl
  - Soon Agilent identifiers will also be available

Gene annotation

- Note:

  Ensembl does an independent mapping of affy probe sequences to genomes. If there is no clear match then that probe is not assigned to a gene.

Gene annotation

- **getGene** returns a MartTable object containing:
  - Gene symbol
  - Description
  - Chromosome name
  - Band
  - Start position
  - End position
  - BioMartID

Annotation of affy id's

#Assume the following affy id's were found upregulated in our experiment

```
> upregulated <-
  c("210708_x_at","202763_at","211464_x_at")
```

Annotation of affy id's

```
> gene <- getGene( id = upregulated, array =
   "hgu133plus2", mart = mart)
```

```
 > gene
 object of class martTable
 slot id
[1] "210708_x_at" "202763_at"   "211464_x_at"
 slot table
 $symbol
[1] "CASP10" "CASP3"  "CASP6"
```

Annotation of affy id's

$description

[1] "Caspase-10 precursor (EC 3.4.22.-) (CASP-10) (ICE-like apoptotic protease 4) (Apoptotic protease Mch-4) (FAS-associated death domain protein interleukin-1B-converting enzyme 2) (FLICE2). [Source:Uniprot/SWISSPROT;Acc:Q92851]"

[2] "Caspase-3 precursor (EC 3.4.22.-) (CASP-3) (Apopain) (Cysteine protease CPP32) (Yama protein) (CPP-32) (SREBP cleavage activity 1) (SCA-1). [Source:Uniprot/SWISSPROT;Acc:P42574]"

[3] "Caspase-6 precursor (EC 3.4.22.-) (CASP-6) (Apoptotic protease Mch-2). [Source:Uniprot/SWISSPROT;Acc:P55212]"

Annotation of affy id's

```
$band
[1] "q33.1" "q35.1" "q25"

$chromosome
[1] "2" "4" "4"

$start
[1] 201873361 185924000 110967389

$end
[1] 201919616 185945750 110982233

$martID
[1] "ENSG00000003400" "ENSG00000164305" "ENSG00000138794"
```

Annotation of affy id's

```
> getAffyArrays(mart)
```

	V1	V2
1	canine	cfamiliaris
2	zebrafish	drerio
3	hg_focus	hsapiens
4	hg_u133_plus_2	hsapiens
5	hg_u133a_2	hsapiens
6	hg_u133a	hsapiens
7	hg_u133b	hsapiens
8	hg_u95av2	hsapiens
9	hg_u95b	hsapiens
10	hg_u95c	hsapiens
11	hg_u95d	hsapiens
12	hg_u95e	hsapiens
13	u133_x3p	hsapiens

Annotation of affy id's

14	mg_u74av2	mmusculus
15	mg_u74bv2	mmusculus
16	mg_u74cv2	mmusculus
17	mouse430_2	mmusculus
18	mouse430a_2	mmusculus
19	mu11ksuba	mmusculus
20	mu11ksubb	mmusculus
21	rat230_2	rnorvegicus
22	rg_u34a	rnorvegicus
23	rg_u34b	rnorvegicus
24	rg_u34c	rnorvegicus

Annotation of other id's

```
>getGene(id=100,species="hsapiens",type="entrezgene"
  ,mart=mart)

  An object of class "martTable"
  Slot "id":
  [1] "100"

 Slot "table":
 $symbol
 [1] "ADA"

 $description
  [1] "Adenosine deaminase (EC 3.5.4.4) (Adenosine
   aminohydrolase). [Source:Uniprot/SWISSPROT;Acc:P00813]"
```

Annotation of other id's

```
$band
[1] "q13.12"

$chromosome
[1] "20"

$start
[1] 42681578

$end
[1] 42713790

$martID
[1] "ENSG00000196839"
```

Annotation of other id's

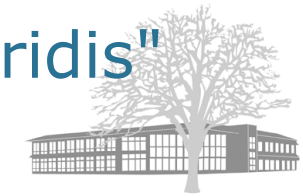
Valid identifier types:

- entrezgene
- hugo
- embl
- ensembl
- affy (no need to specify type as array does this)
- refseq

getSpecies

- Get all species present in Ensembl or VEGA

```
> getSpecies(mart)
 [1] "agambiae"      "amellifera"    "celegans"
"cfamiliaris"
 [5] "cintestinalis" "dmelanogaster" "drerio"
"frubripes"
 [9] "ggallus"       "hsapiens"      "mmusculus"
"ptroglodytes"
[13] "rnorvegicus"   "scerevisiae"   "tnigroviridis"
"xtropicalis"
```



Retrieval of GO information

- The Gene Ontology project provides a controlled vocabulary to describe gene and gene product attributes in any organism.

getGO function

- **getGO** retrieves:
  - GO id
  - GO term
  - Evidence code

GO evidence codes

IMP	Inferred from mutant phenotype
IGI	Inferred from genetic interaction
IPI	Inferred from physical interaction
ISS	Inferred from sequence similarity
IDA	Inferred from direct assay
IEP	Inferred from expression pattern
IEA	Inferred from electronic annotation
TAS	Traceable author statement
NAS	Non-traceable author statement
ND	No biological data available
IC	Inferred by curator

getGO

```
> getGO(id="1939_at",array="hgu95av2",mart=mart)
```

An object of class "martTable"

Slot "id":

```
 [1] "1939_at" "1939_at" "1939_at" "1939_at" "1939_at" "1939_at" "1939_at" "1939_at"
     "1939_at" "1939_at“  "1939_at" "1939_at" "1939_at" "1939_at" "1939_at" "1939_at"
     "1939_at" "1939_at" "1939_at" "1939_at“ "1939_at" "1939_at" "1939_at" "1939_at"
     "1939_at" "1939_at" "1939_at"
```

Slot "table":

SGOID

```
 [1] "GO:0005739" "GO:0005730" "GO:0051262" "GO:0051097" "GO:0046902"
     "GO:0045786" "GO:0030308" "GO:0030154" "GO:0008635" "GO:0008630"
     "GO:0008628" "GO:0008283" "GO:0007569" "GO:0007050" "GO:0006915"
     "GO:0006355" "GO:0006310" "GO:0006289" "GO:0006284" "GO:0000075"
     "GO:0008270" "GO:0005524" "GO:0005515" "GO:0005507" "GO:0004518"
     "GO:0003700" "GO:0000739"
```

getGO

```
> $description
 [1] "mitochondrion"
 [2] "nucleolus"
 [3] "protein tetramerization"
 [4] "negative regulation of helicase activity"
 [5] "regulation of mitochondrial membrane
permeability"
 [6] "negative regulation of cell cycle"
 [7] "negative regulation of cell growth"
 [8] "cell differentiation"
 [9] "caspase activation via cytochrome c"
[10] "DNA damage response, signal transduction
 resulting in induction of apoptosis"
```

getGO

```
$evidence
 [1] "IDA" "IDA" "TAS" "TAS" "TAS" "IEA" "IMP" "TAS"
     "IDA" "TAS"
```

OMIM

- Online Mendelian Inheritance in Man.
- Catalogue of human genes and genetic disorders

OMIM

- **getOMIM** retrieves:
  - OMIM id
  - Disease
  - BioMart id

```
> getOMIM(id="1939_at",array="hgu95av2",
  mart=mart)

  An object of class "martTable"
  Slot "id":
  [1] "1939_at" "1939_at"
  Slot "table":
  $OMIMID
  [1] 191170 191170
  $disease
  [1] "Colorectal cancer, 114500 (3)" "Li-Fraumeni
  syndrome (3)"
  $martID
  [1] "ENSG00000141510" "ENSG00000141510"
```

INTERPRO

- InterPro is an integrated resource of protein families, domains and functional sites.

http://www.ebi.ac.uk/interpro

INTERPRO


UniProt
Universal Protein Resource
prosite
Pfam
PRINTS
Protein Fingerprints Database
ProDom
PANTHER
Classification System


SMART
TIGR
tigr fams
PIRSF
Superfamily
Gene3D
Domain Architecture Classification

INTERPRO

```
>getINTERPRO(id="1939_at",array="hgu95av2",
  mart=mart)

An object of class "martTable"
Slot "id":
[1] "1939_at" "1939_at" "1939_at" "1939_at"

Slot "table":
$INTEPROID
[1] "IPR002117" "IPR011615" "IPR010991" "IPR001472"
```

INTERPRO

```
$short description
[1] "P53"                "p53_DNA_bind"
   "p53_tetrameristn" "NLS_BP"


$description
[1] "p53 tumor antigen"                          "p53, DNA-
   binding"
[3] "p53, tetramerisation"                       "Bipartite
   nuclear localization signal"
```

Sequences

- Genomic sequences of up to 100Mb can be retrieved with the **getSequence** function
- Export sequences in FASTA format with **exportFASTA** function

getSequence

> seq<-getSequence(species="hsapiens", chromosome = 19, start = 18357968, end = 18360987, mart = mart)

Seq
 object of class martTable
slot id
[1] "19_18357968_18360987"
slot table
chromosome
[1] "19"
start
[1] 18357968
end
[1] 18360987
sequence
"AGTCCCAGCTCAGAGCCGCAACCTGCACAGCCATGCCCGGGCAAGAACTCAGGACGGTGAATGGCTCTCAGATGCTCCTGGTG
TTGCTGGTGCTCTCGTGGCTGCCGCATGGGGGCGCCCTGTCTCTGGCCGAGGCGAGCCGCGCAAGTTTCCCGGGACCCTCAGA
GTTGCACTCCGAAGACTCCAGATTCCGAGAGTTGCGGAAACGCTACGAGGACCTGCTAACCAGGCTGCGGGCCAACCAGAGCTG
GGAAGATTCGAACACCGACCTCGTCCCGGCCCCTGCAGTCCGGATACTCACGCCAGAAGGTAAGTGAAATCTTAGAGATCCCCT
CCCACCCCCCAAGCAGCCCCCATATCTAATCAGGGATTCCTCATCTTGAAAAGCCCAGACCTACCTGCGTATCTCTCGGGCCGC
CCTTCCCGAGGGGCTCCCCGAGGCCTCCCGCCTTCACCGGGCTCTGTTCCGGCTGTCCCCGACGGCGTCAAGGTCGTGGGAC
GTGACACGACCGCTGCGGCGTCAGCTCAGCCTTGCAAGACCCCAGGCGCCCGCGCTGCACCTGCGACTGTCGGGGCCGCCGT
CGCAGTCGGACCAACTGCTGGCAGAATCTTCGTCCGCACGGCCCCAGCTGGAGTTGCACTTGCGGCCGCAAGCCGCCAGGGGG

SNP

- Single Nucleotide Polymorphisms (SNPs) are common DNA sequence variations among individuals.

  e.g. A**A**GGCTAA and A**T**GGCTAA

- biomaRt uses the SNP mart of Ensembl which is obtained from dbSNP

getSNP

> snp<-getSNP(species="hsapiens", chromosome = 19, start = 18357968, end = 18360987,mart = mart)

```
> snp
 object of class martTable
 slot table
 $snpStart
[1] 18358024 18358137 18358141 18358162 18358903 18359246 18359591
18359624 [9] 18359718 18359808
$allele
 [1] "C/G" "G/A" "A/T" "G/T" "G/A" "G/A" "T/G" "T/C" "A/G" "T/G"
$coding
[1] 1 1 1 1 NA NA NA NA NA NA
$intronic
[1] NA NA NA NA 1 1 1 1 1 1
$syn
[1] 0 1 0 0 NA NA NA NA NA NA
$utr5
[1] NA NA NA NA NA NA NA NA NA NA
$utr3
[1] NA NA NA NA NA NA NA NA NA NA
```

Homology mapping

The **getHomolog** function enables mapping of many types of identifiers from one species to the same or another type of identifier in another species.

getHomolog

- Example 1:

  from **Entrez-Gene** id in **Homo sapiens** to **RefSeq** id in **Mus musculus**:

```
> getHomolog(id = 2,
                 from.species = 'hsapiens',
                 to.species = 'mmusculus',
                 from.type = 'entrezgene',
                 to.type = 'refseq',
                 mart = mart)
```

getHomolog

An object of class "martTable"

Slot "id":

[1] "2" "2" "2"

Slot "table":

$MappedID

[1] "NM_001013775" "NM_008645"
"NM_007376"

getHomolog

Example 2:

Get homolog for 1939_at from array hgu95av2 as affy id on the canine array

```
> getHomolog(id = "1939_at",
                 from.array = "hgu95av2",
                 to.array = “canine",
                 mart = mart )
```

getHomolog

```
Slot "id":
 [1] "1939_at" "1939_at"

 Slot "table":
 $MappedID
 [1] "1582452_at" "1590246_at"
```

Make subset of genes of interest combined with data analysis

getFeature function

Filter on:

- gene location
- symbol
- OMIM
- GO

getFeature

Select all features which correspond to BRCA2 on affy array hgu95av2:

```
>getFeature(   symbol="BRCA2",
               array="hgu95av2",
               mart=mart)
```

getFeature

```
An object of class "martTable"
Slot "id":
[1] "1990_g_at" "1989_at"

Slot "table":
$symbol
[1] "BRCA2" "BRCA2"

$description
[1] "Breast cancer type 2 susceptibility protein.
   [Source:Uniprot/SWISSPROT;Acc:P51587]"
[2] "Breast cancer type 2 susceptibility protein.
   [Source:Uniprot/SWISSPROT;Acc:P51587]"
```

getFeature

Select all RefSeq id's involved in diabetes mellitus:

```
>getFeature(   OMIM="diabetes mellitus",
               type="refseq",
               species="hsapiens",
               mart=mart)
```

getFeature

```
 An object of class "martTable"
Slot "id":
 [1] "NM_000160" "NM_000207" "NM_001042"
   "NM_000208" "NM_000457" "NM_178850"
 [7] "NM_000408" "NM_002103" "NM_000545"
   "NM_000545" "NM_000340" "NM_005544"


Slot "table":
$OMIMID
 [1] 138033 176730 138190 147670 600281 600281
   138430 138570 142410 142410
[11] 138160 147545
```

getFeature

$description
[1] "Diabetes mellitus, type II (3)"
[2] "Diabetes mellitus, rare form (1)"
[3] "Diabetes mellitus, noninsulin-dependent (3)"
[4] "Diabetes mellitus, insulin-resistant, with acanthosis nigricans (3)"
[5] "Non-insulin-dependent diabetes mellitus, 125853 (3)"
[6] "Non-insulin-dependent diabetes mellitus, 125853 (3)"
...

getFeature

Select all EntrezGene ids located on Y chromosome

```
>getFeature(chromosome="Y",
           type="entrezgene",
           species="hsapiens",
           mart=mart)
```

getFeature

```
An object of class "martTable"
Slot "id":
 [1] "55344"  "8225"   "28227"  "6736" …
Slot "table":
$chromosome
 [1] "Y" "Y" "Y" "Y"  …
$start
 [1]  132996  160025  264972 2698391  …
$end
 [1]  160020  170886  317627 2699005  …
```

getFeature

Select all affy id's of genes located on chromosome 21 between basepair 30Mb and 35Mb

```
> getFeature(     chromosome=21,
                  start=30000000,
                  end = 35000000,
                  array="hgu95av2",
                  mart=mart)
```

getFeature

```
An object of class "martTable"
Slot "id":
 [1] "33610_at"   "33611_g_at" "34559_at"   "37460_at"
    "38370_at“ …
Slot "table":
$chromosome
 [1] "21" "21" "21" "21" "21” …
$start
 [1] 30508196 30508196 30613592 31414352 31414352
    …
$end
 [1] 30510223 30510223 30614224 31853161 31853161
    …
```

getFeature

Select all affy id's from mouse430a2 array that have a GO term attached to it containing "cell cycle"

```
> getFeature(     GO="cell cycle",
                  array="mouse430a2",
                  mart=mart)
```

getFeature

```
Slot id
"1418404_at"   "1417897_at"   "1416206_at"
   "1451803_a_at" "1423092_at" …..
Slot table
$GOID              ….
$description
  [1] "cell cycle checkpoint"
  [2] "negative regulation of cell cycle"
  [3] "negative regulation of cell cycle"
  [4] "regulation of cell cycle"
  [5] "cell cycle"
  [6] "cell cycle"
   …..
```

Ensembl Cross-references

- Powerful function to map between all possible cross-references in Ensembl
- Can also be used to map between different affy arrays within one species

Ensembl Cross-references

- getPossibleXrefs
  - Retrieves all possible cross-references

```
> xref <- getPossibleXrefs(mart = mart)
> xref[1:10, ]
species xref
[1,] "agambiae" "embl"
[2,] "agambiae" "pdb"
[3,] "agambiae" "prediction_sptrembl"
[4,] "agambiae" "protein_id"
[5,] "agambiae" "uniprot_accession"
[6,] "agambiae" "uniprot_id"
```

Ensembl Cross-references

- getXref
  - Retrieves the cross-references

```
> getXref(id = "1939_at", from.species = "hsapiens",
to.species = "hsapiens", from.xref = "affy_hg_u95av2",
to.xref = "affy_hg_u133_plus_2", mart = mart)
```

Ensembl Cross-references

An object of class "martTable"

Slot "id":

[1] "1939_at" "1939_at"

Slot "table":

$from.id

[1] "1939_at" "1939_at"

$to.id

[1] "211300_s_at" "201746_at"

$martID

Comparison to other annotation packages

- biomaRt is complementary to other annotation packages
- biomaRt advantages:
  - Up-to-date data retrieval
  - Comprehensive, covers multiple chips
  - One package to annotate many species

Comparison to other annotation packages

biomaRt disadvantages:

- Possibility that you need to update the package when there is a new release of a BioMart database (Working on using database description in XML to prevent this)
- Reproducibility, annotation might change compared to a previous run
- Need to be online or have local install of the BioMart database of interest.

Future developments

- Retrieve more information such as protein data, gene expression data from Arrayexpress, …
- Include more BioMart databases
- Use of XML database description present in each BioMart database

Exercise

- Select all affy probes from array mouse430a2 that have human homologs that have a known role in hypertension.

Acknowledgements

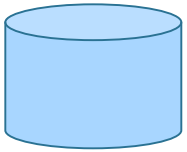
- EBI
  - Wolfgang Huber
  - Arek Kasprzyk
  - Ewan Birney
  - Alvis Brazma
- ESAT-SCD KULeuven
  - Yves Moreau
  - Bart De Moor
- NIH/NHGRI
  -Sean Davis
- Bioconductor users

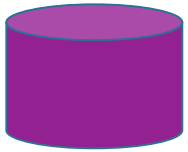
BioMart idea …..

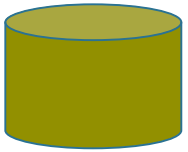
Phenotypes

Expression

Genomes







Give me all genes with Phenotype X

Give me all Genes Expressed in brain

Give me all genes associated with coding SNPs

BioMart integration layer

Give me all genes with Phenotype X, expressed in brain And associated with coding SNPs

Annotation of affy id's

> go<-getGO(id = upregulated[1], array = "hgu133plus2", mart = mart)

Object of class martTable
slot "id"
[1] "210708_x_at" "210708_x_at" "210708_x_at" "210708_x_at" "210708_x_at" "210708_x_at" "210708_x_at" "210708_x_at"

slot "table"
$GOID
[1] "GO:0005515" "GO:0008234" "GO:0030693" "GO:0006508" "GO:0042981" "GO:0030693" "GO:0006917"

$description
[1] "protein binding" "cysteine-type peptidase activity" "caspase activity"
[4] "proteolysis and peptidolysis" "regulation of apoptosis" "caspase activity"
[7] "induction of apoptosis"

$evidence
[1] "IEA" "IEA" "IEA" "IEA" "IEA" "TAS" "TAS"

Annotation of affy id's

```
omim<-getOMIM( id = "203140_at", array = "hgu133plus2", mart = mart)

 > omim
object of class martTable
slot id
[1] "203140_at" "203140_at"

slot table
$OMIMID
[1] 109565 109565

$disease
[1] "Lymphoma, B-cell (2)" [2] "Lymphoma, diffuse large cell (3)"
```