Browsing Genomic Information with Ensembl

Stefan Wyder

January 2018

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

Ensembl is a data base + genome browser which integrates 4 types of ressources:

- gene annotation
- variation (SNPs, structural, …)
- regulation (for selected species)
- comparative genomics

Entry page

main menu

help

EnsemblPlants ▾ HMMER | BLAST | BioMart | Tools | Downloads | Documentation | Website help

Login/Register

Search Ensembl Plants...

Search: All species for Go

e.g. Carboxy* or chx28

search

Favourite genomes

Arabidopsis thaliana
TAIR10

Oryza sativa Japonica
IRGSP-1.0

Triticum aestivum
TGACv1

Hordeum vulgare
Hv_IBSC_PGSB_v2

Zea mays
AGPv4

Physcomitrella patens
ASM242v1

Edit favourites

All genomes

-- Select a species --

select species

View full list of all Ensembl Plants species

New Ensembl Genomes Archive Sites

Ensembl Genomes now has archive sites for all divisions:

- http://oct2017-bacteria.ensembl.org
- http://oct2017-fungi.ensembl.org
- http://oct2017-metazoa.ensembl.org
- http://oct2017-plants.ensembl.org
- http://oct2017-protists.ensembl.org

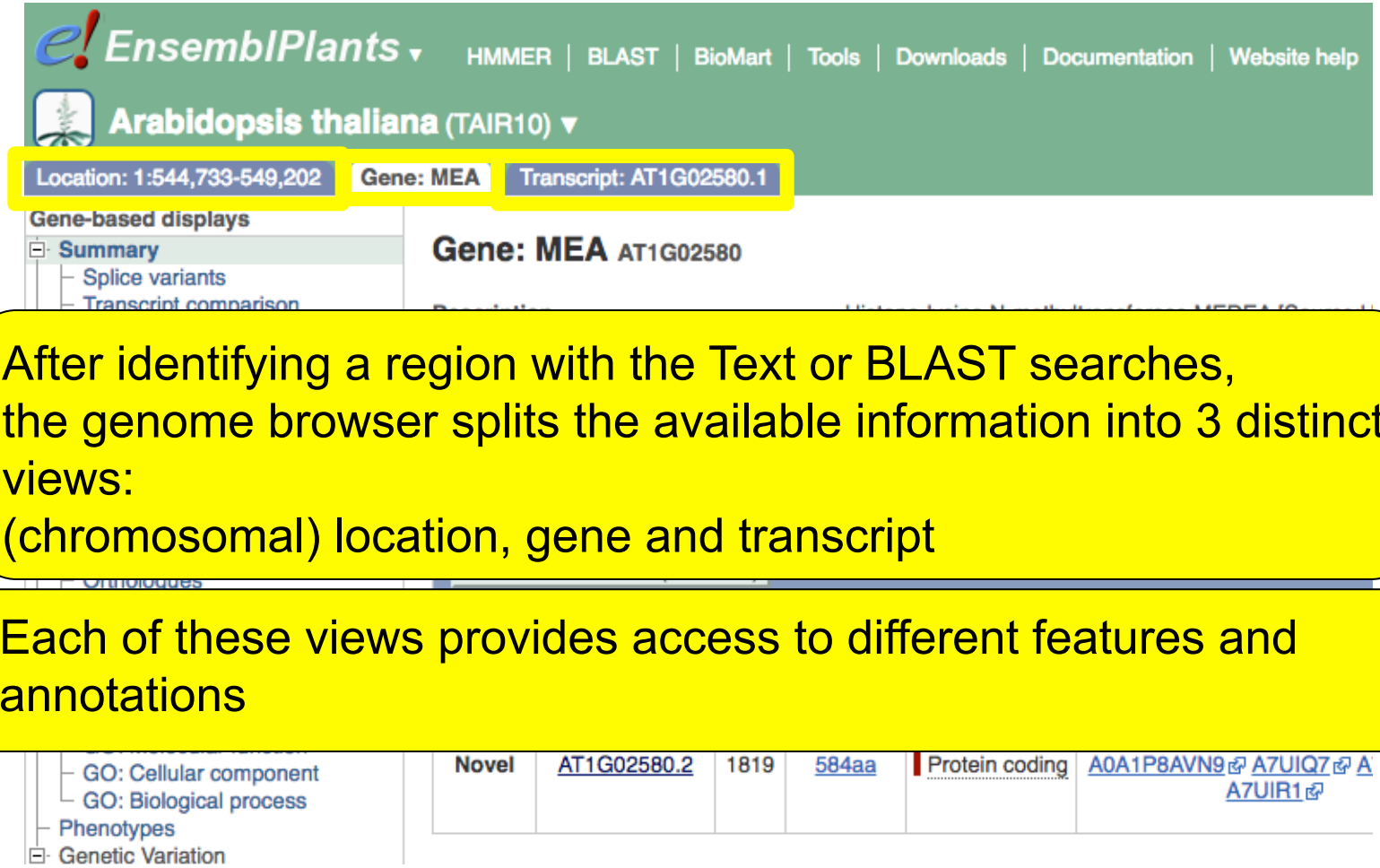
allowing researchers to access data from old releases and display track hubs for previous assemblies. Archive sites are searchable and have BioMarts available. Archival REST servers will not initially be available, but will be added in future.

The first release of the archive sites contains content from Release 37. New archive sites will be released at least once a year, under URLs indicating the date of the data they contain.The previously existing archive for Ensembl Plants, http://archive.plants.ensembl.org, will continue to be available at this URL, but also as http://mar2016-plants.ensembl.org, in accordance with the new naming scheme. As previously, data from all recent releases will continue to be available for download at ftp://ftp.ensemblgenomes.org.

New and updated genomes

This bumper release of Ensembl Plants brings eight new genomes, including cassava, cotton, cucumber, green bean, sunflower and yam. In addition we have updated to latest and greatly improved (v3.0) sorghum genome assembly from JGI as well as updated

3 tabs with different views/information



After identifying a region with the Text or BLAST searches, the genome browser splits the available information into 3 distinct views:
(chromosomal) location, gene and transcript

Each of these views provides access to different features and annotations

Chromosomal location view

Location tab

Gene tab

Transcript tab

side menu

all elements are clickable

Arabidopsis thaliana (TAIR10)

Location: 1:544,733-549,202 | Gene: MEA | Transcript: AT1G02580.2

Location-based displays
- Whole genome
- Chromosome summary
- Region overview
- Region in detail
- Comparative Genomics
  - Synteny
  - Alignments (image)
  - Alignments (text)
  - Region Comparison
- Genetic Variation
  - Resequencing
  - Linkage Data
- Markers
- Other genome browsers
  - TAIR
  - Phytozome

Configure this page

Custom tracks

Export data

Share this page

Bookmark this page

Ensembl Plants is produced in collaboration with Gramene

Chromosome 1: 544,733-549,202

Region in detail

Location: 1:544733-549202

Go

Gene Legend

Protein Coding: protein coding

Imports and Miscellaneous: Fgenesh

Variant Legend

frameshift variant

stop retained variant

missense variant

5 prime UTR variant

splice region variant

3 prime UTR variant

There are currently 76 tracks turned off.

Ensembl Plants Arabidopsis thaliana version 91.11 (TAIR10) Chromosome 1: 544,733 - 549,202

Gene view

Location: 1:544,733-549,202 | Gene: MEA | Transcript: AT1G02580.1

Gene-based displays

- Summary
  - Splice variants
  - Transcript comparison
  - Gene alleles
- Sequence
  - Secondary Structure
- Gene families
- Literature
- Orthologues
- Paralogues
- Pan-taxonomic Compara
  - Gene Tree
  - Orthologues
- Ontologies
  - GO: Molecular function
  - GO: Cellular component
  - GO: Biological process
- Phenotypes
- Genetic Variation
  - Variant table
  - Variant image
  - Structural variants
- Gene expression
- Pathway
- Regulation
- External references
- Supporting evidence

lists isoforms

Gene: MEA AT1G02580

Description	Histone-lysine N-methyltransferase MEDEA [Source:UniProtKB/Swiss-Prot;Acc:O65312]
Synonyms	At1g02580, EMB173, FIS1, MEDEA, SDG5, SET5, T14P4.11
Location	Chromosome 1: 544,733-549,202 forward strand.
About this gene	This gene has 2 transcripts (splice variants), 23 orthologues and 8 paralogues.
Transcripts	Hide transcript table

Show/hide columns (1 hidden) Filter

Name	Transcript ID	bp	Protein	Biotype	UniProt	RefSeq	Flags
Novel	AT1G02580.1	2341	689aa	Protein coding	A0A178WGQ4 A7UIQ9 A7UIR6 C0KG89 C0KG93 C0KG94 C0KG95 C0KG97 C0KG98 C0KG99 O65312	NM_001331345.1 NM_100139.4 NP_563658.1	
Novel	AT1G02580.2	1819	584aa	Protein coding	A0A1P8AVN9 A7UIQ7 A7UIQ8 A7UIR1	NM_001331345.1 NM_100139.4 NP_001322965.1	

Summary

Name	MEA (UniProtKB Gene Name)
UniProtKB	This gene has proteins that correspond to the following UniProtKB identifiers: O65312
Gene type	Protein coding
Annotation method	Gene annotation by ARAPORT through a process of automatic and manual curation.

Transcript view

Location: 1:544,733-549,202 | Gene: MEA | Transcript: AT1G02580.1

Transcript-based displays

- Summary
- Sequence
  - Exons
  - cDNA
  - Protein
- Protein Information
  - Protein summary
  - Domains & features
  - Variants
- Genetic Variation
  - Variant table
  - Variant image
  - Population comparison
  - Comparison image
- External References
  - General identifiers
  - Oligo probes
- Supporting evidence
- ID History
  - Transcript history
  - Protein history

Configure this page

Custom tracks

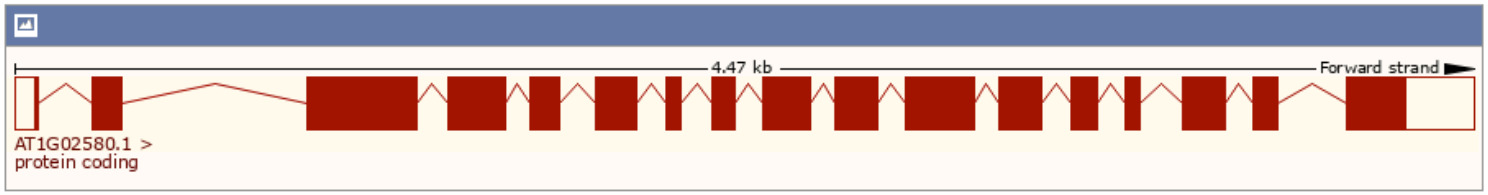
Export data

Share this page

Transcript: AT1G02580.1

Description	Histone-lysine N-methyltransferase MEDEA [Source:UniProtKB/Swiss-Prot;Acc:O65312]
Location	Chromosome 1: 544,733-549,202 forward strand.
About this transcript	This transcript has 17 exons, is annotated with 19 domains and features, is associated with 108 variations and maps to 3 oligo probes.
Gene	This transcript is a product of gene AT1G02580 Show transcript table

Summary



Statistics	**Exons:** 17, **Coding exons:** 17, **Transcript length:** 2,341 bps, **Translation length:** 689 residues
Uniprot	This transcript corresponds to the following Uniprot identifiers: O65312
Version	AT1G02580.1.
Type	Protein coding
Annotation Method	Gene annotation by ARAPORT through a process of automatic and manual curation.

Gene view | Gene tree

Location: 1:544,733-549,202 | Gene: MEA

Gene-based displays
- Summary
  - Splice variants
  - Transcript comparison
  - Gene alleles
- Sequence
  - Secondary Structure
- Gene families
- Literature
- Plant Compara
  - Genomic alignments
  - Gene tree
  - Gene gain/loss tree
  - Orthologs
  - Paralogs
- Pan-taxonomic
- Structural variants
- Gene expression
- Pathway
- Regulation
- External references
- Supporting evidence
- ID History
  - Gene history

Configure this page

Custom tracks

Export data

Share this page

Bookmark this page

Ensembl Plants is produced in collaboration with Gramene

Click on "Gene tree"

Gene: MEA AT1G02580

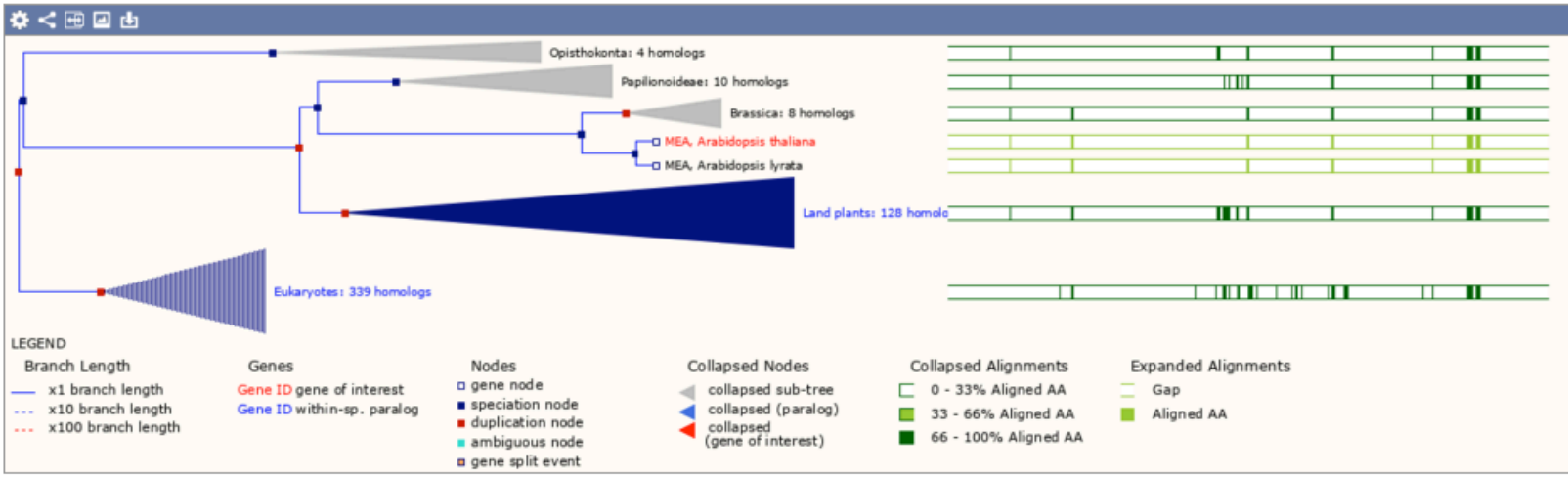
Description	Histone-lysine N-methyltransferase MEDEA [Source:UniProtKB/Swiss-Prot;Acc:Q65312]
Synonyms	At1g02580, EMB173, FIS1, MEDEA, SDG5, SET5, T14P4.11
Location	Chromosome 1: 544,733-549,202 forward strand.
About this gene	This gene has 2 transcripts (splice variants), 23 orthologues and 8 paralogues.
Transcripts	Show transcript table

Gene tree

GeneTree EPIGT00880000131274

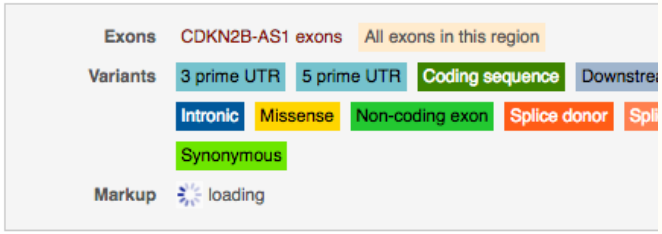
	491
des	361
des	100
des	22
nts	7

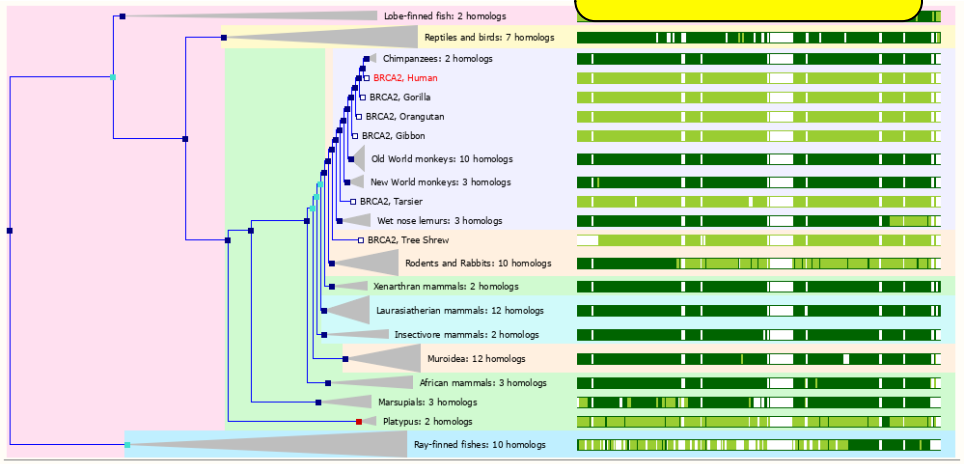
Show annotations table

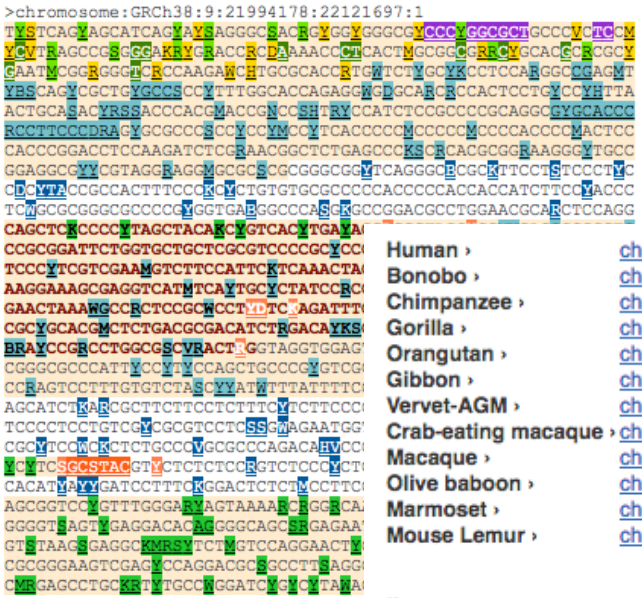


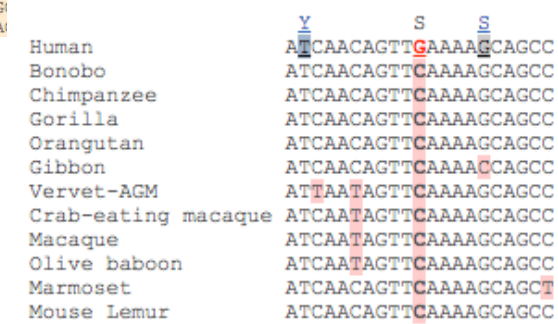
Many other data displays, e.g.

gene tree

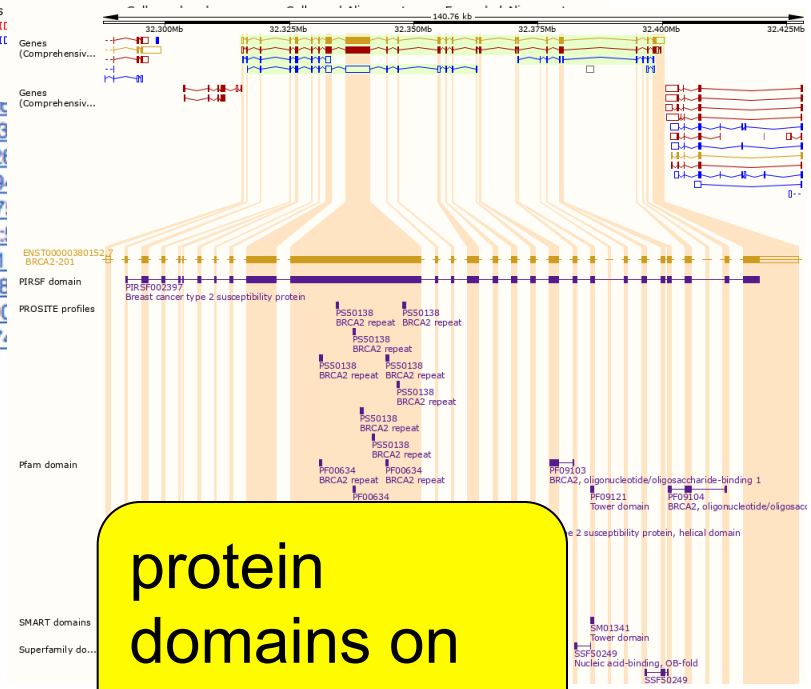








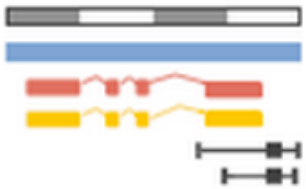
SNPs



protein domains on exons

Gene annotation

- Genomic sequence
- Gene / transcript / protein models
- External references
- Mapped sequences
  - cDNAs, proteins, repeats, markers, probes, etc.



Methods in gene annotation

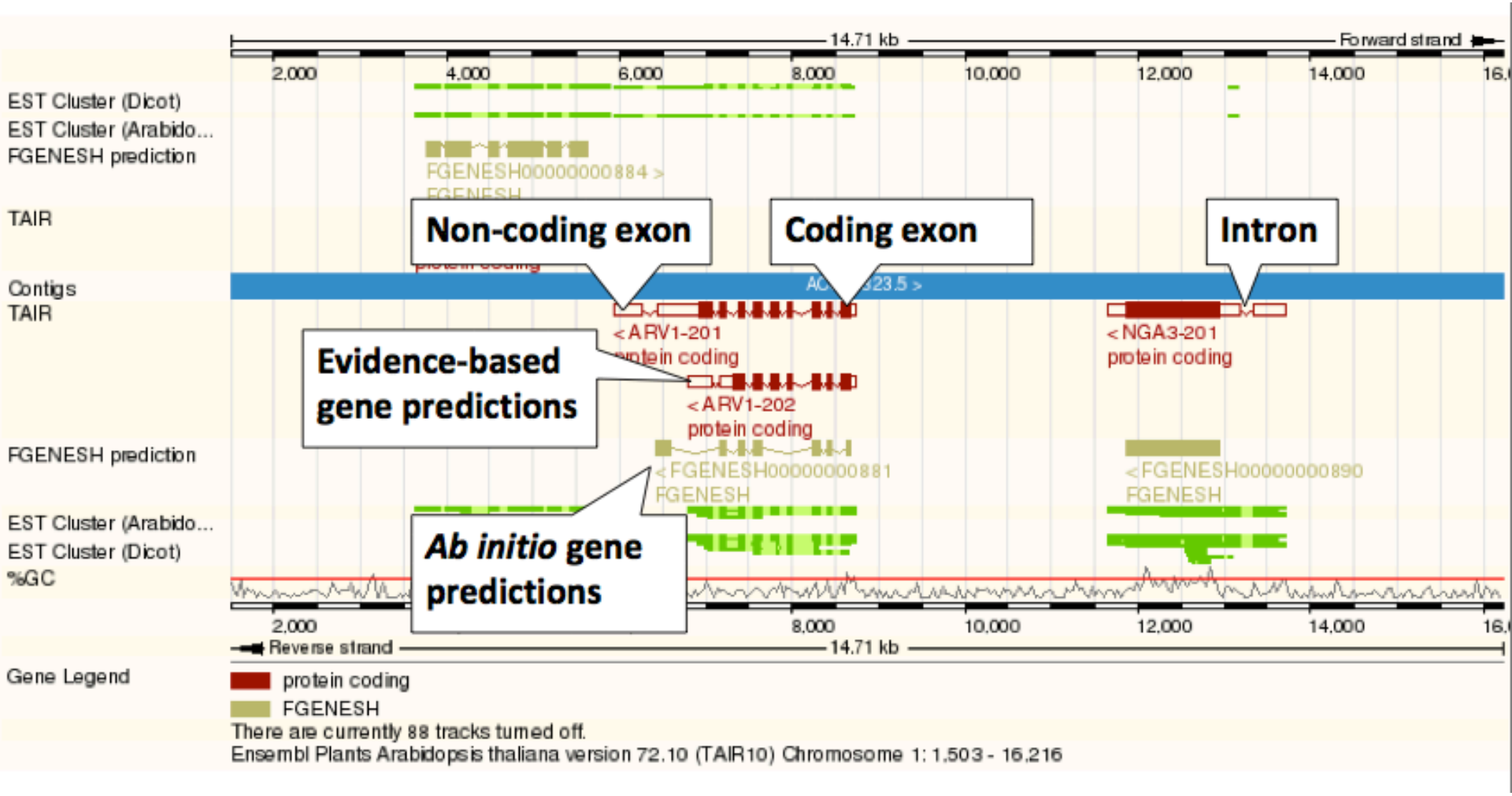
Automatic annotation
Using known proteins/ESTs/cDNAs

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

Manual annotation
To correct errors in automatic annotation

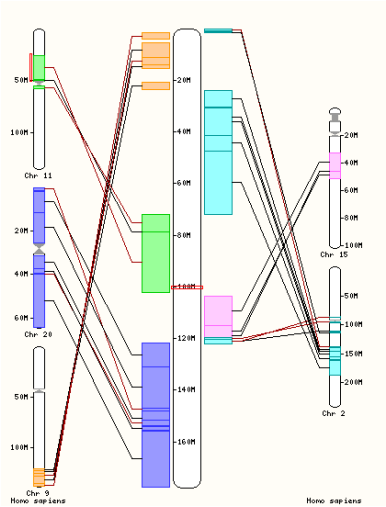
***Ab initio* predictions**
(finding apparent ORFs in the sequence)

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

Comparative data

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



Synteny between Rat chromosome 3 and Human

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:

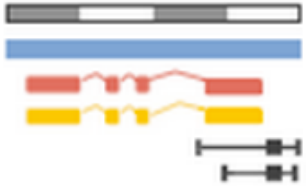
  MUS (*Mus musculus*) for mouse: ENSMUSG###
  DAR (*Danio rerio*) for zebrafish: ENSDARG###, etc.

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

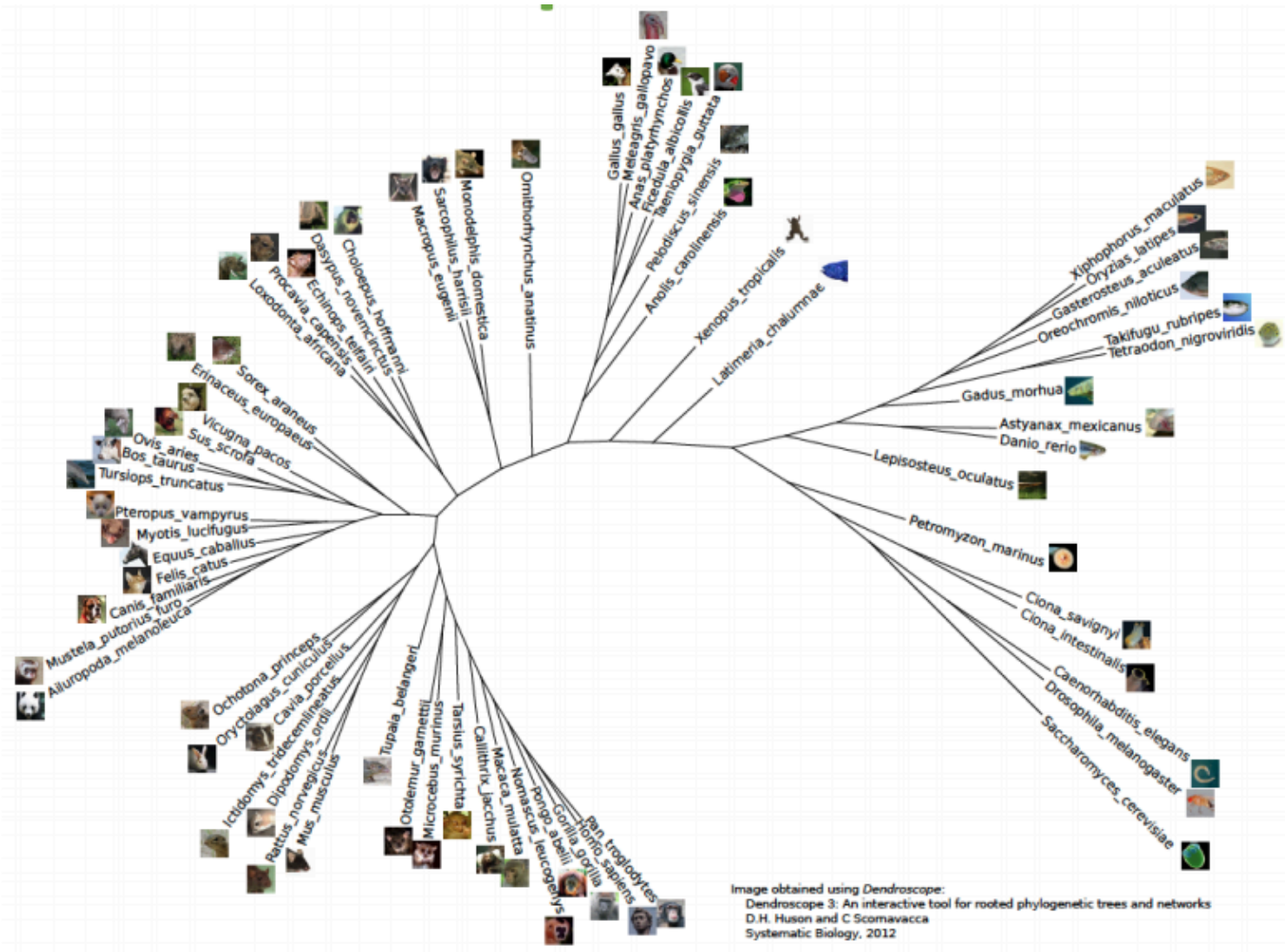
More Ensembl features

- Variation display and VEP
- Display of user data
- Completely open source

- 4-5 updates (versions/freezes) a year
- archived versions

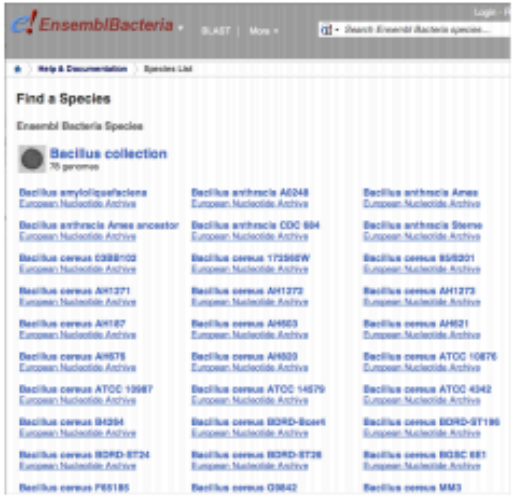


124 vertebrate species in Ensembl v91

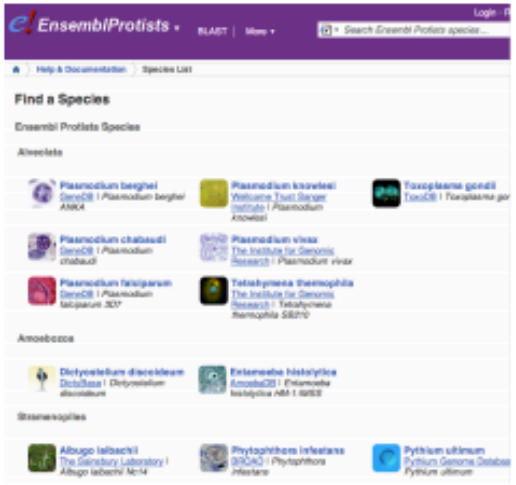


+ early access Pre! websites

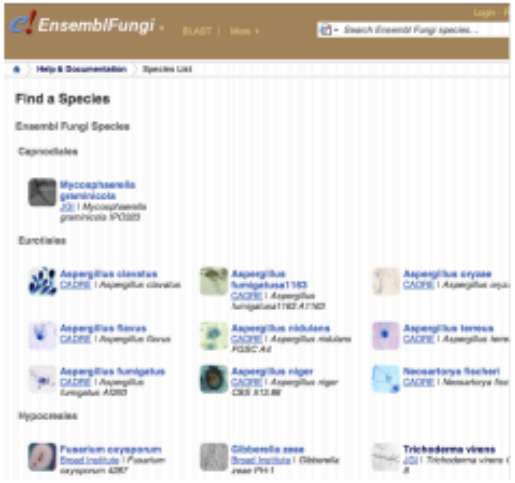
Non-vertebrates on Ensembl Genomes



Bacteria (44,046 genomes)

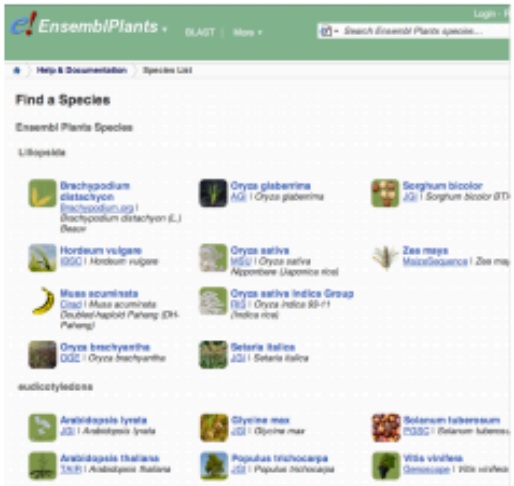


Protists (189)



Fungi (811)

Metazoa (69)



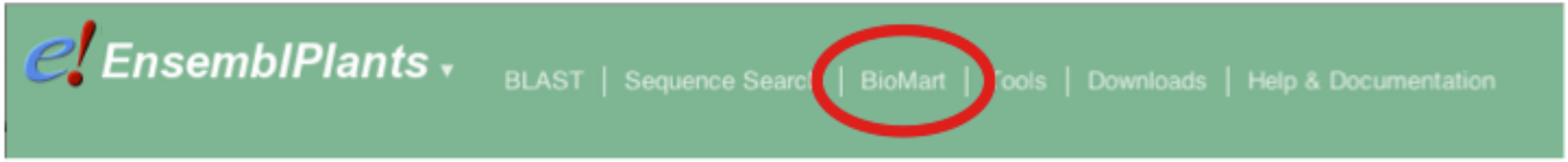
Plants (53)

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

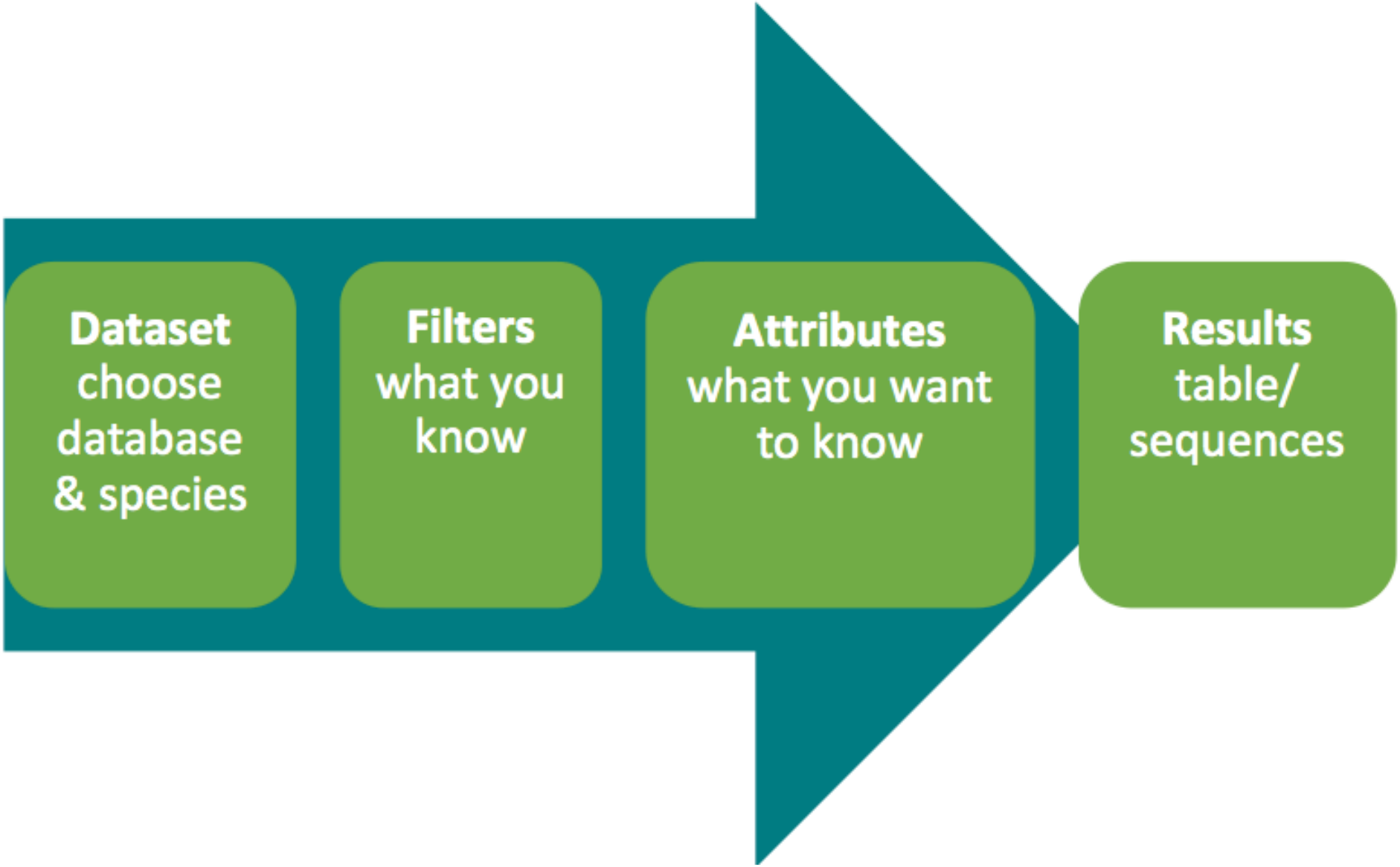


What can I do with BioMart?

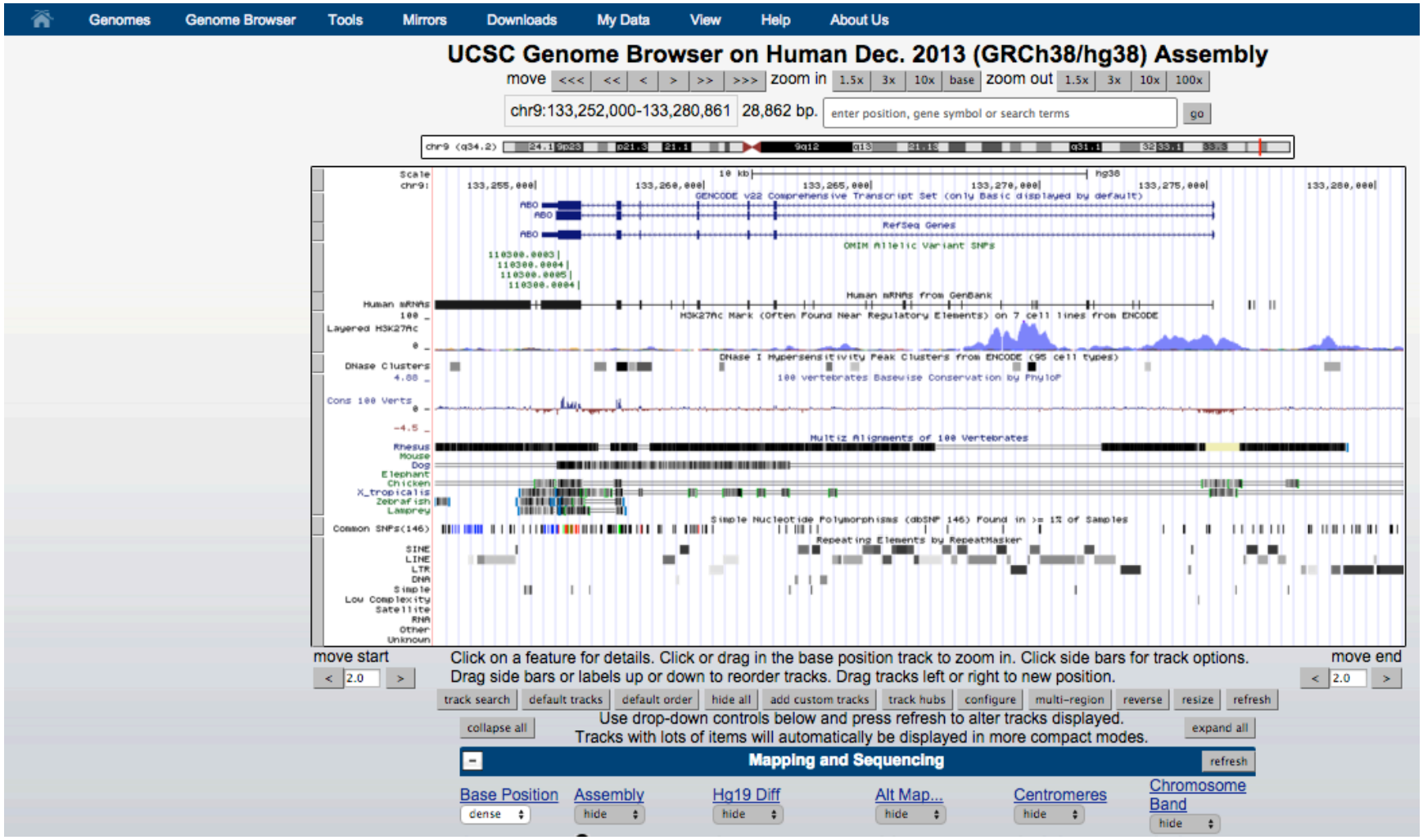
Extract data for a large number of genes, e.g.:

- Export list of genes in a region
- Convert IDs
- Retrieve fasta sequences of all introns
- Get orthologues and paralogues

Biomart: 4 steps



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)

Sources

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

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