

biomaRt

BioC Developers' Forum - 15/08/2019

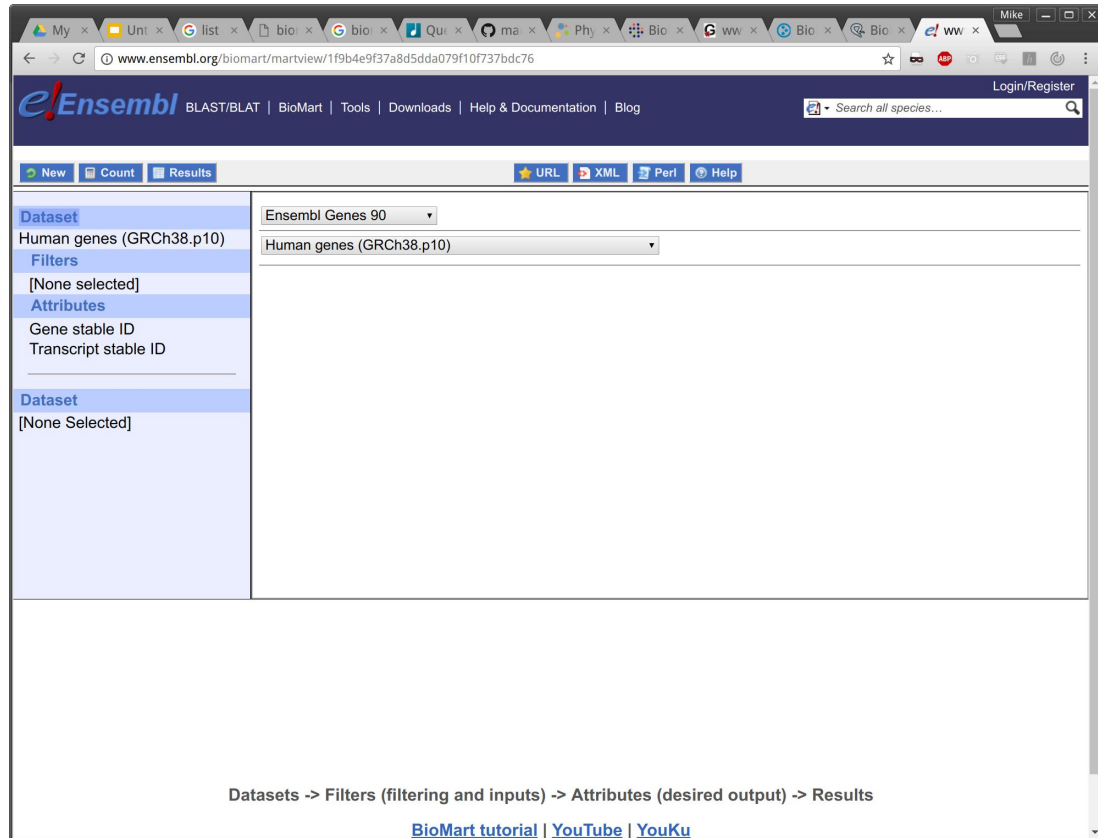
Mike Smith

  @grimbough

BioMart & biomaRt

BioMart

- BioMart & Ensembl



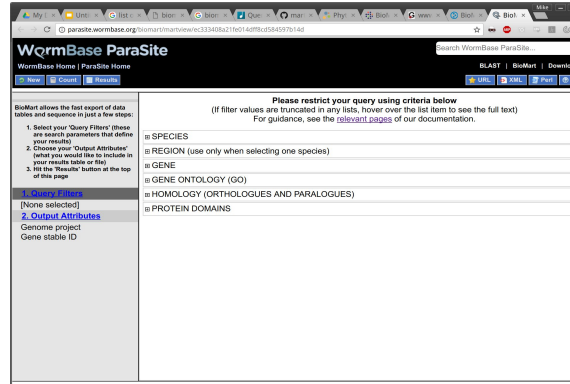
The screenshot shows the Ensembl BioMart interface in a web browser. The URL is www.ensembl.org/biomart/martview/1f9b4e9f37a8d5dda079f10f737bdc76. The page features the Ensembl logo and navigation links for BLAST/BLAT, BioMart, Tools, Downloads, Help & Documentation, and Blog. A search bar is present with the text "Search all species...". Below the navigation, there are buttons for "New", "Count", and "Results". The main interface is divided into several sections:

- Dataset:** A dropdown menu showing "Ensembl Genes 90".
- Human genes (GRCh38.p10):** A dropdown menu showing "Human genes (GRCh38.p10)".
- Filters:** A section with "[None selected]".
- Attributes:** A section with "Gene stable ID" and "Transcript stable ID".
- Dataset:** A section with "[None Selected]".

At the bottom of the interface, there is a summary of the workflow: "Datasets -> Filters (filtering and inputs) -> Attributes (desired output) -> Results". Below this, there are links for "BioMart tutorial", "YouTube", and "YouKu".

BioMart

- BioMart \neq Ensembl



WormBase ParaSite

WormBase Home | ParaSite Home

BLAST | BioMart | Downloads

NEW | COUNT | RESULTS

FAST | XML | PDF | HELP

Please restrict your query using criteria below
(If filter values are truncated in any lists, hover over the list item to see the full text)
For guidance, see the [relevant pages](#) of our documentation.

- SPECIES
- REGION (use only when selecting one species)
- GENE
- GENE ONTOLOGY (GO)
- HOMOLOGY (ORTHOLOGUES AND PARALOGUES)
- PROTEIN DOMAINS

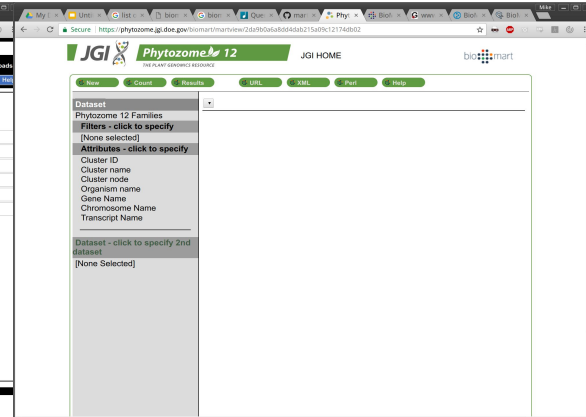
1. Select your Query Filter (these are search parameters that define your results)
2. Choose your 'Output Attributes' (what you would like to include in your results table or file)
3. Hit the 'Search' button at the top of this page

1. Filter

[None selected]

2. Output Attributes

Genome project
Gene stable ID



JGI Phytozome 12 JGI HOME

bio::mart

NEW | COUNT | RESULTS

URL | XML | PDF | HELP

Dataset: Phytozome 12 Families

Filters - click to specify

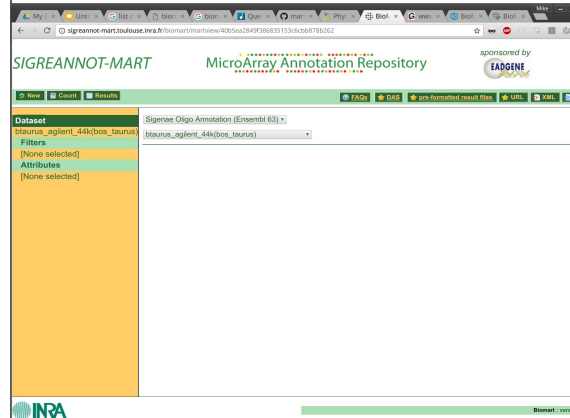
[None selected]

Attributes - click to specify

Cluster ID
Cluster name
Cluster node
Organism name
Gene Name
Chromosome Name
Transcript Name

Dataset - click to specify 2nd dataset

[None Selected]



SIGREANNOT-MART MicroArray Annotation Repository

sponsored by EADGENT

NEW | COUNT | RESULTS

FAST | DAS | GFF3-formatted result files | URL | XML | PDF

Dataset: Sigen Oligo Annotation (Ensembl 63)

blaurus_qlent_44k(bos_taurus)

Filters

[None selected]

Attributes

[None selected]

INRA

BioMart version



CIP INTERNATIONAL POTATO CENTERS

bio::mart

Home

SEARCH

[Go]

(You can search by Accession number, Accession name, Female parent, Male parent, Genus, Species name or Substrain)

DB POTATO

Public Experimental Trials Mapping Population In situ PVS

DB Potato (Public)

DB SWEETPOTATO

Public Experimental Trials Molecular Marker Experiment

DB Sweetpotato (Public)

DB Sweetpotato Gen Index

DB ARTC

DB ARTC (Public)

Disclaimer

This database is compiled from various sources and while great care has been taken in its preparation, CIP cannot guarantee that the information on this database is 100% accurate. Users are advised to use this information at their own risk. CIP declines any responsibility for any damage that arises from use. However, please report any errors or discrepancies to info@cip.org.

This new system is based on the bio::mart database system. You can find help to explain each filter and attribute variable in the data dictionary.

63 Regions

On Map - New Data

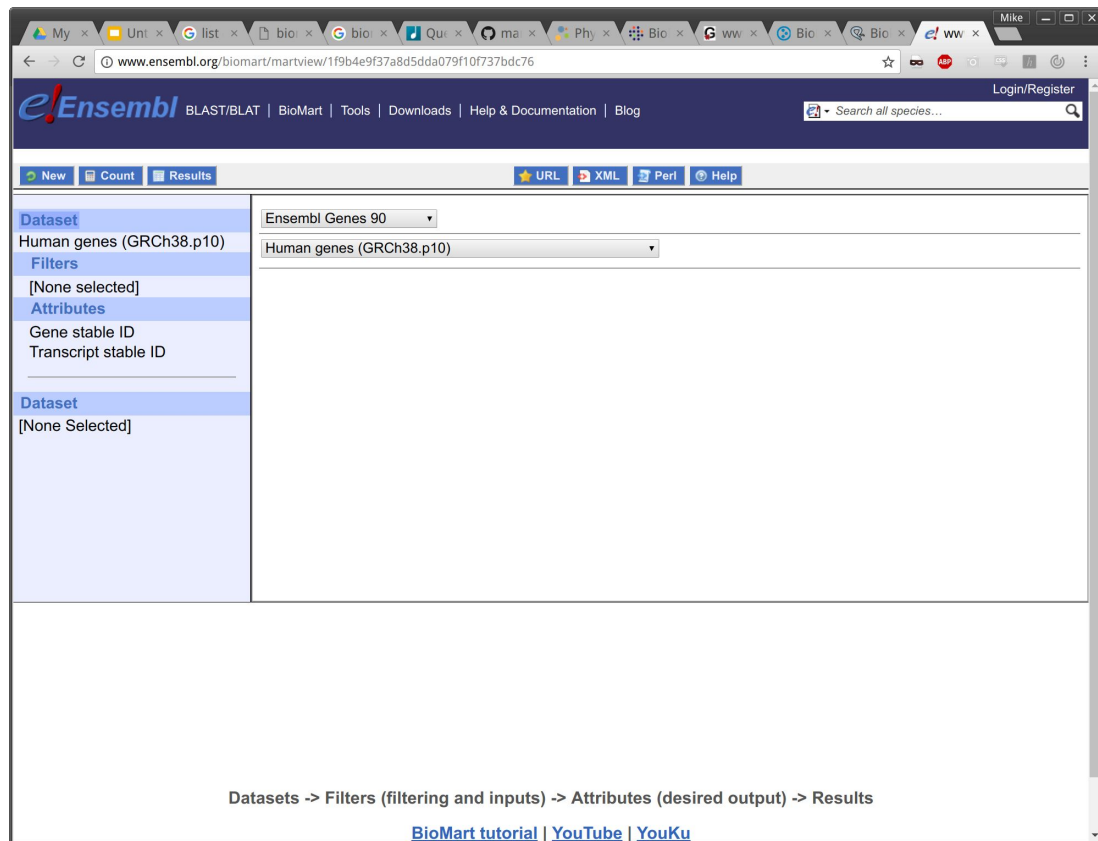
Powered by bio::mart

© 2016 International Potato Center. Av. La Molina 1055, La Molina - Peru.

genplasmdb.cip.cqar.org/index.jsp

BioMart

- BioMart \neq Ensembl
- Vast majority of the time
BioMart = Ensembl
- Existing databases are either:
 - Dieing
 - Being absorbed into Ensembl e.g. COSMIC
 - Moving to other platforms e.g. InterMine

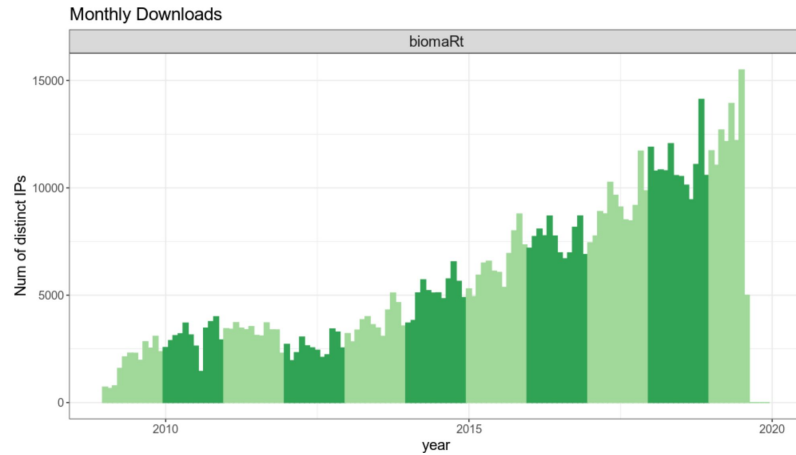


The screenshot shows the Ensembl BioMart interface. The browser address bar displays the URL: www.ensembl.org/biomart/martview/1f9b4e9f37a8d5dda079f10f737bdc76. The Ensembl logo is visible in the top left, and navigation links for BLAST/BLAT, BioMart, Tools, Downloads, Help & Documentation, and Blog are in the top right. A search bar contains the text "Search all species...". Below the navigation bar, there are buttons for "New", "Count", and "Results". A toolbar includes options for "URL", "XML", "Perl", and "Help". The main interface is divided into several sections: "Dataset" (Human genes (GRCh38.p10)), "Filters" ([None selected]), "Attributes" (Gene stable ID, Transcript stable ID), and another "Dataset" section ([None Selected]). The right side of the interface shows a dropdown menu for "Ensembl Genes 90" and another dropdown for "Human genes (GRCh38.p10)". At the bottom, a flow diagram reads: "Datasets -> Filters (filtering and inputs) -> Attributes (desired output) -> Results". Below this, there are links for "BioMart tutorial", "YouTube", and "YouKu".

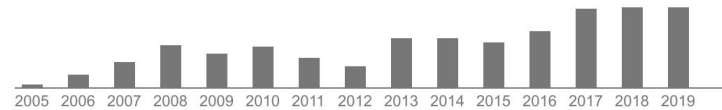
biomaRt

- R package for querying BioMart instance programmatically
 - Originally developed by Steffen Durinck in 2005
- Split between Ensembl specific & generic functionality
- 118 BioC packages list biomaRt in DESCRIPTION

biomaRt is still used

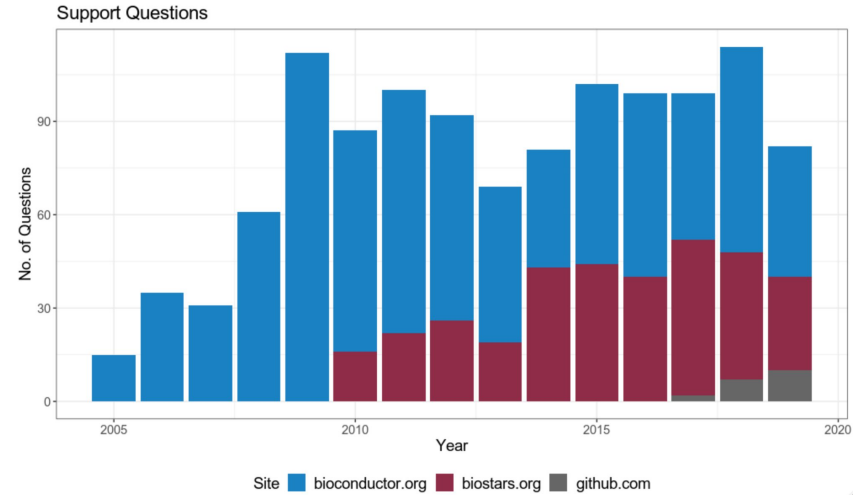


Total citations [Cited by 864](#)

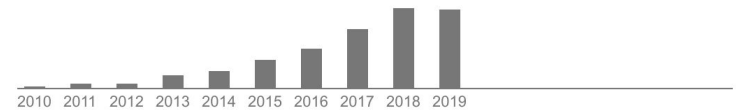


BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis.

S Durinck, *et al* - Bioinformatics, 2005



Total citations [Cited by 803](#)



Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt

S Durinck, *et al* - Nature protocols, 2009

Recent Issues

- Increasing server instability since Ensembl 96 release (Apr 2019)
 - Queries fail with 'Error 500'
 - Larger queries fail more frequently - user has to resume from start
 - Mirror sites generally more responsive
- Unexpect renaming of attributes e.g. `entrezgene` to `entrezgene_id`
- These affect downstream packages too

Package changes

- Queries are broken down into batches, results of each batch written to tempdir
 - Re-running query picks up where it left of
 - Not persistent
- Results of complete queries cached using BiocFileCache
 - Identified via hash based on query parameters / function arguments
- Re-enable automatic mirror redirection
 - If a host is specified this is obeyed
- Doesn't solve build failures - will the cache remain between builds?

Conversations with Ensembl

- Problems affecting web interface too
- Infrastructure struggling with scaling up number of genomes
- BioMart is effectively dead - no updates here
- 'Massive' R queries cause issues
 - IPs banned
 - Unclear to me how frequent this is vs other scaling problems
- Replacement is 'coming' - Spring 2020
 - I've been hearing this for a long time....
- Hopefully will know about variable renaming before release time!

Suggestions for the mid-term

- Impose hard or soft limits on query size?
- Actively promote other resources?
 - ensembl db, org.*.db packages, Ensembl resources e.g. REST API, VEP, Bulk download
 - How/Where? Messages, vignette?
- Use other resources ‘under the hood’?
- Develop something new for most common use cases?



Mike Smith
@grimhough

Do you use [@ensembl](#) BioMart? Thinking about the future updates for **biomaRt** [@Bioconductor](#) package and trying to get a feel for the most popular use cases. Based on support questions here are my feelings on common tasks. Do you do these? No? Please reply with your examples



135 votes · Final results

11:29 am · 7 May 2019 · [Twitter Web Client](#)