Browsing Genomic Information with Ensembl

Stefan Wyder

April 2016

Universität Zürich[UZH]

URPP Evolution in Action

Objectives

- What is Ensembl?
- How to navigate the Ensembl browser
- What type of data can you get in Ensembl?
- How to 'data mine' Ensembl
- Where to go for help and documentation

Why do we need genome browsers?

CGGCCTTTGGGCTCCGCCTTCAGCTCAAGACTTAACTTCCCTCCCAGCTGTCCCAGATGACGCCATCTGAAATTTCTTGGAA
ACACGATCACTTTAACGGAATATTGCTGTTTTGGGGAAGTGTTTTACAGCTGCTGGGCACGCTGTATTTGCCTTACTTAAGC
CCCTGGTAATTGCTGTATTCCGAAGACATGCTGATGGGAATTACCAGGCGGCGTTGGTCTCTAACTGGAGCCCTCTGTCCCC
ACTAGCCACGCGTCACTGGTTAGCGTGATTGAAACTAAATCGTATGAAAATCCTCTTCTCTAGTCGCACTAGCCACGTTTCG
AGTGCTTAATGTGGCTAGTGGCACCGGTTTGGACAGCACAGCTGTAAAATGTTCCCATCCTCACAGTAAGCTGTTACCGTTC
CAGGAGATGGGACTGAATTAGAATTCAAACAAATTTTCCAGCGCTTCTGAGTTTTACCTCAGTCACATAATAAGGAATGCAT
CCCTGTGTAAGTGCATTTTGGTCTTCTGTTTTGCAGACTTATTTACCAAGCATTGGAGGAATATCGTAGGTAAAAATGCCTA
TTGGATCCAAAGAGAGGCCAACATTTTTTGAAATTTTTAAGACACGCTGCAACAAAGCAGGTATTGACAAATTTTATATAAC
TTTATAAATTACACCGAGAAAGTGTTTTCTAAAAAATGCTTGCTAAAAACCCAGTACGTCACAGTGTTGCTTAGAACCATAA
ACTGTTCCTTATGTGTGTATAAATCCAGTTAACAACATAATCATCGTTTGCAGGTTAACCACATGATAAATATAGAACGTCT
AGTGGATAAAGAGGAAACTGGCCCCTTGACTAGCAGTAGGAACAATTACTAACAAATCAGAAGCATTAATGTTACTTTATGG
CAGAAGTTGTCCAACTTTTTGG GAGTGCTTTTGTTTATG
TAGCTTACCATATTAGAAATTT ATCCCAGCACTTTGGGA
GGCCGAGGTGGGCGGATCACTT CTATCTCTACTAAAAAT
ACAAAAAATGTGCTGCGTGTGG GAGAATCGCTTGAACCC
TGGAGGCAGAGGTTGCAGTGAG ACTCTGTCTCAAAACAA
ACAAACAAACAAAAAACTAAGA GCATATTATCACAACAT
TCTTAGGAAAAATAACTTTTTGAAAACAAGTGAGTGGAATAGTTTTTACATTTTTGCAGTTCTCTTTAATGTCTGGCTAAAT
AGAGATAGCTGGATTCACTTATCTGTGTCTAATCTGTTATTTTGGTAGAAGTATGTGAAAAAAAATTAACCTCACGTTGAAA
AAAGGAATATTTTAATAGTTTTCAGTTACTTTTTGGTATTTTTCCTTGTACTTTGCATAGATTTTTCAAAGATCTAATAGAT
ATACCATAGGTCTTTCCCATGTCGCAACATCATGCAGTGATTATTTGGAAGATAGTGGTGTTCTGAATTATACAAAGTTTCC
AAATATTGATAAATTGCATTAAACTATTTTAAAAATCTCATTCATTAATACCACCATGGATGTCAGAAAAGTCTTTTAAGAT
TGGGTAGAAATGAGCCACTGGAAATTCTAATTTTCATTTGAAAGTTCACATTTTGTCATTGACAACAAACTGTTTTCCTTGC
AGCAACAAGATCACTTCATTGATTTGTGAGAAAATGTCTACCAAATTATTTAAGTTGAAATAACTTTGTCAGCTGTTCTTTC
AAGTAAAAATGACTTTTCATTGAAAAAATTGCTTGTTCAGATCACAGCTCAACATGAGTGCTTTTCTAGGCAGTATTGTACT
TCAGTATGCAGAAGTGCTTTATGTATGCTTCCTATTTTGTCAGAGATTATTAAAAGAAGTGCTAAAGCATTGAGCTTCGAAA
TTAATTTTTACTGCTTCATTAGGACATTCTTACATTAAACTGGCATTATTATTACTATTATTTTTAACAAGGACACTCAGTG
GTAAGGAATATAATGGCTACTAGTATTAGTTTGGTGCCACTGCCATAACTCATGCAAATGTGCCAGCAGTTTTACCCAGCAT

Large amounts of raw DNA sequence data

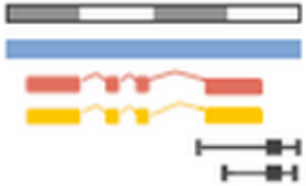
We need to make the data mean something…

http://ensembl.org

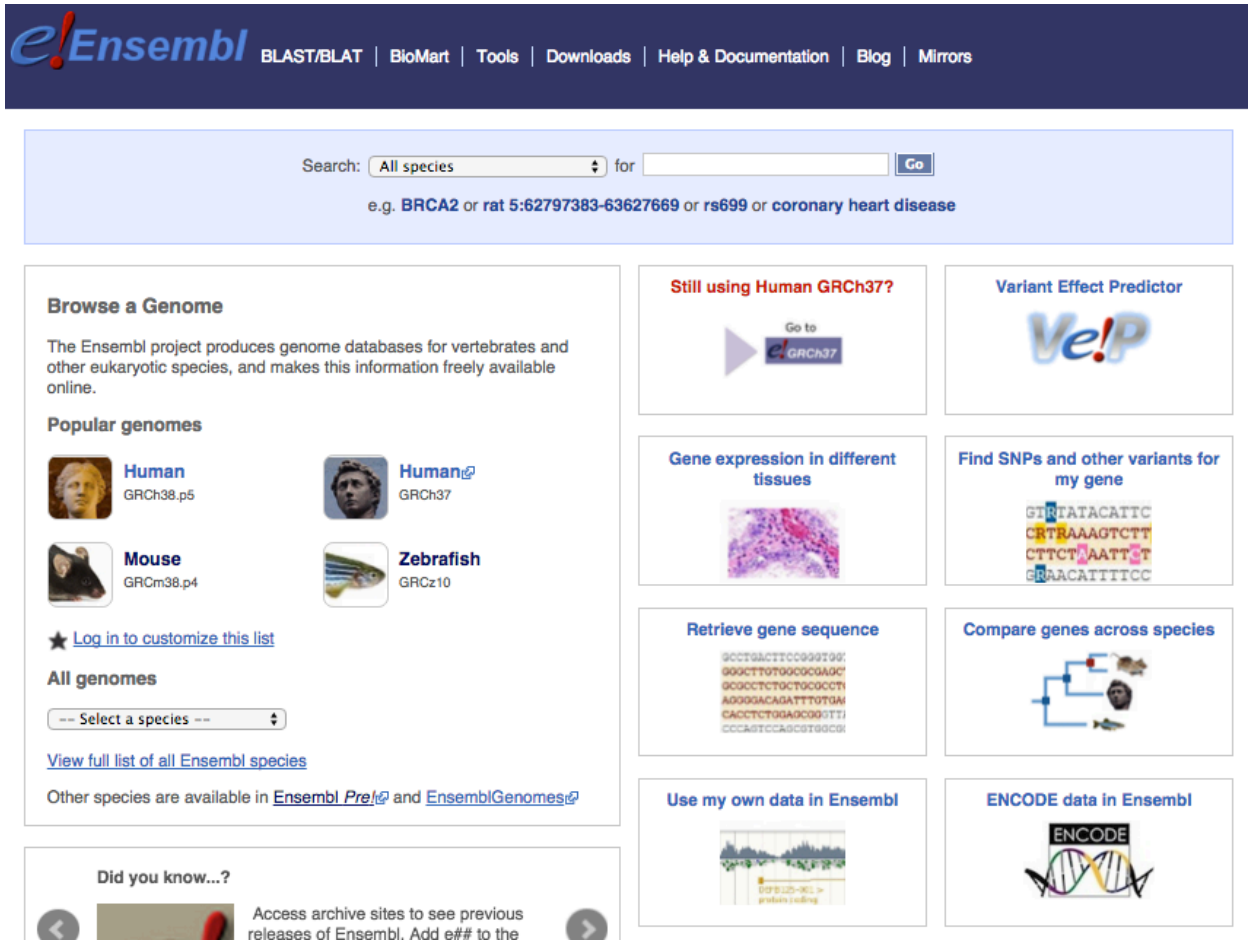
http://genome.ucsc.edu

Ensembl features

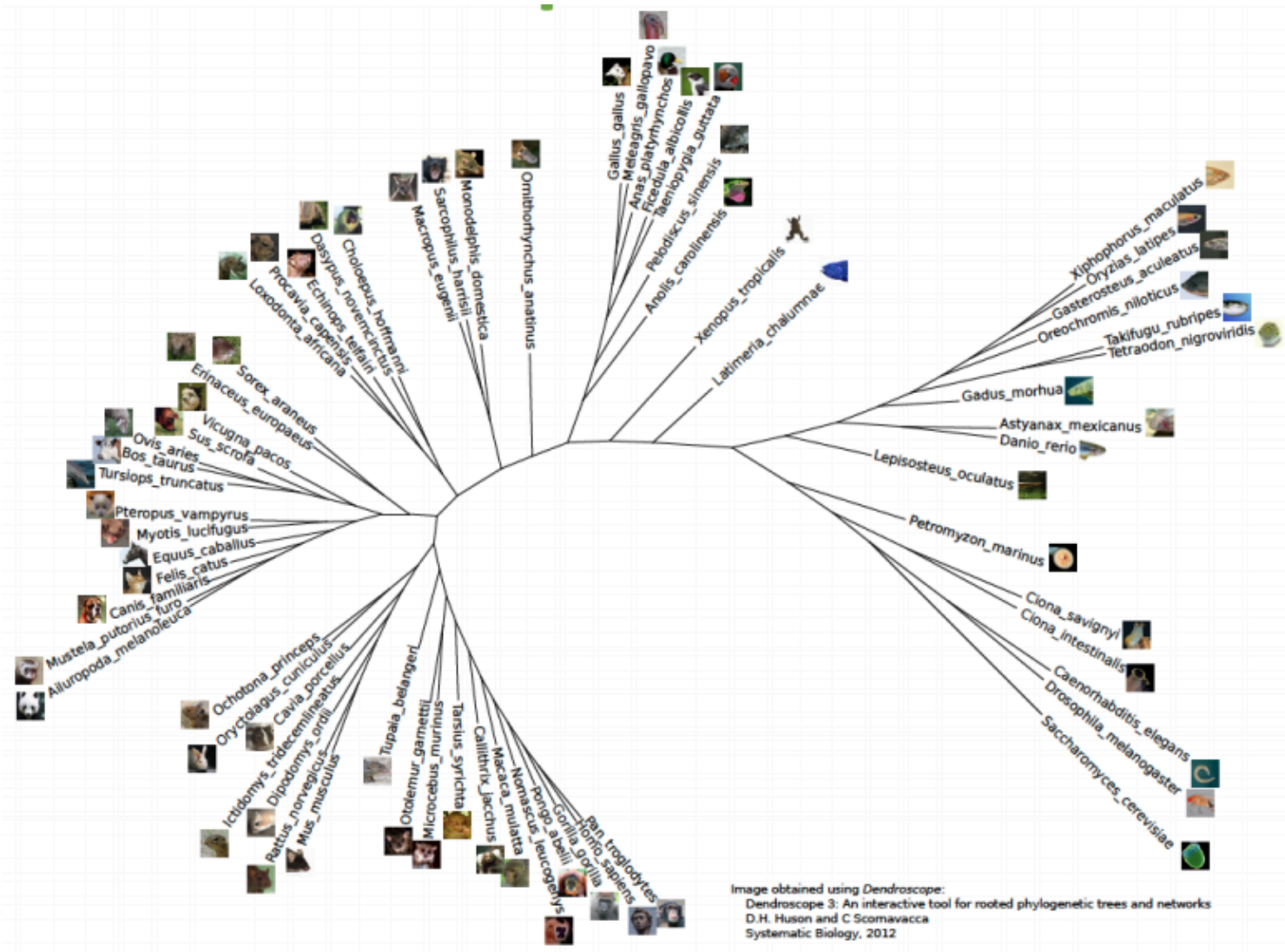
- Gene builds
- Gene trees
- Variation display and VEP
- Display of user data
- BioMart (data export)
- Programmatic access via API
- Completely open source



Exploring the Ensembl Genome Browser

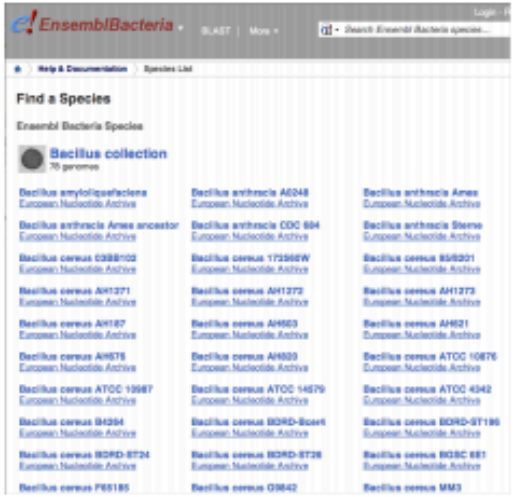


87 Vertebrate Species in Ensembl v84

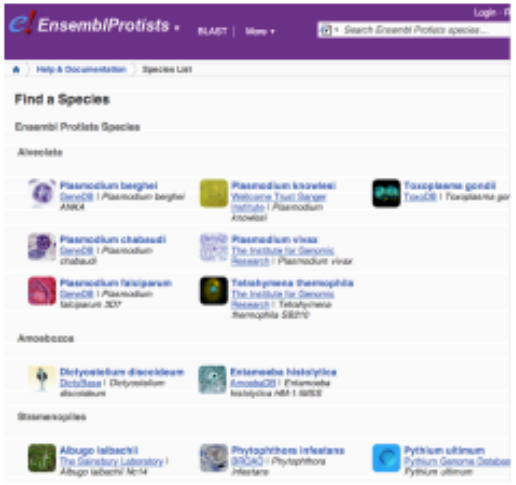


+ early access Pre! websites

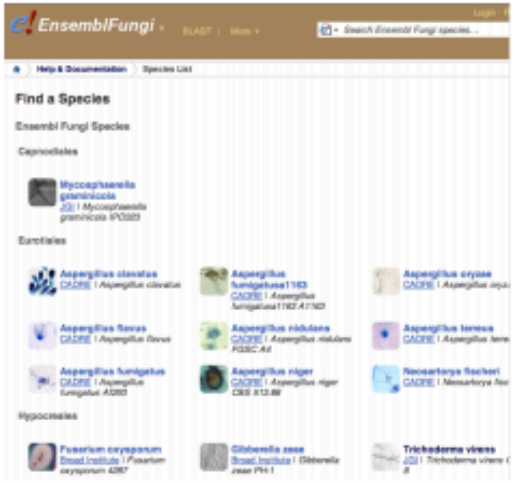
Non-vertebrates on Ensembl Genomes



Bacteria (39,584 genomes)

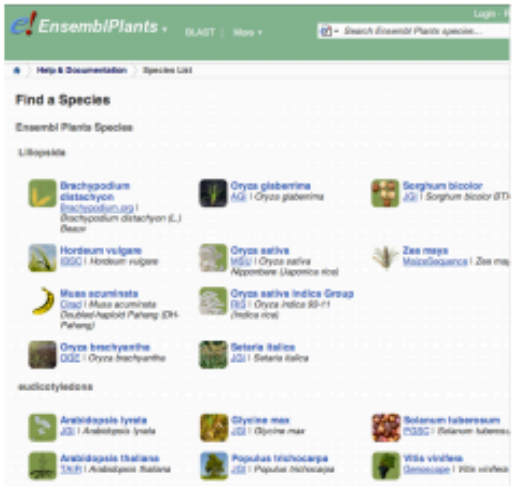


Protists (158)



Fungi (589)

Metazoa (65)



Plants (39)

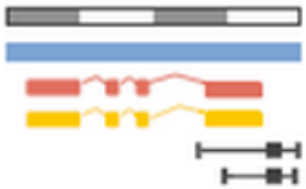
Ensembl vs Ensembl Genomes

	Ensembl	EnsemblGenomes
Released	2000	2009
Species	Vertebrates (fly, worm and yeast as outgroups)	Non-vertebrates (protists, plants, fungi, metazoa, bacteria)
Annotation	by Ensembl	in collaboration with the scientific communities
URL	www.ensembl.org	www.ensemblgenomes.org

- Joint project between EMBL-EBI and Sanger
- Funded primarily by the Wellcome Trust

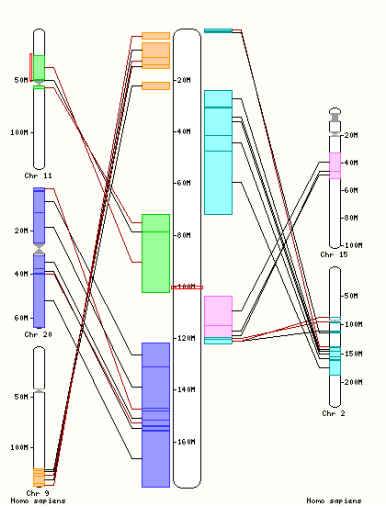
Data

- Genomic sequence
- Gene / transcript / protein models
- External references
- Mapped sequences
  - cDNAs, proteins, repeats, markers, probes, etc.
- Variation data:
  - sequence variants
  - structural variants (selected species)



Data

- Comparative data:
  - Orthologues and paralogues (between plants and pan-taxonomic)
  - Protein families
  - Whole genome pairwise alignments (selected species)
  - Synteny (selected species)
  - whole genome multiple alignment



Synteny between Rat chromosome 3 and Human

Methods in gene annotation

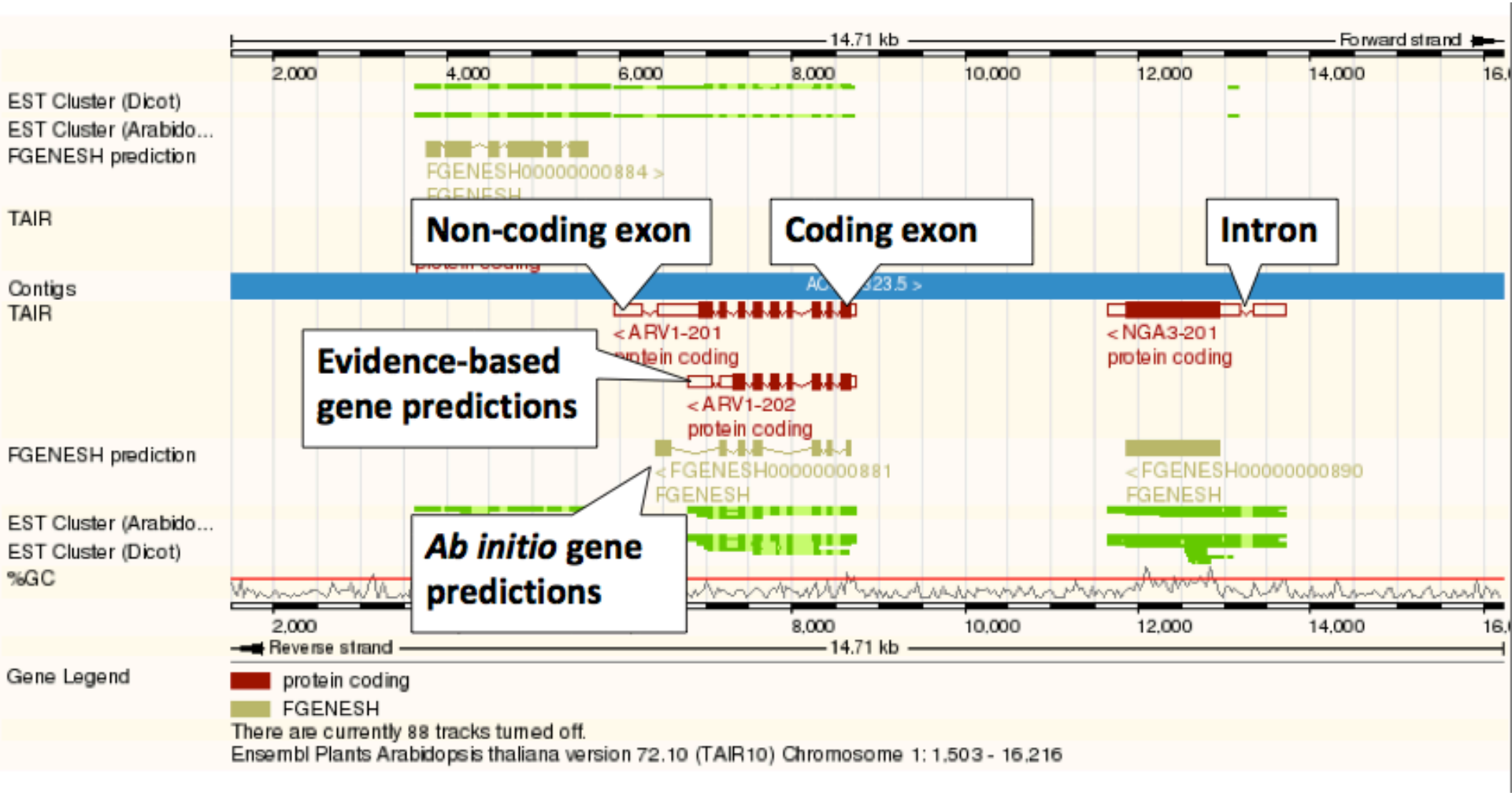
Automatic annotation
Using known proteins/ESTs/cDNAs

Homology annotation
Using known proteins/ESTs/cDNAs from other species

Manual annotation
To correct errors in automatic annotation

***Ab initio* predictions**
(finding apparent ORFs in the sequence)

Gene view


14.71 kb
Forward strand
2,000
4,000
6,000
8,000
10,000
12,000
14,000
EST Cluster (Dicot)
EST Cluster (Arabido...
FGENESH prediction
FGENESH00000000884 >
TAIR
Non-coding exon
Coding exon
Intron
Contigs
TAIR
< ARV1-201
protein coding
< NGA3-201
protein coding
Evidence-based gene predictions
< ARV1-202
protein coding
FGENESH prediction
< FGENESH00000000881
FGENESH
< FGENESH00000000890
FGENESH
EST Cluster (Arabido...
EST Cluster (Dicot)
%GC
Ab initio gene predictions
Reverse strand
Gene Legend
protein coding
FGENESH
There are currently 88 tracks turned off.
Ensembl Plants Arabidopsis thaliana version 72.10 (TAIR10) Chromosome 1: 1,503 - 16,216

Transcript view

Arabidopsis thaliana ▼ | Location: 5:17,164,141-17,165,918 | Gene: DFR | Transcript: AT5G42800.1

Transcript tab

Transcript-based displays
- Transcript summary
- Supporting evidence
- Sequence
  - Exons (6)
  - cDNA
  - Protein
- EBI Protein Summary
- Protein Structure
- External References
  - General identifiers (142)
  - Oligo probes (2)
- Ontology
  - Ontology graphs
    - plant anatomical entity (15
    - biological process (4)
    - plant structure developmer
    - cellular component (1)
    - molecular function (6)
  - Ontology table (37)
- Genetic Variation
  - Variation Table
  - Population comparison
  - Comparison image
- Protein Information
  - Protein summary
  - Domains & features (5)
  - Variations (71)
- External Data
  - Personal annotation
- ID History
  - Transcript history
  - Protein history

Configure this page

Manage your data

Export data

Bookmark this page

Changed side menu

Description dihydroflavonol-4-reductase Source: EMB

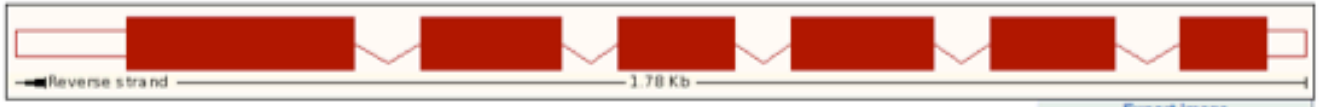
Location Chromosome 5: 17,164,141-17,165,91

Gene This transcript is a product of gene AT5G42800 - This gene has 1 transcript

Show/hide columns | Filter

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype
AT5G42800.1	AT5G42800.1	1358	AT5G42800.1	382	Protein coding

Transcript summary help



Export image

Statistics **Exons:** 6 **Transcript length:** 1,358 bps **Translation length:** 382 residues

Ensembl version AT5G42800.1.1

Type Known protein coding

Prediction Method Gene annotation by TAIR through a process of automatic and manual curation.

Ensembl Plants release 14 - May 2012 © EBI

About Ensembl Genomes | Contact Us | EMBL-EBI Terms of use | Privacy | Cookies | Help

Names in Ensembl

- ENSG### Ensembl Gene ID
- ENST### Ensembl Transcript ID
- ENSP### Ensembl Peptide ID
- ENSE### Ensembl Exon ID

- For other species than human a suffix is added:

  MUS (*Mus musculus*) for mouse: ENSMUSG###
  DAR (*Danio rerio*) for zebrafish: ENSDARG###, etc.

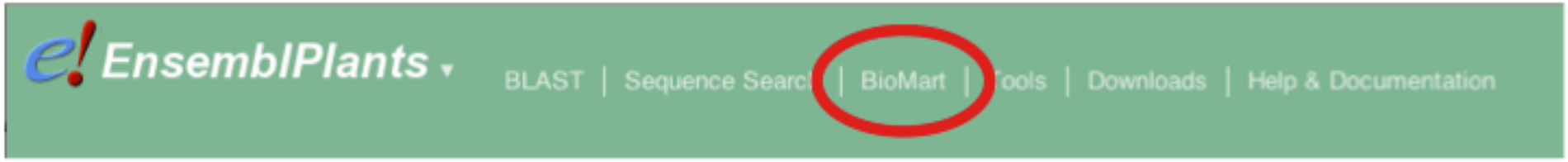
- Imported gene annotations keep the IDs (e.g. Arabidopsis, fly, C. elegans)

Access to data

- Web browser
  - http://ensemblgenomes.org/
  - http://ensembl.org
- BioMart 'Data mining tool'
  - http://ensembl.org/biomart/martview/
- FTP download site
  - ftp://ftp.ensemblgenomes.org/pub/
  - http://ensembl.org/info/data/ftp/
- Public MySQL server
  - mysql.ebi.ac.uk:4157:anonymous
- Ensembl APIs/ via your favourite programming language
  - http://ensembl.org/info/docs/api/
  - http://rest.ensembl.org/

BioMart

- Data export tool
- Quick table generator (e.g. Excel)
- Web interface to mine ensembl data
- Integrated with BioConductor

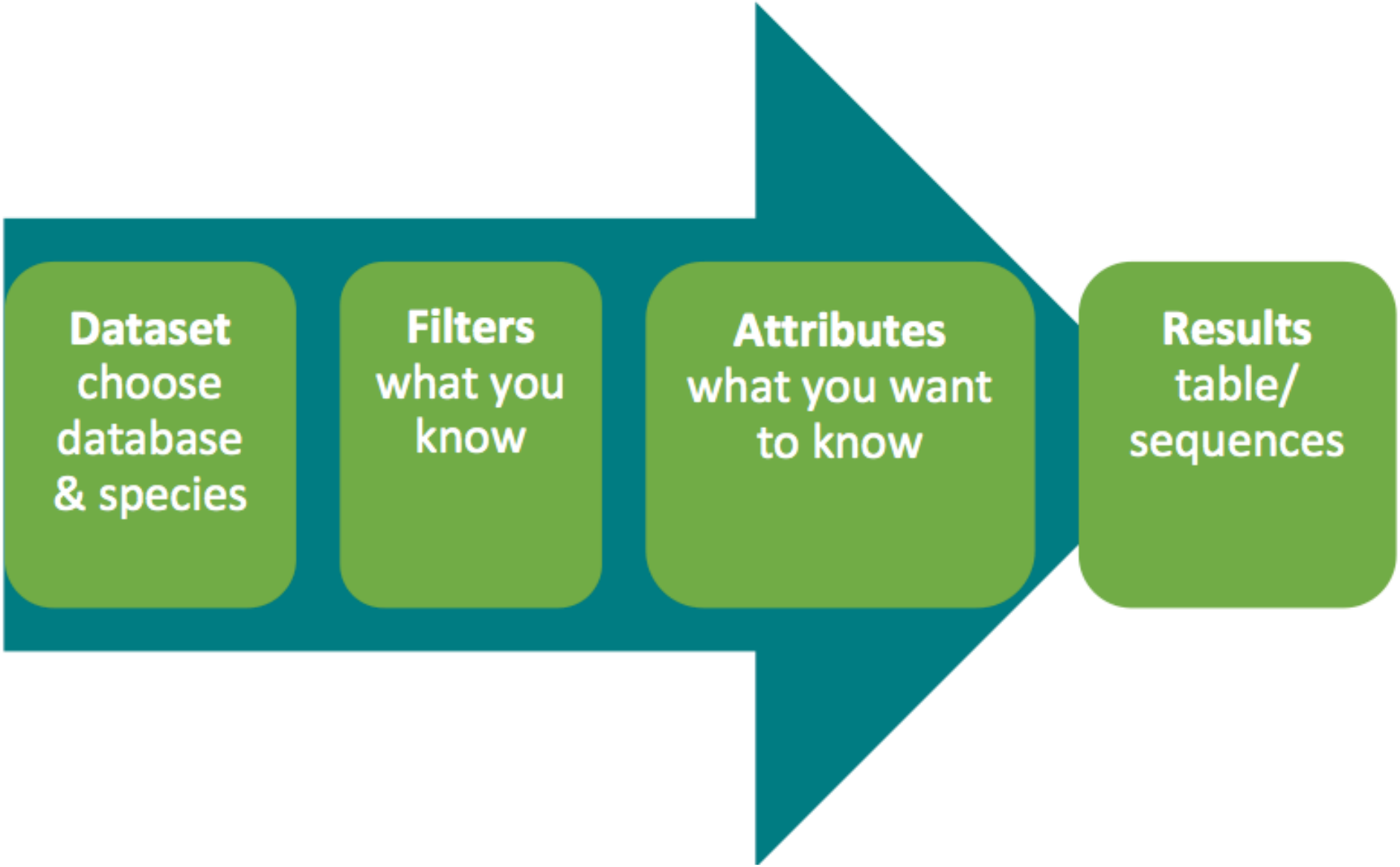


What can I do with BioMart?

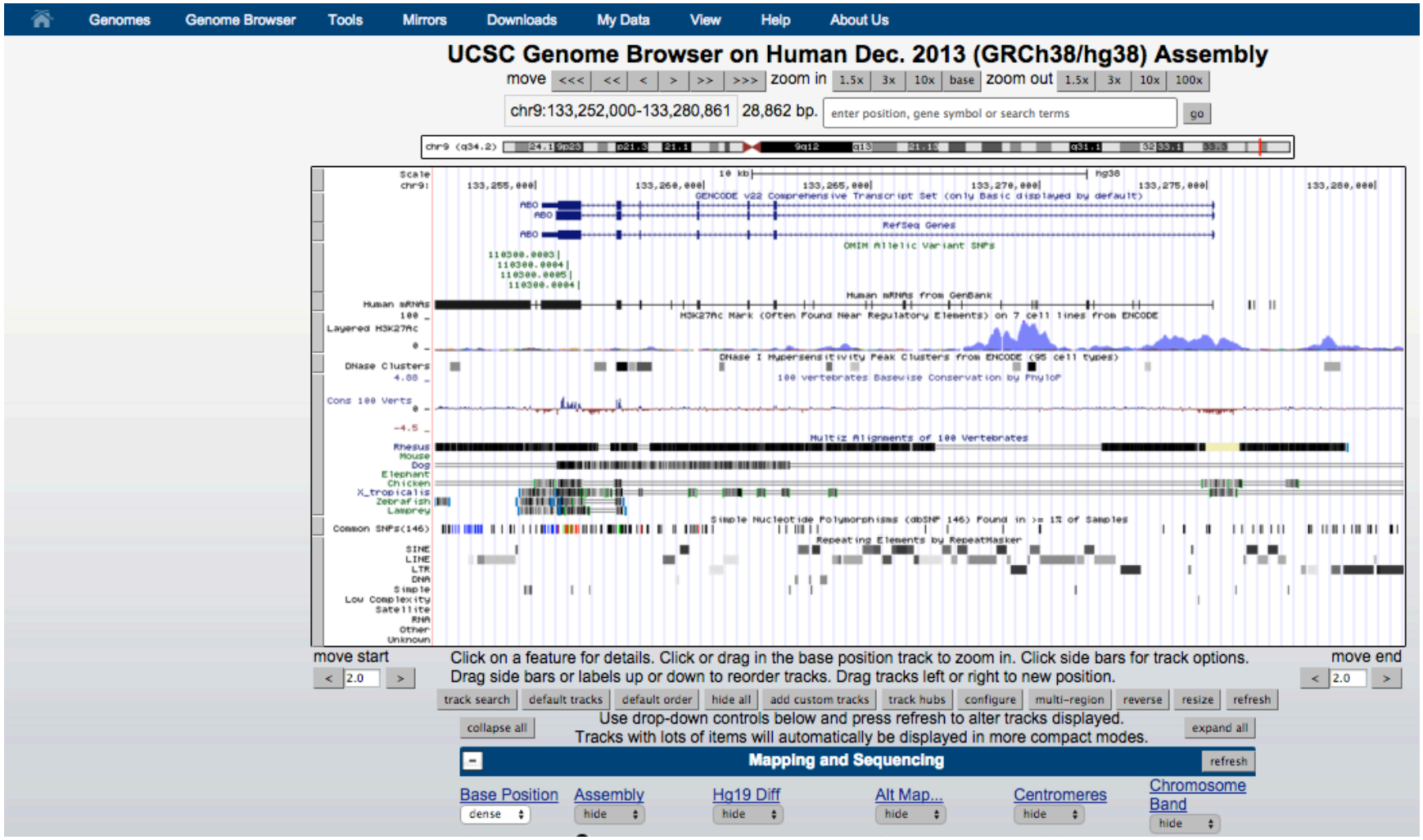
Extract data for a large number of genes, eg:

- Export list of genes in a region
- Convert IDs
- Extract sequences
- Get orthologues and paralogues

4 steps


Dataset
choose
database
& species
Filters
what you
know
Attributes
what you want
to know
Results
table/
sequences

UCSC Genome Browser



http://genome.ucsc.edu

What Distinguishes Ensembl from the UCSC Browser?

- The gene set. Automatic annotation based on mRNA and protein information.
- Species coverage (UCSC only animals)
- Comparative analysis (gene trees)
- BioMart (vs Table Browser)
- Programmatic access via the Perl API (open source)
- Integration with other databases (DAS)

Summary

Ensembl is a genome browser which integrates:

- gene annotation
- variation
- regulation (for selected species)
- comparative genomics

Sources

- slides by Dan Bolser / Bert Overduin
- http://www.ensembl.org/info/website/tutorials/index.html