Training materials

- Ensembl training materials are protected by a CC BY license:

  creativecommons.org/licenses/by/4.0/

- If you wish to re-use these materials, please credit Ensembl for their creation
- If you use Ensembl for your work, please cite our papers:

  ensembl.org/info/about/publications.html

Mining e!Ensembl data with BioMart and VEP

Michal Szpak
Ensembl Outreach Officer

13.08.2020

EMBL-EBI

This workshop

1. Introduction to Ensembl
2. BioMart
   - Introduction
   - Demo
3. Variant Effect Predictor (VEP)
   - Introduction
   - Demo

e!Ensembl's BioMart

EMBL-EBI

What is BioMart?

A tool in your browser to:

- Export data with no programming required
- Build queries with a few mouse clicks
- Generate custom data tables and sequence files

Why use BioMart?

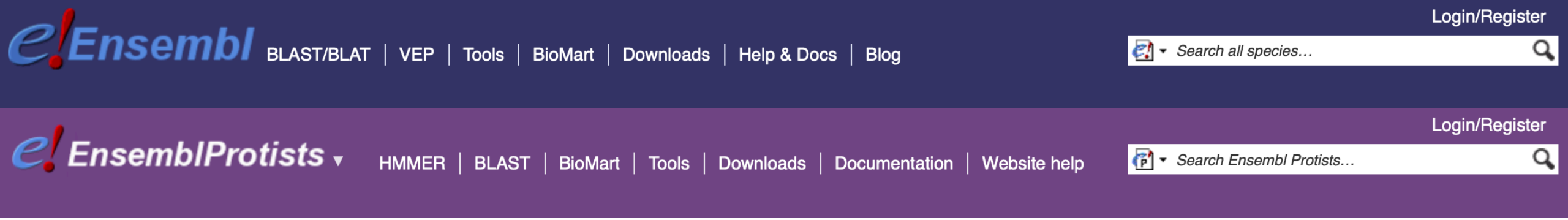
Why use BioMart?

EMBL-EBI

Why use BioMart?

For things that would be time consuming/difficult with the Ensembl browser:

- Query multiple things at once:
  - ID conversions
  - Gene locations
  - Download sequences
- Export large amounts of data



Where can I find BioMart?

ensembl.org/biomart/martview

protists.ensembl.org/biomart/martview

For which genomes is BioMart available?

- Ensembl
  44,048 genomes
- EnsemblFungi
  (some exceptions)
- EnsemblMetazoa
  78 genomes
- EnsemblPlants
  67 genomes
- EnsemblProtists
  (some exceptions)

How do I use BioMart?

The four steps

Dataset	Filters	Attributes	Results
Choose database & species	Narrow down the dataset	Specify your output	View & export

How do I use BioMart?
1. Dataset

Dataset

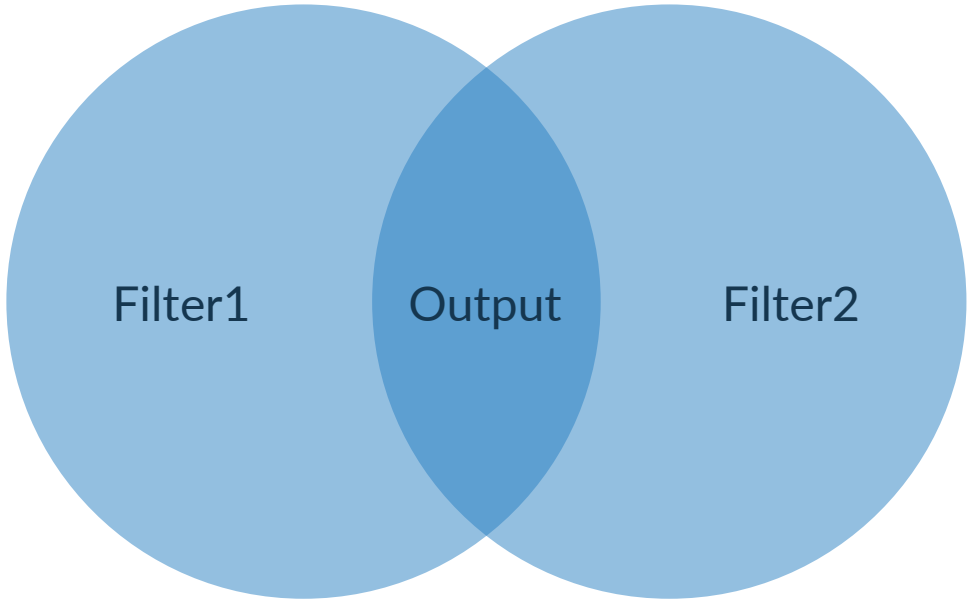
- Define the database you want to search with your filters:
  - Genes
  - Variation
  - Regulation
  - Mouse strains (genes)
- Define the species

How do I use BioMart?
2. Filters

Define a (large) set of genes/variants by combination of filters, e.g.:

- Region
- List of IDs
- Function (GO term)
- Phenotype

How do I use BioMart?
2. Filters



How do I use BioMart?
3. Attributes

Define the data you want to export (your output), e.g.:

- IDs
- Features
- Variants
- Orthologues/paralogues
- Sequences

How do I use BioMart?
4. Results

View and export data table/sequence in a number of formats:

- html
- csv
- tsv
- xls
- fasta

biomaRt: Bioconductor package for BioMart

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data using the R statistical programming language

- Easy installation:

```
source("http://bioconductor.org/biocLite.R")
biocLite("biomaRt")
```

More information

Kinsella *et al.*

Ensembl BioMarts: a hub for data retrieval across taxonomic space

Database: the Journal of Biological Databases and Curation (2011)

europepmc.org/abstract/MED/21785142

Smedley *et al.*

BioMart – biological queries made easy

BMC Genomics (2009) 10:22

europepmc.org/abstract/MED/19144180

Hands on

We're going to look at a set of six human genes, *ESPN*, *MYH9*, *USH1C*, *CISD2*, *THRB* and *WHRN*, and find out:

- Their NCBI gene IDs
- Their function via GO terms
- Their cDNA sequences

Dataset	Filters	Attributes	Results
Genes *Homo sapiens*	Gene names	NCBI IDs GO terms cDNA seq Gene names!	View & export

Variant Effect Predictor VeP

EMBL-EBI

What does the VEP do?

A tool to predict and annotate the functional consequences of variants
(SNPs, insertions, deletions, structural variants)

Input:

- Variant coordinates
- VCF
- Variant IDs
- HGVS
- SPDI

VeP
e!Ensembl

Output:

- Affected gene, transcript and protein sequence
- Pathogenicity
- Frequency data
- Regulatory consequences
- Splicing consequences
- Literature citations

Variant coordinates (Ensembl default)

```
1 182712 182712 A/C 1
2 265023 265023 C/T 1
3 319781 319781 A/- 1
19 110748 110747 -/T 1
```

VCF

```
#CHROM POS     ID        REF ALT
20     14370   rs6054257 G     A
20     17330   .         T     A
20     1110696 rs6040355 A     G,T
20     1230237 .         T     .
```

Variant IDs

```
rs41293501
COSM327779
FANCD1:c.475G>A
rs373400041
```

HGVS notation

```
AGT:c.803T>C
9:g.22125504G>C
ENST00000003084:c.1431_1433delTTC
19:g.110747_110748insT
```

SPDI

```
NC_000001.11:65567:A:C
NC_000002.12:265022:C:T
NC_000003.12:319780:A:
```

VEP features

Transcript consequence predictions:

- Ensembl / GENCODE
- GENCODE basic
- RefSeq
- Merged

Overlap with regulatory regions:

- ENCODE
- BLUEPRINT
- NIH Epigenomics Roadmap
- Can be limited to specific cell types

Known variants:

- dbSNP
- COSMIC
- ClinVar
- ESP
- HGMD-Public
- PhenCode

Allele frequencies:

- 1000 Genomes
- ESP
- gnomAD

Pathogenicity predictions:

- SIFT
- PolyPhen
- dbNSFP
- CADD
- Condel
- LoFTool
- FATHMM
- Mutation-Taster
- Many more!

Phenotype, disease, clinical significance:

- OMIM
- Orphanet
- GWAS catalog
- ClinVar

VEP plugins

- Add extra functionality
- Can extend, filter or manipulate the output
- May make use of external data or code
- Available on the web tool and with the script
- Many plugins for pathogenicity predictions

https://github.com/Ensembl/VEP_plugins
https://www.ensembl.org/info/docs/tools/vep/script/vep_plugins.html#plugins_existing

Using the VEP

- Clinical diagnostics
- Industry
- International consortia

SickKids illumina FABRIC GENOMICS CANCER RESEARCH UK Personalis NHS CONGENICA GENOME BASED MEDICINE

Web interface

- Point-and-click interface
- Suits smaller volumes of data

Documentation

Launch VeP

Command line tool

- More options and flexibility
- For large volumes of data

Documentation

Clone from GitHub

Download (zip)

Pull Docker image from DockerHub

REST API

- Language-independent API
- Simple URL-based queries

Documentation

VEP REST API

Set up a cache

- Speed up your VEP script with an offline cache
- Keep your data private
- Use prebuilt caches for Ensembl species
- Or make your own cache from GTF and FASTA files – even for genomes not in Ensembl!

More information

McLaren *et al.*

The Ensembl Variant Effect Predictor

Genome Biology (2016) 17:122

europepmc.org/abstract/MED/27268795

Hands on

We're going to look at a set of four variants to find out what genes they hit and what effect they have on them:

```
9 128328461 128328461 A/- + var1
9 128322349 128322349 C/A + var2
9 128323079 128323079 C/G + var3
9 128322917 128322917 G/A + var4
```

Help and docs

Courses online:
www.ebi.ac.uk/training/online/subjects/11

Tutorials:
www.ensembl.org/info/website/tutorials

Animations:
www.youtube.com/user/EnsemblHelpdesk
u.youku.com/Ensemblhelpdesk

Email us: helpdesk@ensembl.org

Public mailing lists: dev@ensembl.org, announce@ensembl.org

Follow us

www.facebook.com/Ensembl.org

@Ensembl @Ensemblgenomes

www.ensembl.info

Publications

List of publications:

ensembl.org/info/about/publications.html

Yates *et al.*

Ensembl 2020

Nucleic Acids Research (2020) 48(D1):D682-D688

europepmc.org/abstract/MED/31691826

Ensembl 2019

EMBL-EBI

Acknowledgements:
the entire Ensembl team

Adam Frankish, Ahamed Imran Abdul Salam, Alessandro Vullo, Alexandra Bignell, Ameya Chaubal, Amonida Zadissa, Andrea Winterbottom, Andrew Berry, Andrew Parton, Andrey Azov, Andy Yates, Anja Thormann, Anmol Jaywant Hemrom, Anne Lyle, Anne Parker, Aoife McMahon, Astrid Gall, Benjamin Moore, Bethany Flint, Bianca Schmitt, Brandon Walts, Bruno Contreras-Moreira, Carla Cummins, Carlos Garcia Giron, Claire Davidson, Cristina Guijarro, Dan Sheppard, Daniel Zerbino, David Thybert, Denye Ogeh, Diana Lemos, Elizabeth Lewis, Emily Perry, Erin Haskell, Fergal Martin, Fiona Cunningham, Gareth Maslen, Gareth Williams, Garth Ilsley, Glen Threadgold, Guy Naamati, Helen Schuilenburg, IF Barnes, Ilias Lavidas, Irina Armean, James Allen, Jamie Allen, Jane Loveland, Joannella Morales, Jonathan Mudge, Jorge Alvarez-Jarreta, Jose Carlos Marugan, Jose Manuel Gonzalez Martinez, Jyothish Bhai, Kamalkumar Jayantilal Dodiya, Kevin Howe, Kieron Taylor, Kostas Billis, Lahcen Campbell, Laurent Gil, Leanne Haggerty, Luca Da Rin Fioretto, Magali Ruffier, Manoj Sakthivel, Manuel Carbajo Martinez, Marc Chakiachvili, Marek Szuba, Marie-Marthe Suner, Mark McDowall, Mateus Patricio, Matthew Hardy, Matthew Russell, Matthieu Barba, Matthieu Muffato, Michael Paulini, Michael Nuhn, Michal Szpak, Mike Kay, Mikkel Christensen, Mira Sycheva, Myrto Kostadima, Nick Langridge, Nishadi De Silva, Osagie Izuogu, Paul Davis, Paul Flicek, Premanand Achuthan, Reham Fatima, Ridwan Amode, Ruth Bennett, Sanjay Boddu, Sarah Donaldson, Sarah Hunt, Shamika Mohanan, Sophie Janacek, Stephen Trevanion, Thibaut Hourlier, Thomas Juettemann, Thomas Maurel, Tiago Grego, Toby Hunt, Tuan Le, Vasili Sitnik, Wasiu Akanni

Funding:

BBSRC

wellcome

Co-funded by the European Union

Open Targets

TRANSFORMING GENETIC MEDICINE INITIATIVE

SAVE THE TASMANIAN DEVIL PROGRAM

National Human Genome Research Institute

EMBL-EBI

Training materials

- Ensembl training materials are protected by a CC BY license:

  creativecommons.org/licenses/by/4.0/

- If you wish to re-use these materials, please credit Ensembl for their creation
- If you use Ensembl for your work, please cite our papers:

  ensembl.org/info/about/publications.html