Variation data in Ensembl Plants

Hands-on workshop
July 2015

Dan Bolser
Denise Carvalho-Silva
Brandon Walts
EMBL-EBI

http://plants.ensembl.org

Overview of the next two hours...

Before tea

Session 1, Browser:

- Introduction
  - Ensembl Plants and plant variation data
- Hands-on
  - Variation in the Ensembl browser
- Displaying your data in Ensembl

After tea

Session 2, Tools:

- Variant Effect Predictor
  - Introduction and hands-on
- Mining variation data with BioMart
  - Introduction and hands-on

Resources

These slides:

http://tinyurl.com/transplant2015

Course booklet:

http://tinyurl.com/transplant2015b

Introduction: An overview of Ensembl Plants

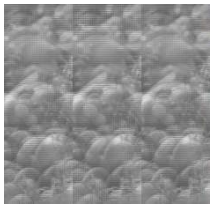
Dan Bolser
Ensembl Plants project leader
EMBL-EBI

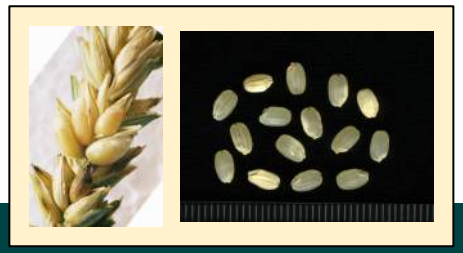
http://plants.ensembl.org

Introduction overview

Background:

- Ensembl Plants
  - History
  - Data
- Recent updates
  - Wheat
  - Tomato
  - Rice





Plant variation data:

- Plant variation data
- Variation in the Ensembl browser
- Tools for processing your own data

Ensembl Plants uses Ensembl technology

Ensembl:

- A platform for **genome browsing**, annotation and analysis.
- Has ‘modules’ for handling:
  - Genomic data, comparative genomics, variations, and epigenomics.
- Multiple points of access to data:
  - Browser-based application, Perl and REST APIs, direct access (MySQL), DAS (client and server), FTP, ..
  - Data mining tool. biomart
- Upload your own data and compare it to the reference seq. and annotation.

Ensembl was originally developed for vertebrate genomes, subsequently extended to non-vertebrate species:

- Ensembl → Ensembl Genomes → Ensembl Plants


Ensembl is developed
jointly by the EBI and
the Wellcome Trust
Sanger Institute

Currently 39 genomes in Ensembl Plants

http://plants.ensembl.org

Dicots in Ensembl Plants: 12 in 7 orders


Arabidopsis lyrata G A
JGI | 81972


Arabidopsis thaliana V P G A
TAIR | 3702



Brassica rapa G A
IVFCAAS | 51351

Brassicales


Glycine max G A
JGI | 3847


Medicago truncatula G A
IMGAG | 3880

Fabales


Populus trichocarpa G A
JGI | 3694

Malpighiales


Prunus persica G A
IPGI | 3760

Rosales


Solanum lycopersicum P G A
ITGSP | 4081


Solanum tuberosum G A
PGSC | 4113

Solanales

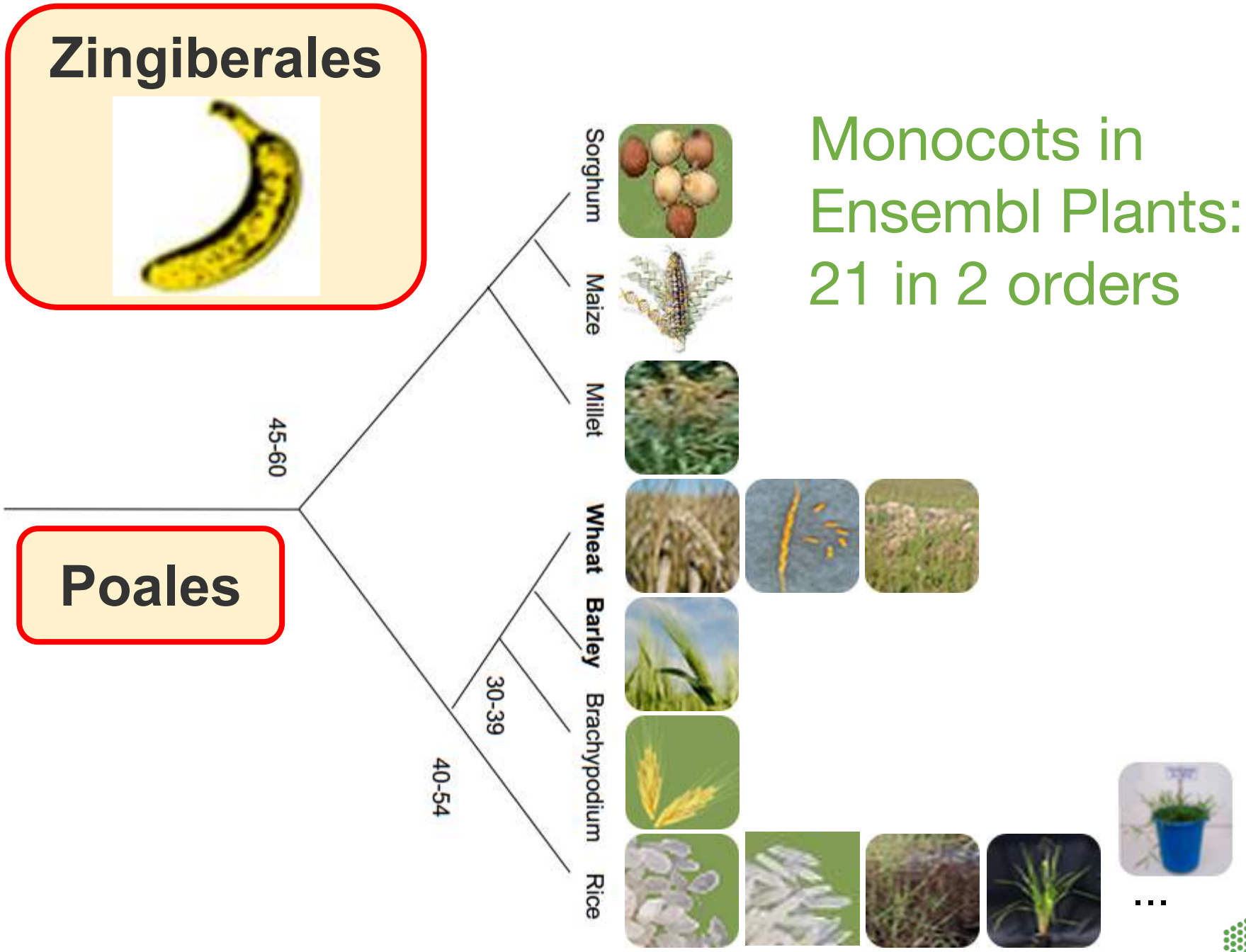

Vitis vinifera V P G A
Genoscope | *Vitis vinifera* | 29760

Vitales


Theobroma cacao (cocoa)
Data Source: The Cacao Genome Database

Malvales

Monocots in Ensembl Plants: 21 in 2 orders



'Others': 6 in ~5 Classes

Rhodophyta

Cyanidioschyzon merolae P G
CMGP | 280699

Chlorophyta

Chlamydomonas reinhardtii P G A
JGI | 3055

Bryophyta

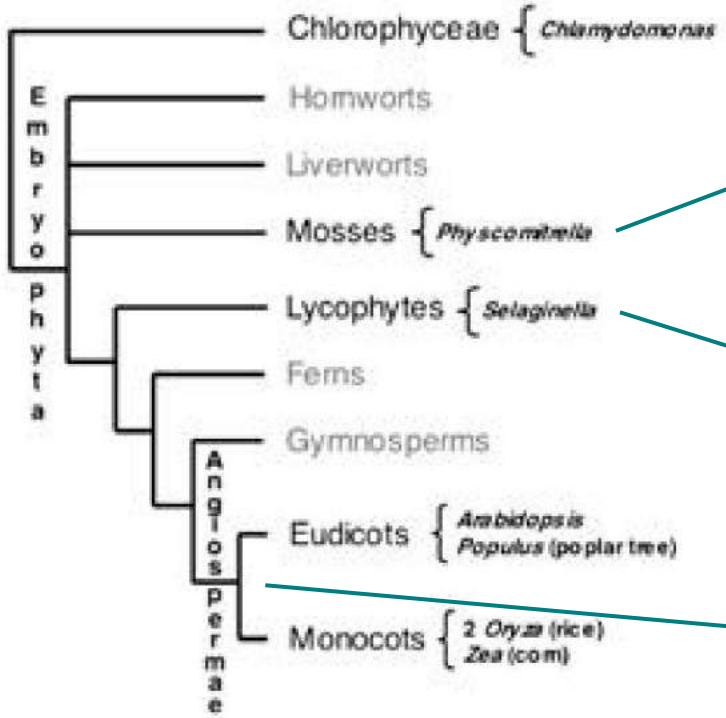
Physcomitrella patens P G A
JGI | 145481

Lycopodiophyta

Selaginella moellendorffii G A
JGI | 88036

Amborellales

Amborella trichopoda P G
AGD | 13333



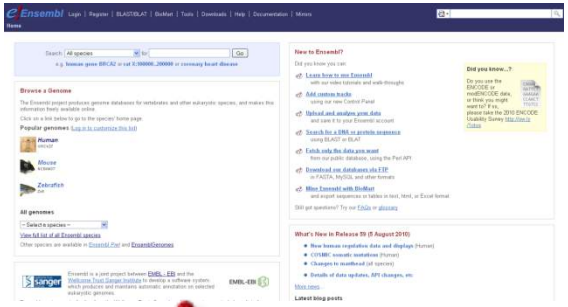
Mónica Medina (2005) Proc. Natl. Acad. Sci. USA 102, 6630-6635

What do we host?

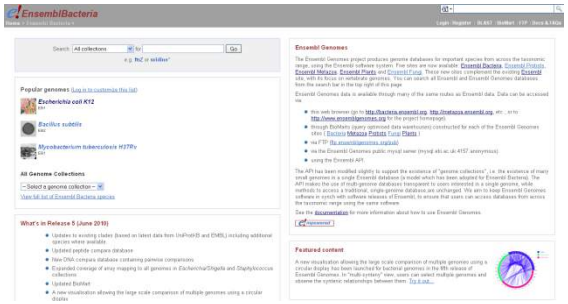
Genomes are selected for inclusion based on:

- Availability of complete genome sequence
- Importance as model organisms
  - e.g. Arabidopsis and Brachypodium.
- Importance for agriculture
  - e.g. potato, sorghum, barley, rice, wheat and Brassica.
- Interest as evolutionary reference points
  - e.g the basal angiosperm, Amborella trichopoda, the aquatic alga Chlamydomonas reinhardtii, the moss Physcomitrella patens and the vascular non-seed spikemoss Selaginella moellendorffii.

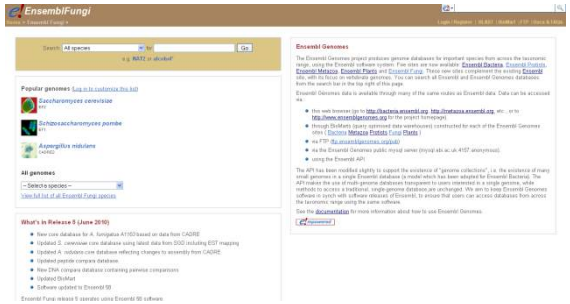
Or all the above!



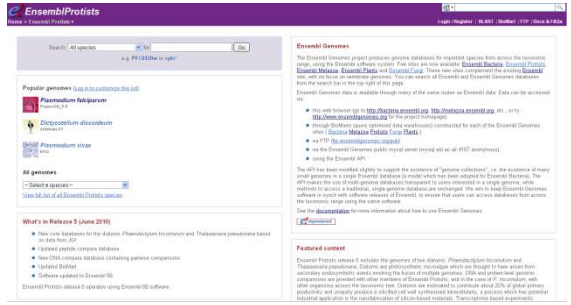
Ensembl



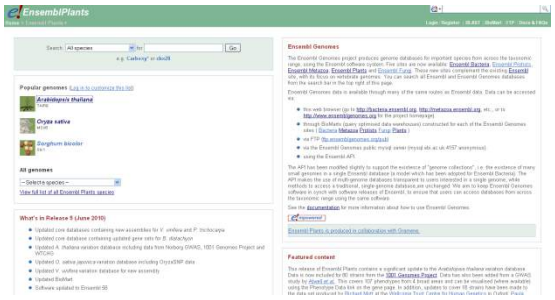
EnsemblBacteria



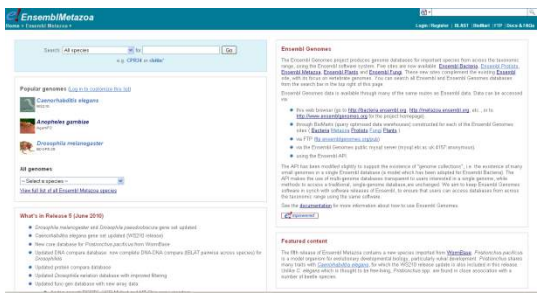
EnsemblFungi



EnsemblProtists



EnsemblPlants



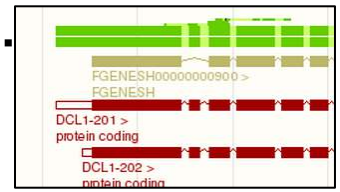
EnsemblMetazoa

Ensembl and agriculture

- Scope of Ensembl Plants includes important crop and model species
- Ensembl Metazoa, Protists, Fungi and Bacteria include important agricultural pathogens, pests, pollinators, symbiotes
- Important farm animals (cow, sheep, pig, chicken, etc.) are among the key species in vertebrate Ensembl
- See also, PhytoPath
  - Integrates genome-scale data from important plant pathogen species with literature-curated information about the phenotypes of host infection.

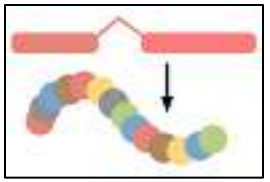
Types of data in Ensembl (Ensembl Plants)

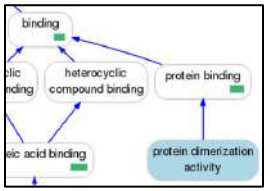
- Genomic sequence
- Gene, transcript, and protein annotations
- External references and ontology terms
- Mapped sequences: cDNAs, proteins, probes, BACs, repeats, markers, ...



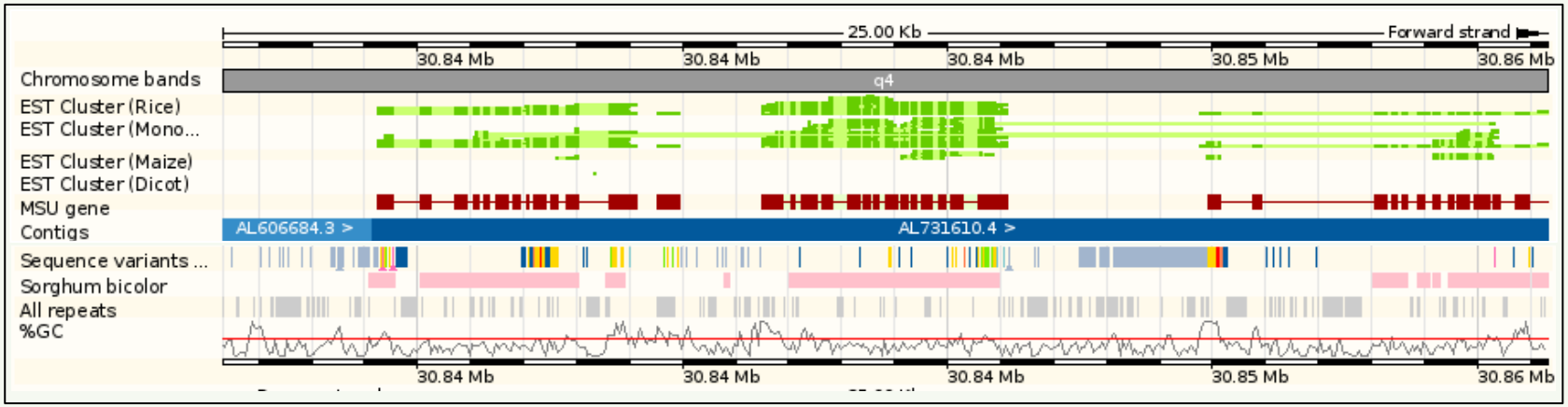
- Variation data:
  - sequence variants
  - structural variants

- Comparative data:
  - gene trees, orthologues, paralogues
  - whole genome alignments and synteny





Genome Browsers



- Assembly structure and sequence
- Genes
- Expressed regions
- Comparative alignments
- Genetic markers
- Variation
- Repeats & transposable elements
- Regulatory & Epigenetic marks

Baseline Annotation

- Transposon discovery
- Gene prediction
- EST Alignment

Recent updates (Release 27)

http://plants.ensembl.org

Release 4 to 5 times a year (3 months)

Release 24 (Nov. 2014)

- New chromosome assembly for barley.
- Added 10 million wheat ‘Inter-homoeologous variations’ (IHVs).
- Added 70 million variations from the tomato 150 genomes project.

Release 25 (Jan. 2015)

- New chromosome assembly for wheat.
- More than 10 million new barley variations from 170 individuals.
- New whole genome alignments between Brassica rapa and oleracea.

Release 4 to 5 times a year (3 months)

Release 26 (Mar. 2015)

- Updated genome for *Medicago truncatula* and *Oryza meridionalis*.
- New RNA-Seq data for barley.
- New whole genome alignments between triticeae using ATAC.

Release 27 (Jun. 2015)

- Tomato genome updated to version 2.5
- New variation data from the wheat HapMap project

Funding (Ensembl Plants)

- Ensembl Genomes Funded by
  - EMBL
  - EU (INFRAVEC, Microme, transPLANT, AllBio)
  - BBSRC (PhytoPath, wheat, barley and midge sequencing, UK-US collaboration, RNAcentral, capital grant)
  - Wellcome Trust (PomBase)
  - NIH/NIAID (VectorBase)
  - NSF (Gramene collaboration)
  - Bill and Melinda Gates Foundation (wheat rust)

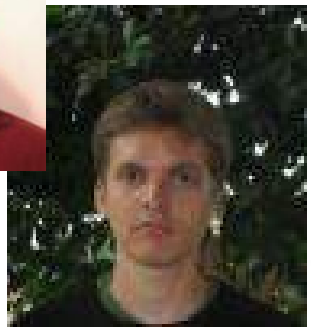
Ensembl Genomes (Ensembl Plants)

- James Allen, Irina Armean, Dan Bolser, Mikkel Christensen, Paul Davies, Christoph Grabmueller, Kevin Howe, Malcolm Hinsley, Jay Humphrey, Arnaud Kerhornou, Paul Kersey, Julia Khobdova, Eugene Kulesha, Nick Langridge, Dan Lawson, Mark McDowall, Uma Maheswari, Gareth Maslen, Michael Nuhn, Chuang Kee Ong, Michael Paulini, Helder Pedro, Anton Petrov, Dan Staines, Mary Ann Tuli, Brandon Walts, Gary Williams
    - If you have a question that is not answered here, please contact me (dbolser@ebi.ac.uk) or our HelpDesk:
        - helpdesk@ensemblgenomes.org

People

- Arnaud Kerhornou
  - Most of the wheat work
- Brandon Walts
  - Barley projection
- Christoph Grabmüller
  - Projection pipeline
- Paul Kersey
  - Wheat inter-homoeologous variations
- The web team
  - Eugene Kulesha
  - Julia Khobova
  - Nick Langridge







On to variation...

Introduction to variation data in Ensembl Plants

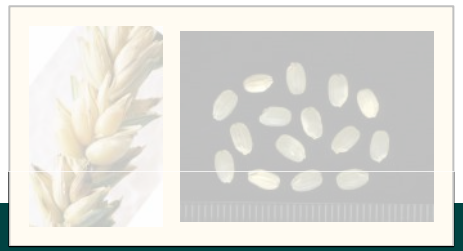
Dan Bolser
Ensembl Plants project leader
EMBL-EBI

http://plants.ensembl.org

Introduction overview

Background:

- Ensembl Plants
  - History
  - Data
- Recent updates
  - Wheat
  - Tomato
  - Rice



Plant variation data:

- Plant variation data
- Variation in the Ensembl browser
- Tools for processing your own data

Current plant variation data

Species	Samps	Pops*	~Variations
Arabidopsis thaliana	1,610	7	14,000,000 **SV: 14,000**
Brachypodium distachyon	3	3	300,000
Hordeum vulgare	188	16	24,000,000
Oryza glaberrima	21	1	800,000
Oryza indica	19	19	5,000,000
Oryza sativa	416	21	6,000,000
Solanum lycopersicum	84	8	71,000,000
Sorghum bicolor	3	1	**SV: 64,000**
Triticum aestivum	248	1	700,000
Vitis vinifera	17	3	500,000
Zea mays	103	1	51,000,000

Some large upcoming variation projects:

- International Maize and Wheat Improvement Center (CIMMYT)
- BMAP (150 brassica)
- 3000 rice cultivars
- Barley GBS on exomes of ~2500 EMS mutants

* Pops can be true genetic populations or, more often, simply ‘collections’.

Variation Data

- SNPs and structural variation
- Natural and structured populations

Species	Variants	Source	Studies
Oryza sativa Group japonica	**3,332,525**	160K SNPs x 20 accessions 1311 SNPs x 395 accessions NCBI dbSNP	McNally et al. (2009). PNAS 106:12273-12278 Zhao et al. 2010. PLoS ONE. 5:e10780
Oryza sativa Group indica	**4,747,883**	NCBI dbSNP	
Zea mays	**50,719,843**	HapMap1: NAM founder lines HapMap2: pre-domesticated & domesticated lines	Gore et al. 2009. Science 326:1115-1117. Chia et al. 2013. Genet 44:803-807.
Arabidopsis thaliana	**14,234,197** **SV: 13,667**	250K SNPs x 1179 accessions 1001 genomes project: 411 resequenced accessions	Atwell et al. 2010. Nature. 465:627-631.
Brachypodium distachyon	**327,988**	3 accessions of Brachypodium sylvaticum	Fox et al. (2013) Applications in Plant Sciences 1 (3):1200011. 2013
Vitis vinifera	**457,404**	Resequencing USDA germplasm collection	Myles S, et al. 2010. PLoS ONE. 5:e8219.
Hordeum vulgare	**12,994,003**	Resequencing 4 accessions plus wild barely	The International Barley Genome Sequencing Consortium. 2012. Nature 491, 711–716
Oryza glaberrima	**7,704,409**	Resequenced 20 accessions African rice & wild progenitor	Oryza Genome Evolution project
Sorghum bicolor	**SV: 64,507**	Structural variants from Database of Genomic Variants archive (dGVA)	Zheng et al. 2011. Genome Biol.. 12:R114.

Arabidopsis

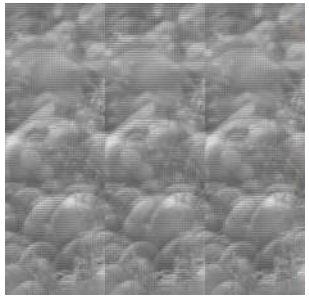
Currently:

- SNPs for 1,179 strains on the Affy 250k SNP chip.
- Resequencing data for 18 Arabidopsis lines.
- A total of 392 strains from the 1001 Genomes Project:
  - 80 strains from the Cao pilot study.
  - 132 strains from the Salk Institute.
  - 180 strains from the Nordborg group at GMI.
- Phenotype data from a GWAS study of 107 phenotypes in 95 inbred lines.

Coming ‘soon’:

- The ‘final’ 1001 data
  - ~1,300 strains and 13,000,000 genotypes.
  - The ‘full imputed’ matrix for ~2,000 strains.

Tomato



Currently:

- Data from resequencing 84 tomato accessions
  - Accessions include cultivated wild relatives representative of the Lycopersicon, Arcanum, Eriopersicon and Neolycopersicon groups.
  - The variation data has been submitted to the ENA with accession ERP004618, and has been locus-level accessioned using the transPLANT variation archive.

Potentially:

- Data from the 350 genomes study...

Wheat

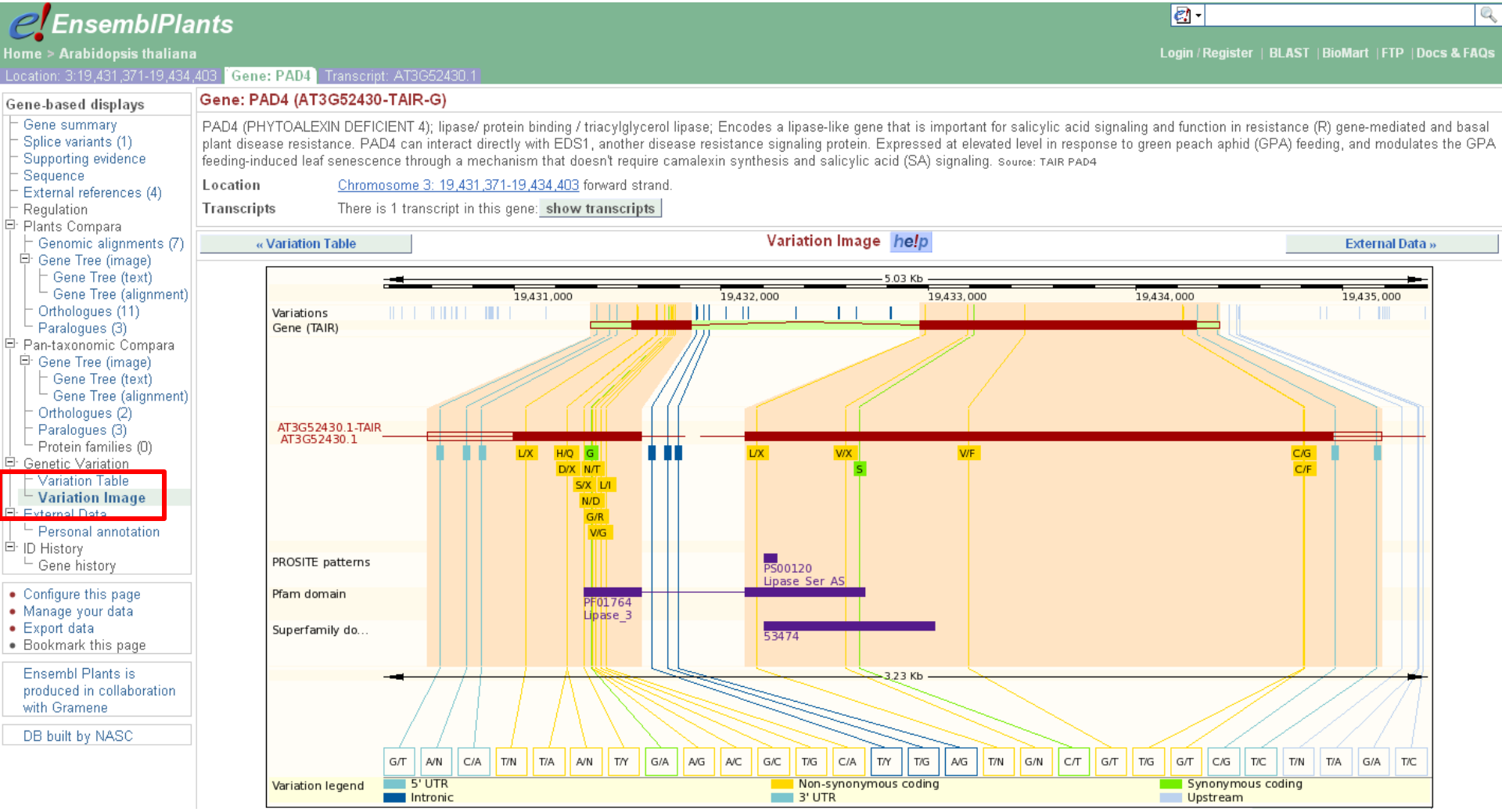
Currently:

- SNP data for ~725,000 loci across ~250 individuals provided by CerealsDB.
- ~1.5 million loci from the Wheat Hapmap project generated by resequencing 62 diverse wheat lines.
- A total of ~2,000,000 inter-homoeologous variants called from whole genome alignments between the A, B and D component genomes.

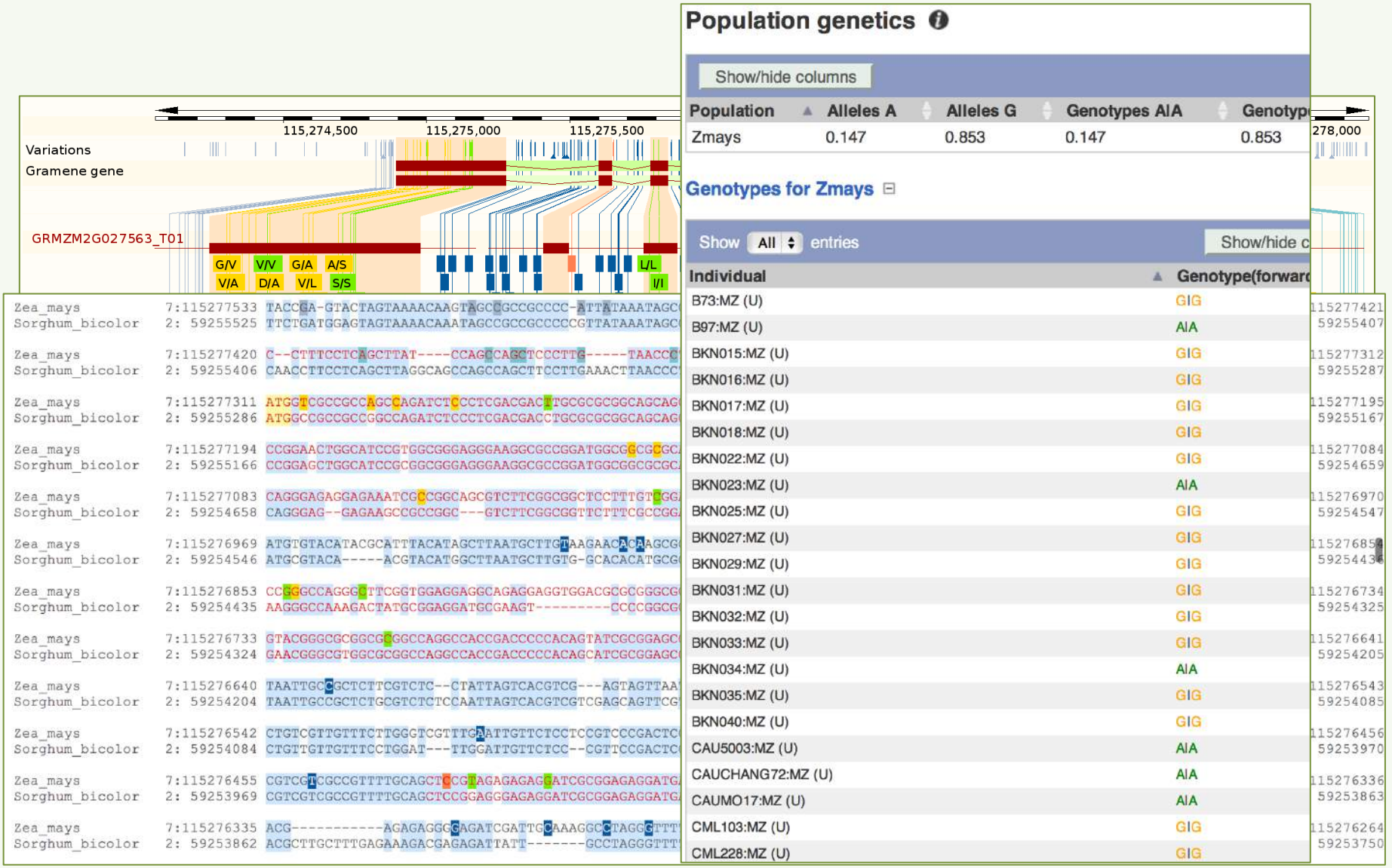
Coming ‘soon’:

- Exome capture of ~2,000 EMS mutants.

A ‘typical’ variation view in Ensembl



Viewing Variation Data



Population genetics

Show/hide columns

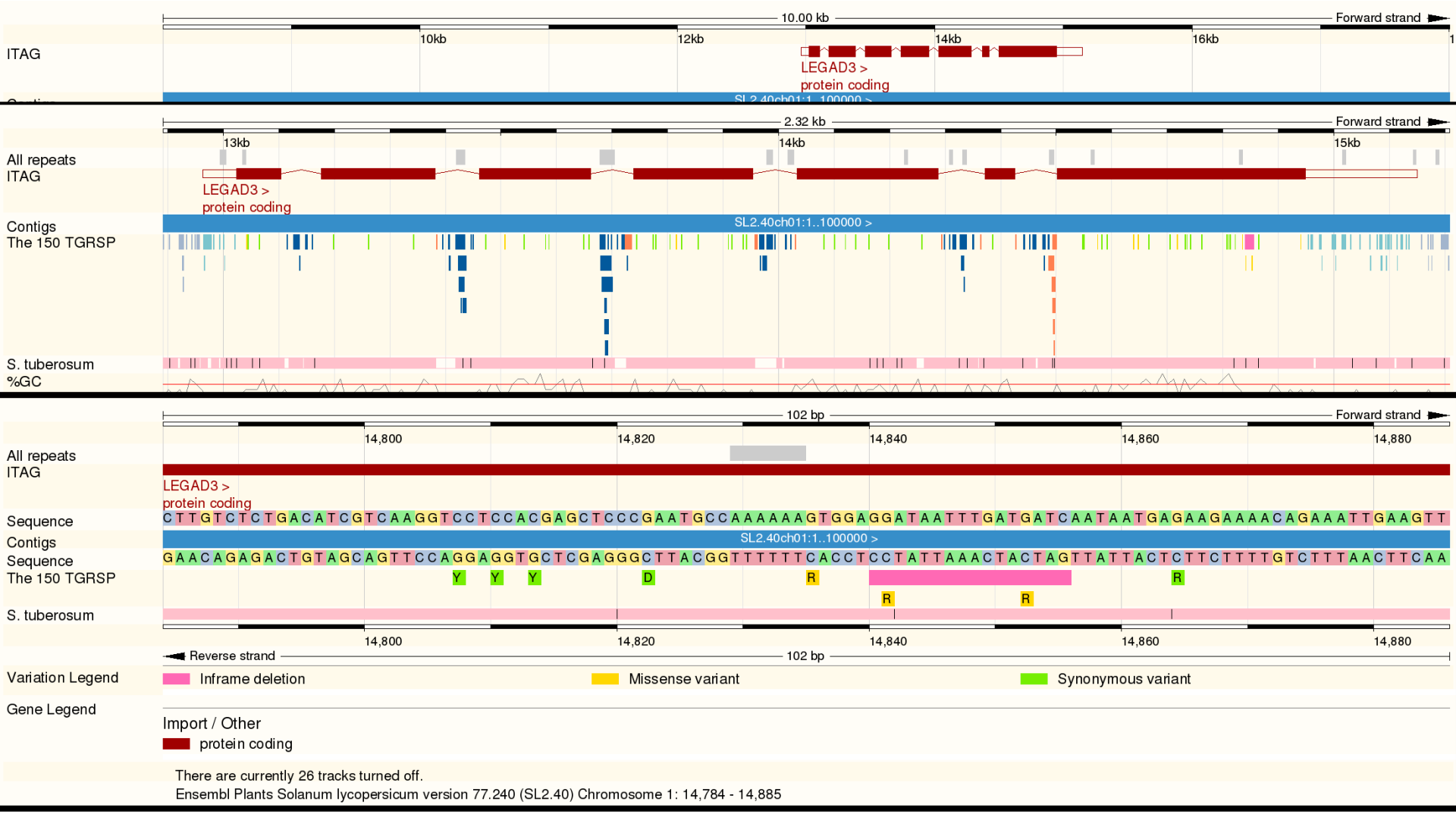
Population	Alleles A	Alleles G	Genotypes AlA	Genotyp
Zmays	0.147	0.853	0.147	0.853

Genotypes for Zmays

Show All entries Show/hide c

Individual	Genotype(forward
B73:MZ (U)	GIG
B97:MZ (U)	AIA
BKN015:MZ (U)	GIG
BKN016:MZ (U)	GIG
BKN017:MZ (U)	GIG
BKN018:MZ (U)	GIG
BKN022:MZ (U)	GIG
BKN023:MZ (U)	AIA
BKN025:MZ (U)	GIG
BKN027:MZ (U)	GIG
BKN029:MZ (U)	GIG
BKN031:MZ (U)	GIG
BKN032:MZ (U)	GIG
BKN033:MZ (U)	GIG
BKN034:MZ (U)	AIA
BKN035:MZ (U)	GIG
BKN040:MZ (U)	GIG
CAU5003:MZ (U)	AIA
CAUCHANG72:MZ (U)	AIA
CAUMO17:MZ (U)	AIA
CML103:MZ (U)	GIG
CML228:MZ (U)	GIG

Looking at some tomato variation data...



Variation displays
- Explore this variation
- Genomic context
  - Genes and regulation (18)
  - Flanking sequence
- Genotype frequency
- Individual genotypes (289)
- Linkage disequilibrium
- Phenotype Data (49)
- Phylogenetic Context
- Citations
- External Data

Configure this page

Add your data

Export data

Bookmark this page

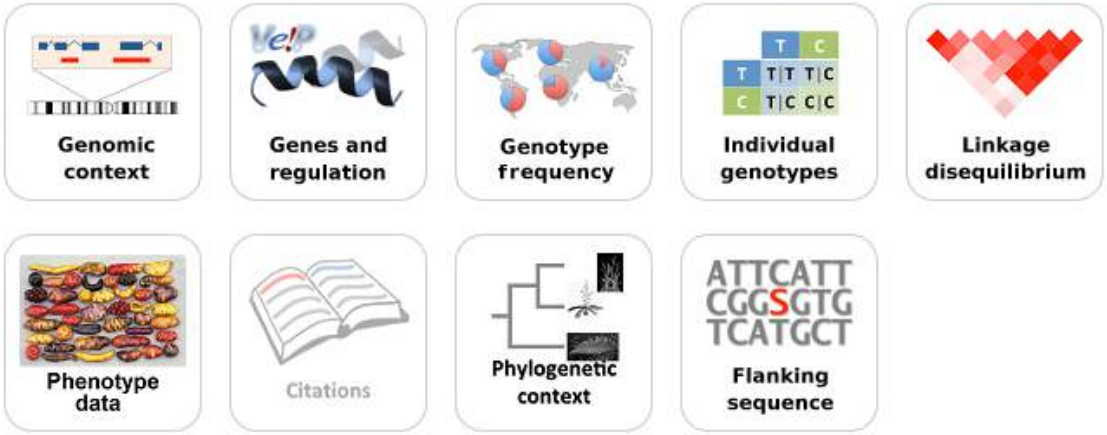
Share this page

Ensembl Plants is produced in collaboration with Gramene

ENSVATH00550254 SNP

Original source	Variation features from Affy_250k, Perlegen_1M, WTCHG and 1001 Genomes, with Ensembl identifiers | Ensembl
Alleles	C/N/T
Location	Chromosome **4:16056694** (forward strand) | View in location tab
Co-located	with **Ensembl** ENSVATH09775155 (C/-)
Synonyms ⊞	This variation has **2** synonyms - click the plus to show
HGVS name	4:g.16056694C>T

Explore this variation



Using the website

- Video: Browsing SNPs and CNVs in Ensembl
- Video: Clip: Genome Variation
- Video: BioMart: Variation IDs to HGNC Symbols
- Exercise: Genomes and SNPs in Malaria

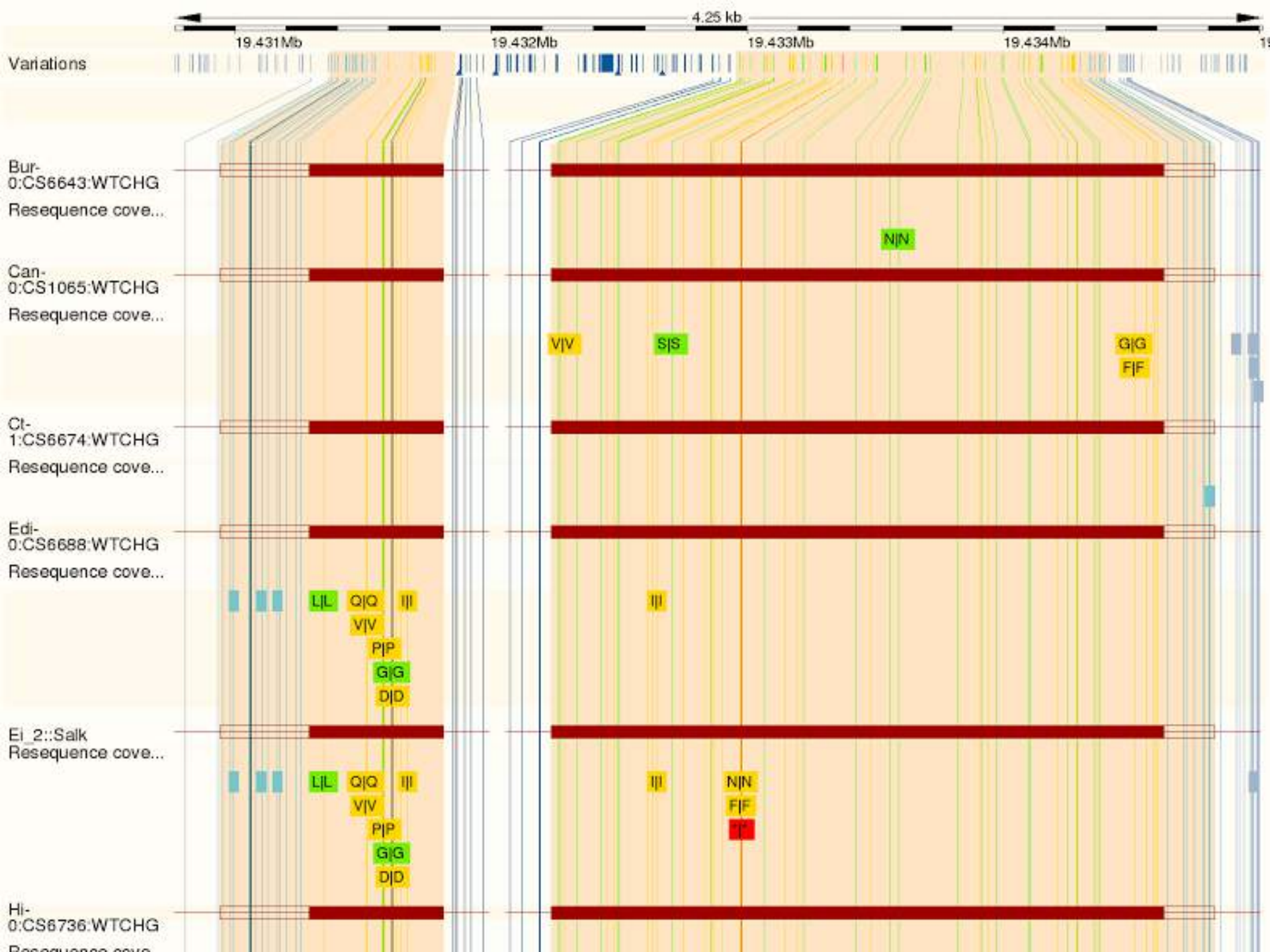
Analysing your data

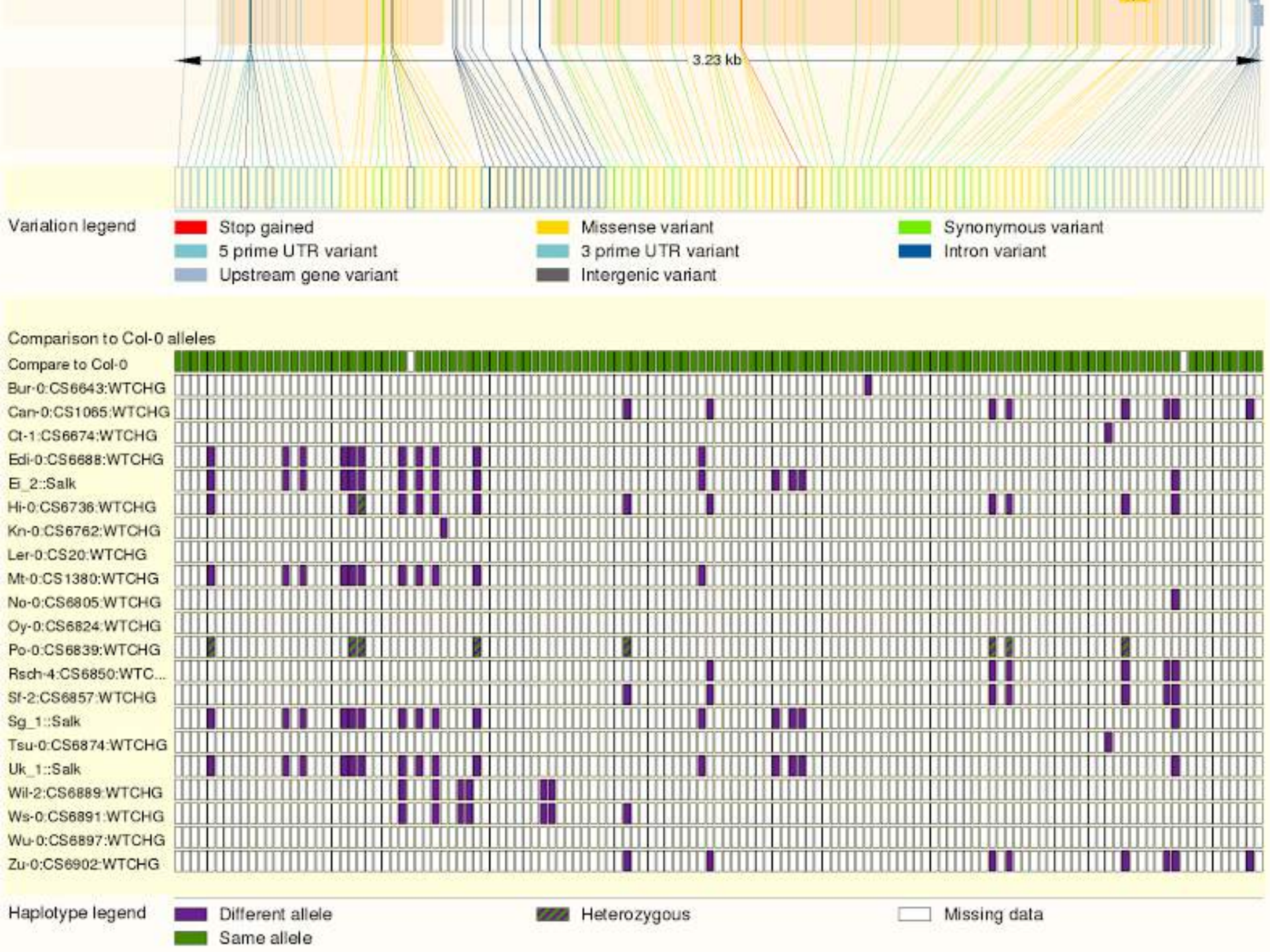
Programmatic access

- Tutorial: Accessing variation data with the Variation API

Reference materials

- Ensembl variation documentation portal
- Ensembl variation data description
- Variation Quick Reference card


4.25 kb
19.431Mb
19.432Mb
19.433Mb
19.434Mb
Variations
Bur-0:CS6643:WTCHG
Resequence cove...
Can-0:CS1065:WTCHG
Resequence cove...
Ct-1:CS6674:WTCHG
Resequence cove...
Edi-0:CS6688:WTCHG
Resequence cove...
Ei_2::Salk
Resequence cove...
Hi-0:CS6736:WTCHG
N|N
V|V
S|S
G|G
F|F
L|L
Q|Q
I|I
V|V
P|P
G|G
D|D
I|I
L|L
Q|Q
I|I
V|V
P|P
G|G
D|D
I|I
N|N
F|F
|


3.23 kb
Variation legend
Stop gained
Missense variant
Synonymous variant
5 prime UTR variant
3 prime UTR variant
Intron variant
Upstream gene variant
Intergenic variant
Comparison to Col-0 alleles
Compare to Col-0
Bur-0:CS6643:WTCHG
Can-0:CS1065:WTCHG
Ct-1:CS6674:WTCHG
Edi-0:CS6688:WTCHG
Ei_2::Salk
Hi-0:CS6736:WTCHG
Kn-0:CS6762:WTCHG
Ler-0:CS20:WTCHG
Mt-0:CS1380:WTCHG
No-0:CS6805:WTCHG
Oy-0:CS6824:WTCHG
Po-0:CS6839:WTCHG
Rsch-4:CS6850:WTC...
Sf-2:CS6857:WTCHG
Sg_1::Salk
Tsu-0:CS6874:WTCHG
Uk_1::Salk
Wil-2:CS6889:WTCHG
Ws-0:CS6891:WTCHG
Wu-0:CS6897:WTCHG
Zu-0:CS6902:WTCHG
Haplotype legend
Different allele
Same allele
Heterozygous
Missing data

EXERCISES

Variation in the Ensembl Browser

Exercise 1 – Exploring a SNP in Arabidopsis

The arabidopsis ATCDSP32 gene is a chloroplastic, 32 kDa, drought-induced stress protein that is proposed to participate in cell redox homeostasis (GO: 0045454).

A. How many variants have been identified in the gene that can cause a change in the protein sequence?
B. What is the ID of the variant that change the residue 60 from Alanine to Threonine? What is the location of this SNP in the Arabidopsis genome? What are its possible alleles?
C. Download the flanking sequence of this SNP in RTF (Rich Text Format). Can you change how much flanking sequence is displayed on the browser?
D. Does this SNP cause a change at the amino acid level for other genes or transcripts?
E. What is the most frequent genotype at this locus in the '1001 Population'?

Exercise 2 – Variation data in the tomato (*S. lycopersicum*) genome

Nearly every aspect of plant biology is dependent on cytochrome P450 enzymes, including metabolism of secondary metabolites in tomato involved in fruit development and ripening.

A. Find a cytochrome P450 gene in tomato known as Solyc02g085360.2 and go to its Location tab. Can you add the data track that shows variation data from the '150 Tomato Genome ReSequencing Project' (TGRSP)?

B. Zoom in around the last exon of this gene. What are the different types of variants seen in that region? What is the location of the only inframe deletion mapped in the region?

C. Click on the splice region variant showed in that view. Why does Ensembl Plants put the G allele first in the string (G/A)?

Exercise 3 – Missense variants in the bread wheat genome

The bread wheat transcript Traes_2AL_5C7E76139.1 is involved in calcium ion binding (GO:0005509). Around 45 variants have been mapped to this transcript.

A. What are the two types of variants annotated in this transcript?
B. Are there any variants predicted to be deleterious? Which amino acid residue is affected and what are the possible amino acids in that position?
C. What are the different sources of EPITAEV06993600 and BA00258972?

Displaying your data in Ensembl Plants

Dan Bolser
Denise Carvalho-Silva
Brandon Walts
EMBL-EBI

http://plants.ensembl.org

Visualise your own data

Upload data:

- Data saved on server
- 5 MB limit
  - Large file formats?

Attach remote files:

- URL-based
  - HTTP or FTP
- No size limit

Upload formats:

BED	genes / features
Gbrowse	genes / features
GFF/GTF	genes / features
PSL	sequence alignments
WIG	continuous-valued data
BedGraph	continuous-valued data
TrackHub	collections of tracks

Attach formats:

BigBed	genes / features
BAM	sequence alignments
BigWig	continuous-valued data
VCF	variants

User added tracks:

- Can be saved or shared
- Only trivial security, do not use for sensitive data!

Live demo

Variant Effect Predictor

Dan Bolser
Denise Carvalho-Silva
Brandon Walts
EMBL-EBI

http://plants.ensembl.org

Variant Effect Predictor

- Maps variants on Ensembl transcripts and predicts their consequences

- Different input formats
  - Ensembl default
  - list of rs IDs
  - VCF
  - Pileup
  - HGVS


BIOINFORMATICS APPLICATIONS NOTE

Databases and ontologies

Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor

William McLaren[1,*], Bethan Pritchard[2], Daniel Rios[1], Yuan Chen[1], Paul Flicek[1] and Fiona Cunningham[1,*]

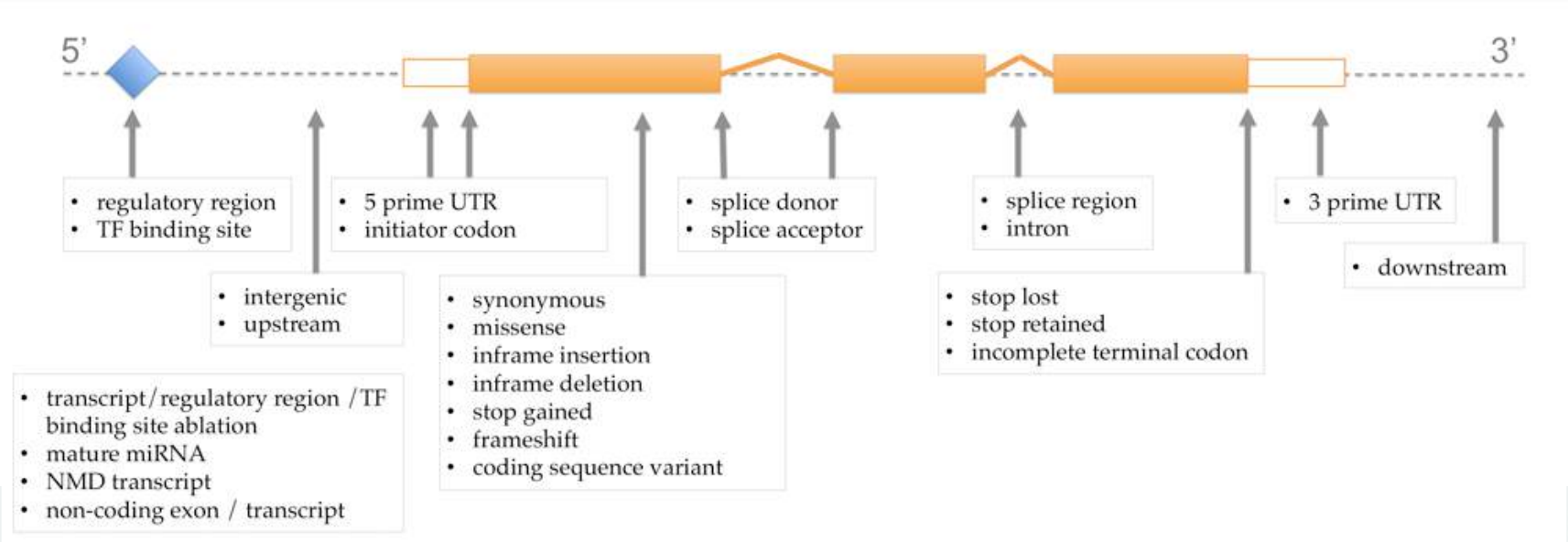
[1]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD and [2]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

PMID: 20562413

- Versions: Rest API, Perl script and online interface

Variant Effect Predictor (VEP)

- Predicts functional consequences of known and unknown variants
- For substitutions, insertions, deletions and structural variants
- Web interface (for up to 750 variants), standalone Perl script, Perl API and REST API



psbO, Oxygen-evolving enhancer protein 1 Traes_2BL_4B8B77E73.1

Variations

Show All entries | Show/hide columns | Filter

Residue	Variation ID	Type	Evidence	Alleles	Ambig. code	Residues	Codons	SIFT
35	BA00804152.2	Synonymous variant		G/A	R	V	GT**C**, GT**T**	-
48	BA00039145	Missense variant		T/C	Y	D, G	G**A**C, G**G**C	0.03
103	BS00179246.2	Missense variant		C/A	M	G, C	**G**GC, **T**GC	0
119	BA00233383.2	Synonymous variant		G/A	R	P	CC**C**, CC**T**	-
124	BA00897028.2	Synonymous variant		C/T	Y	K	AA**G**, AA**A**	-
136	BS00178628.2	Synonymous variant		G/A	R	S	TC**C**, TC**T**	-
139	BA00227489	Synonymous variant		G/C	S	V	GT**C**, GT**G**	-
154	BA00269810.2	Synonymous variant		C/T	Y	K	AA**G**, AA**A**	-
165	BA00829191	Missense variant		G/T	K	L, I	**C**TT, **A**TT	0.11
172	BA00058887.2	Synonymous variant		G/C	S	L	CT**C**, CT**G**	-
181	BA00863963	Missense variant		T/C	Y	K, R	AA**G**, AG**G**	0.27
190	BA00326027	Synonymous variant		A/G	R	Y	TA**T**, TA**C**	-
192	BA00121558	Synonymous variant		G/A	R	A	GC**C**, GC**T**	-
250	BA00796319	Synonymous variant		G/C	S	A	GC**C**, GC**G**	-
253	BA00078190	Synonymous variant		T/C	Y	L	CT**A**, CT**G**	-
264	BA00706748	Synonymous variant		A/G	R	L	**T**TG, **C**TG	-
276	BA00323812	Synonymous variant		T/C	Y	T	AC**A**, AC**G**	-
286	BS00116511	Missense variant		C/T	Y	S, N	AG**C**, AA**C**	0
295	BA00298030.2	Synonymous variant		G/A	R	G	GG**C**, GG**T**	-

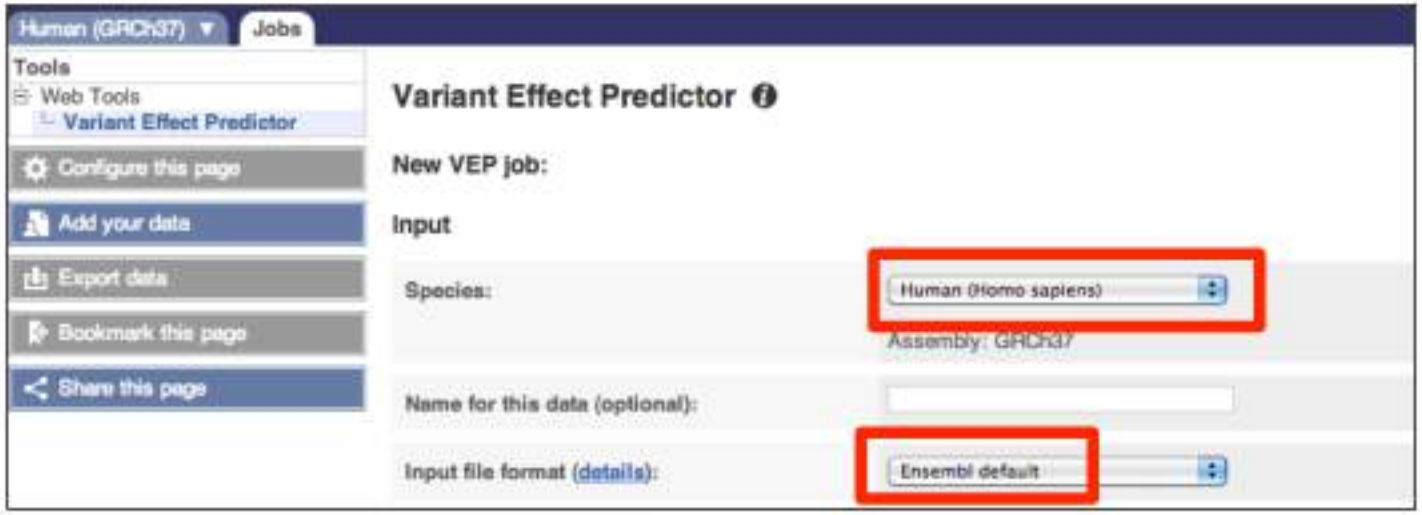
Ensembl tools

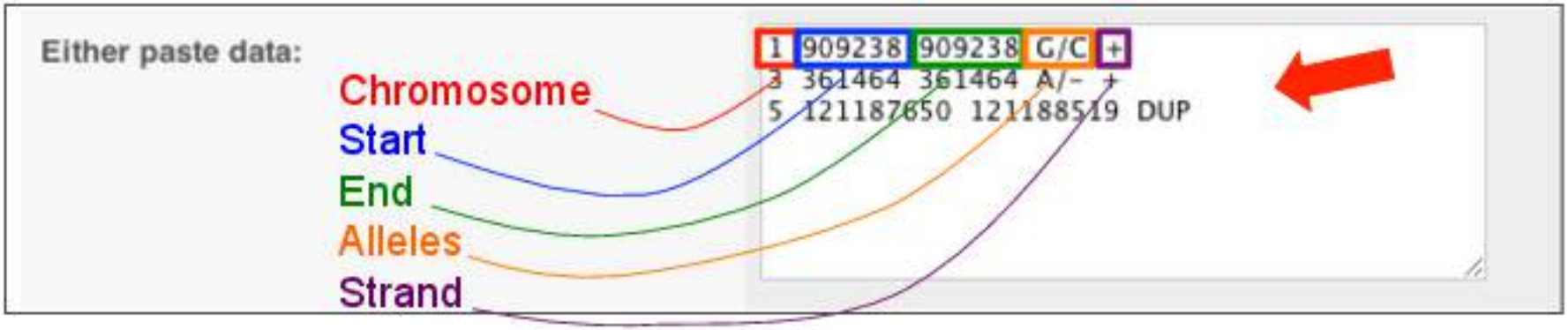
http://www.ensembl.org/tools.html

Name	Description	Online tool	Download code	Documentation
Variant Effect Predictor VeP	Analyse your own variants and predict the functional consequences of known and unknown variants via our Variant Effect Predictor (VEP) tool.	🔧	⬇	ⓘ
BLAST/BLAT	Search our genomes for your DNA or protein sequence.	🔧		ⓘ
BioMart	Use this data-mining tool to export custom datasets from Ensembl.	🔧	⬇	ⓘ
Assembly converter	Map (liftover) your data's coordinates to the current assembly.	🔧	⬇	
ID History converter	Convert a set of Ensembl IDs from a previous release into their current equivalents.	🔧	⬇	
Ensembl Virtual Machine	VirtualBox virtual Machine with Ubuntu desktop and pre-configured with the latest Ensembl API plus Variant Effect Predictor (VEP). **NB: download is >1 GB**	-	⬇	ⓘ

http://www.ensembl.org/vep

Inputting data into VeP





EXERCISE

Displaying your data in Ensembl Plants

Exercise 4 – The Variant Effect Predictor in the bread wheat genome

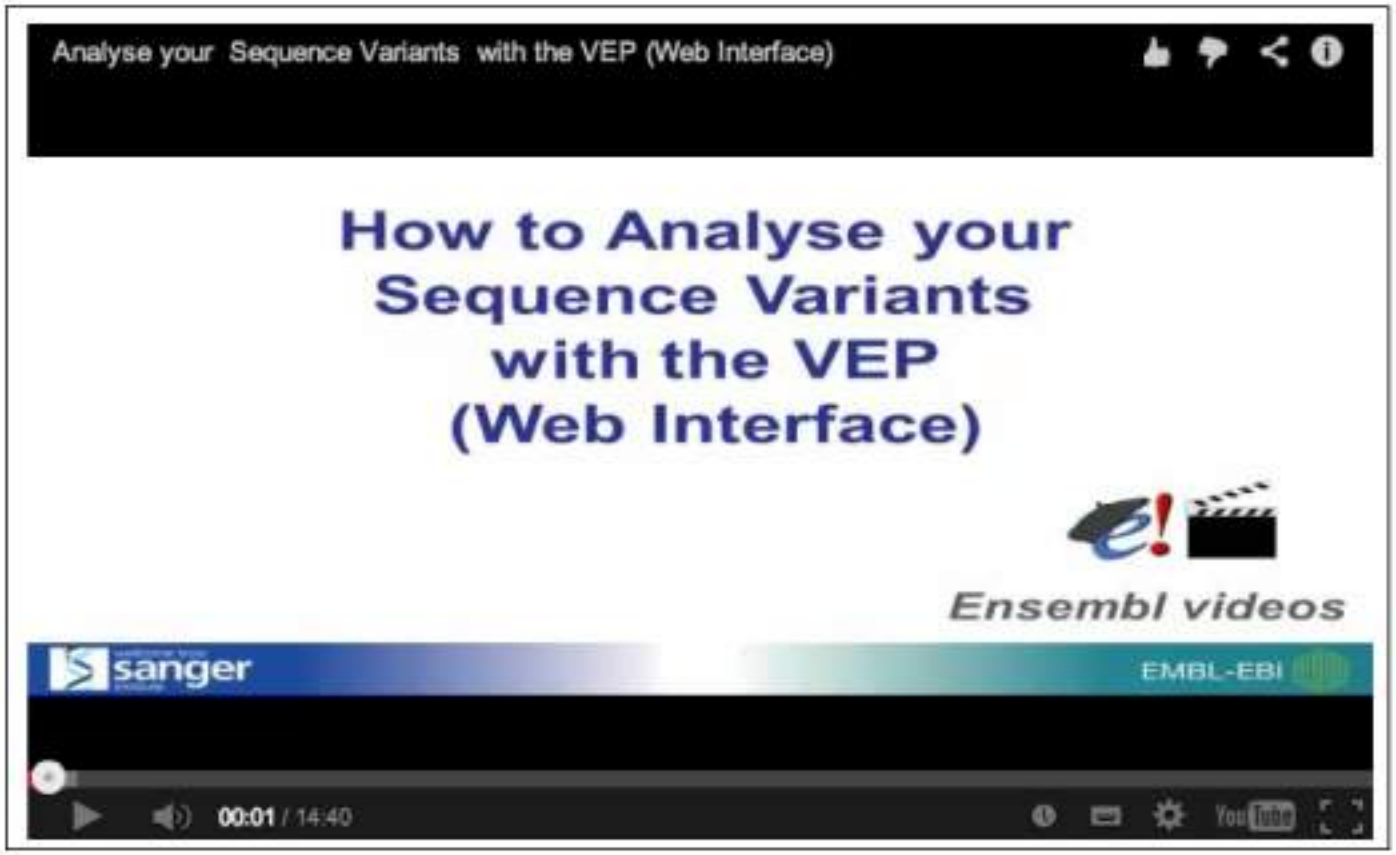
An analysis of 5,000 individuals from two different populations of bread wheat (*T. aestivum*) has identified thousands of polymorphic loci. See a list of a few of them on the right:

Can you use the VEP tool to answer the following?

A. Which genes and transcripts do these variants map to?
B. Which consequence types can be found for these variants?
    - Do any of them cause a change at the amino acid level?

```
chr 2D, genomic coordinate  89551917,
alleles G/A, forward strand

chr 2D, genomic coordinate 148408765,
alleles G/T, forward strand

chr 3D, genomic coordinate 113574123,
alleles C/A, forward strand

chr 3D, genomic coordinate 93827883,
alleles G/A, forward strand

chr 3B, genomic coordinate 727928129,
alleles C/T, forward strand

chr 3B, genomic coordinate 736734474,
alleles C/T, forward strand

chr 6A, genomic coordinate 196872409,
alleles T/G, forward strand

chr 6A, genomic coordinate 196153918,
alleles A/G, forward strand

chr 6A, genomic coordinate 196774882,
alleles G/C, forward strand
```

VEP YouTube video



http://tinyurl.com/vep-video

Ensembl Plants Tools: BioMart

Denise Carvalho-Silva
Ensembl Outreach Officer
EMBL-EBI

http://plants.ensembl.org

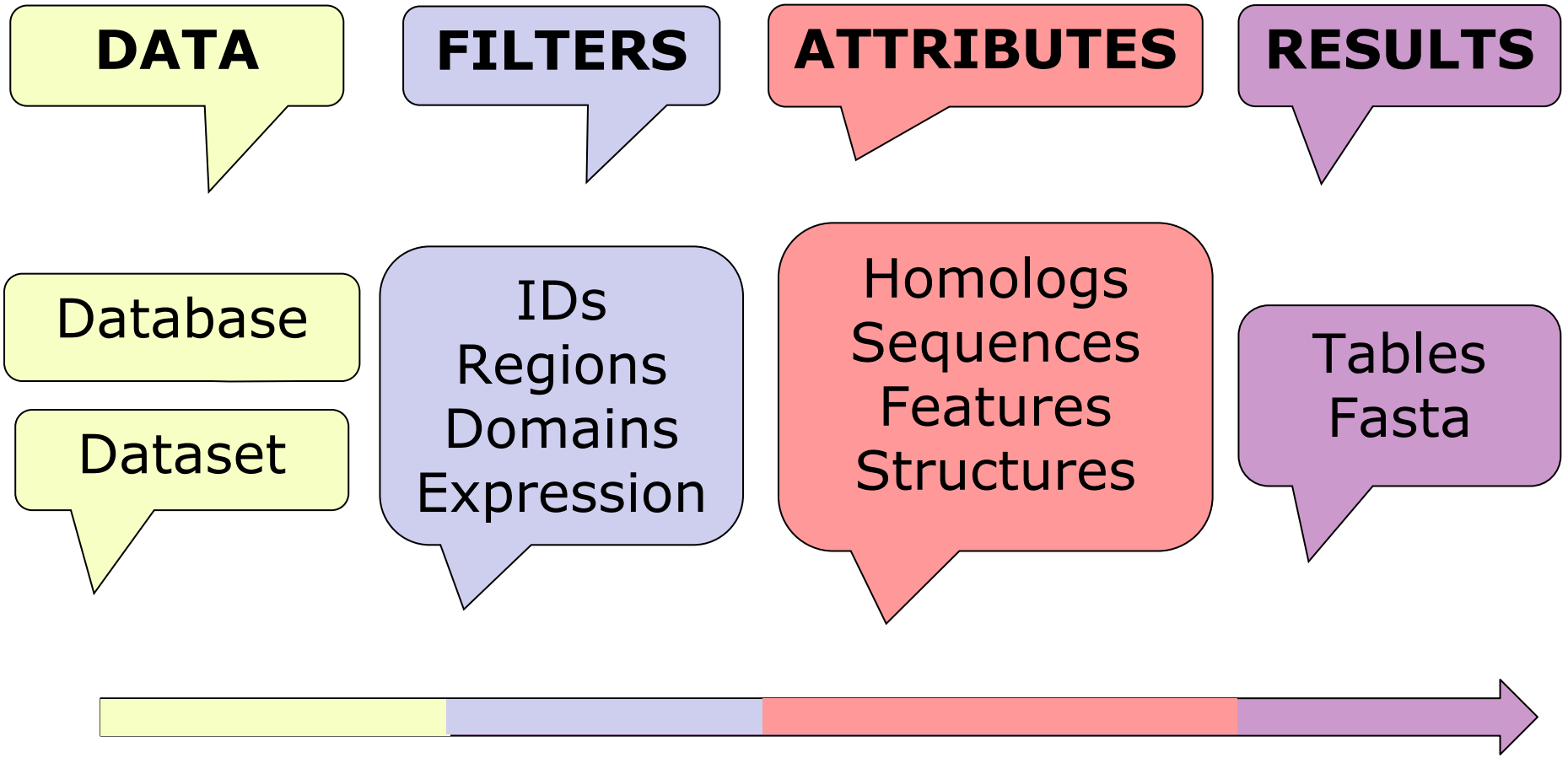
Outline

- Definitions
- The principle: 4 steps
- Tutorial: simple query in Arabidopsis
- Find Ensembl BioMart and BioMart elsewhere
- Sophisticated platforms: mart services, APIs, etc…
- Exercises

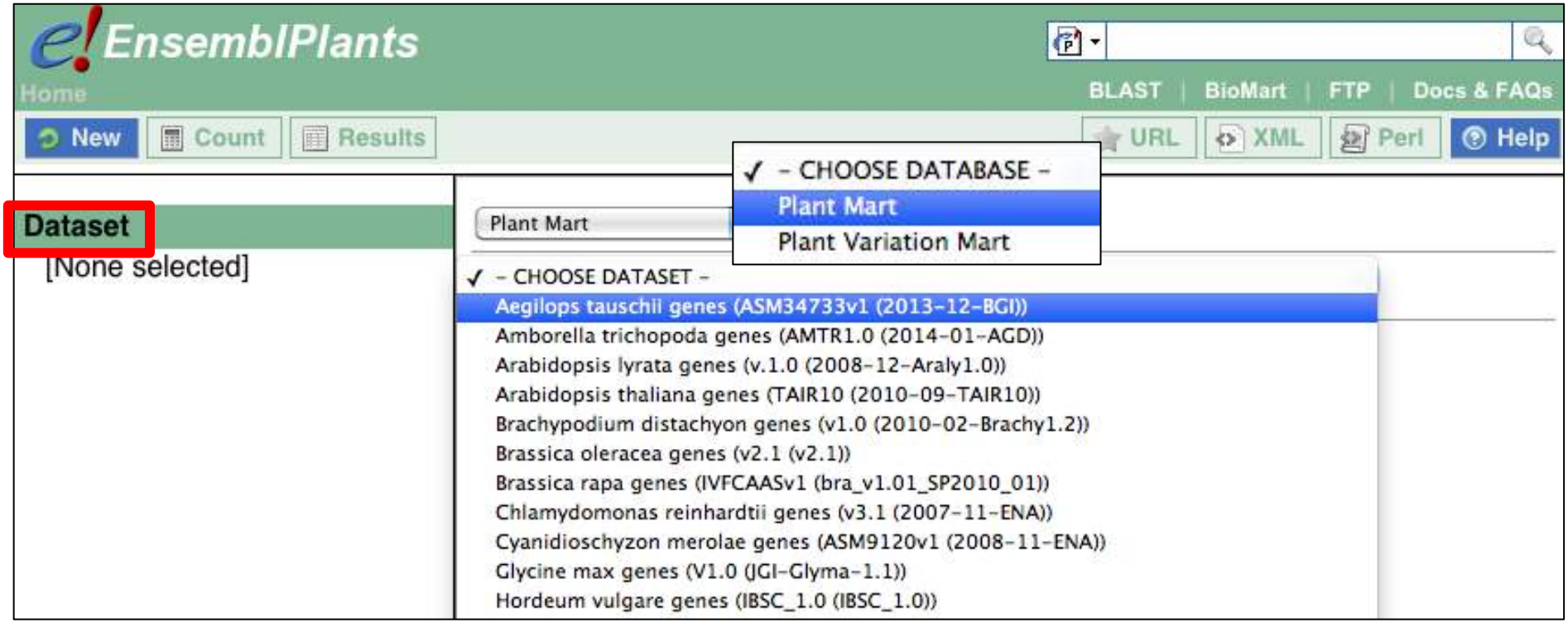
What is BioMart?

- Free service for easy retrieval of Ensembl data
- Data export tool with little/no programming required
- Complex queries with a few mouse clicks
- Output formats (.xls, .csv, fasta, tsv, html)

The four-step principle

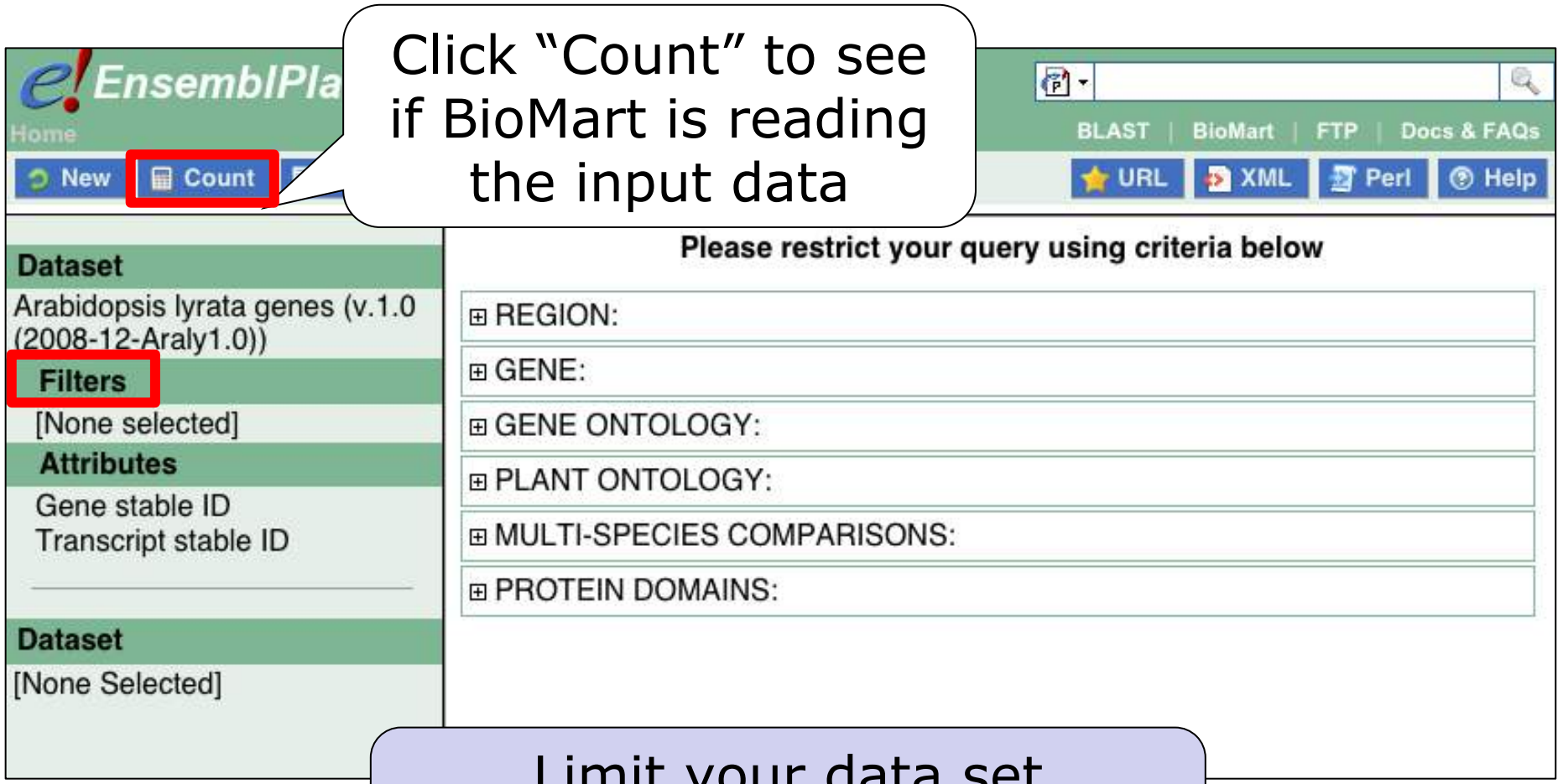


Choosing the data

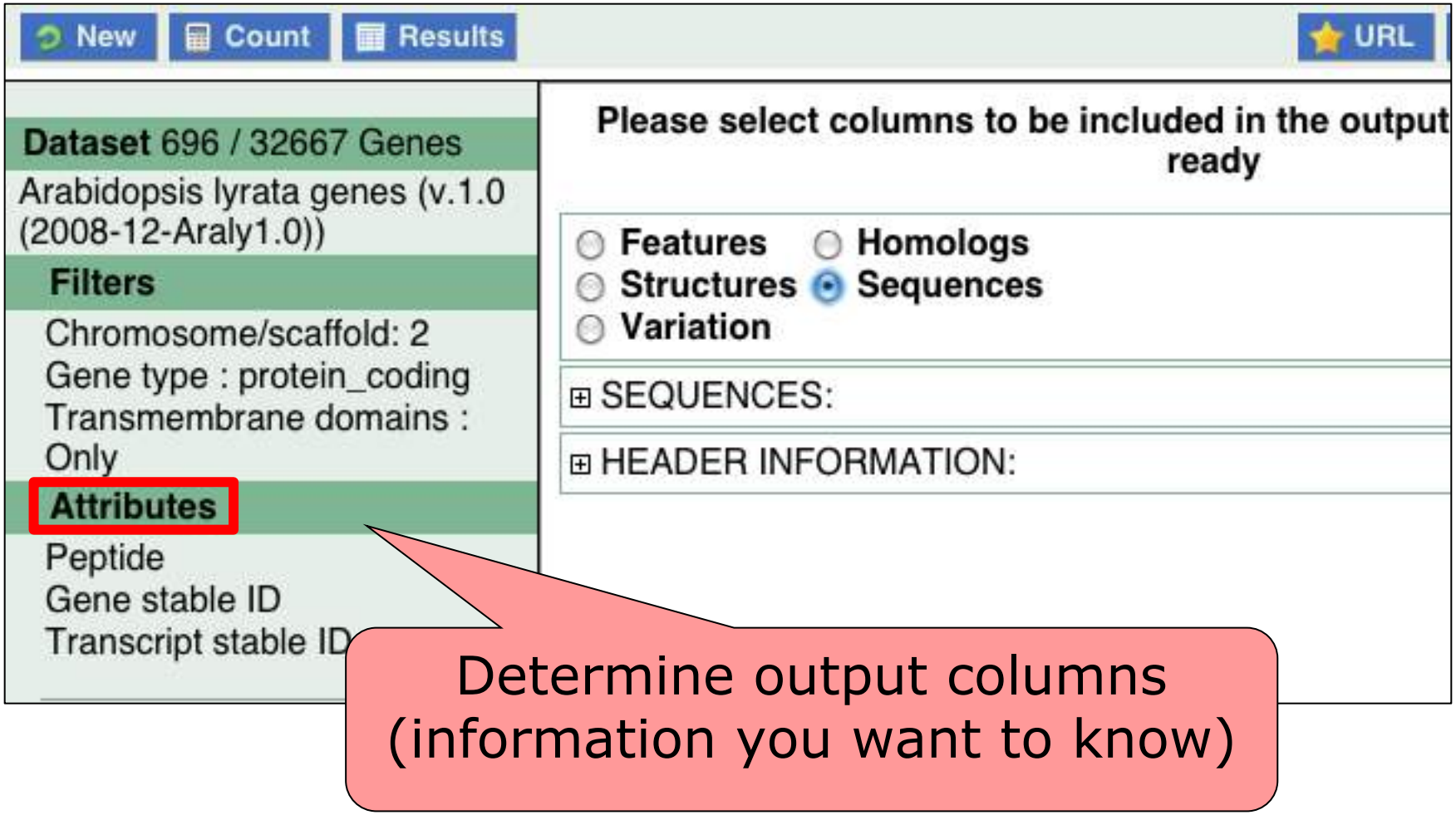


Database and dataset

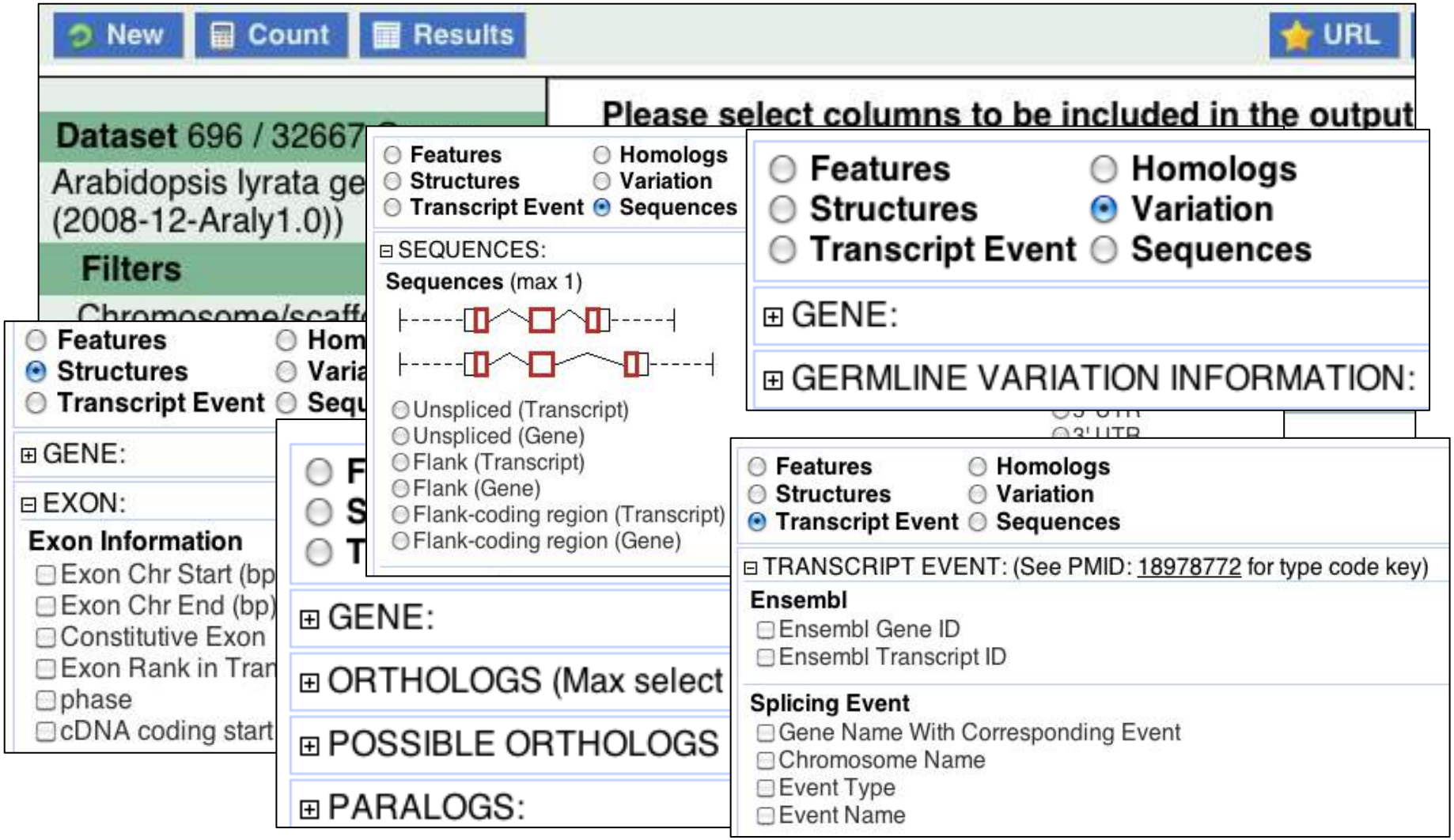
Selecting the filters



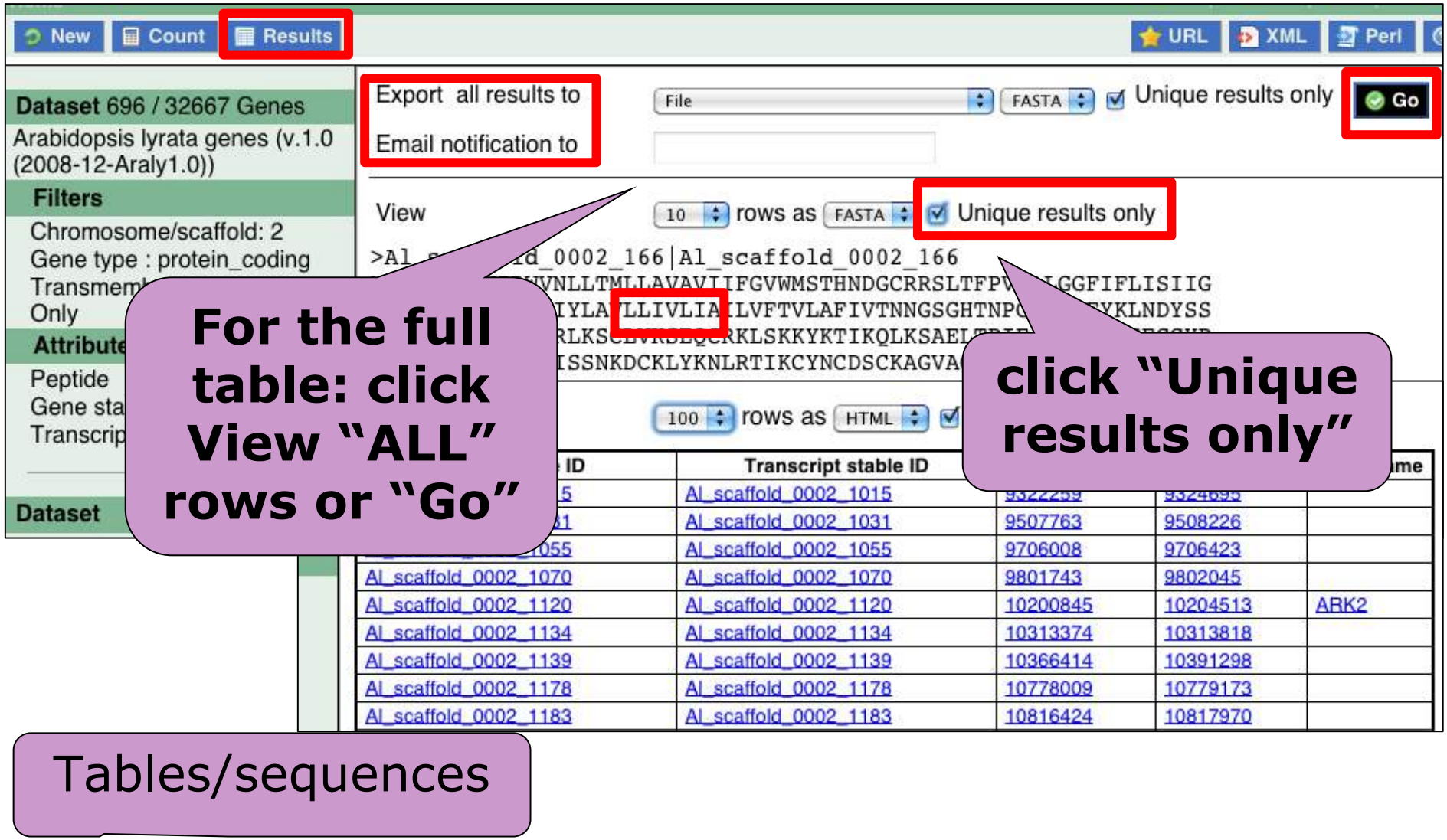
Picking the attributes



The different attributes



Getting the results



This Article | Info for Authors | Subscribe | About

PNAS

Proceedings of the National Academy of Sciences of the United States of America

A genome-wide transcriptional analysis using *Arabidopsis thaliana* Affymetrix gene chips determined plant responses to phosphate deprivation

Julie Misson*, Kashchandra G. Raghothama†, Ajay Jain†, Juliette Jouhet‡, Maryse A. Block‡, Richard Bligny‡, Philippe Ortet§, Audrey Creff*, Shauna Somerville¶, Norbert Rolland‡, Patrick Doumas‖, Philippe Nacry‖, Luis Herrerra-Estrella**, Laurent Nussaume*, and Marie-Christine Thibaud*††

11934–11939 | PNAS | August 16, 2005 | vol. 102 | no. 33

www.pnas.org/cgi/doi/10.1073/pnas.0505266102

Selected ATH1 probes (Affy chip)

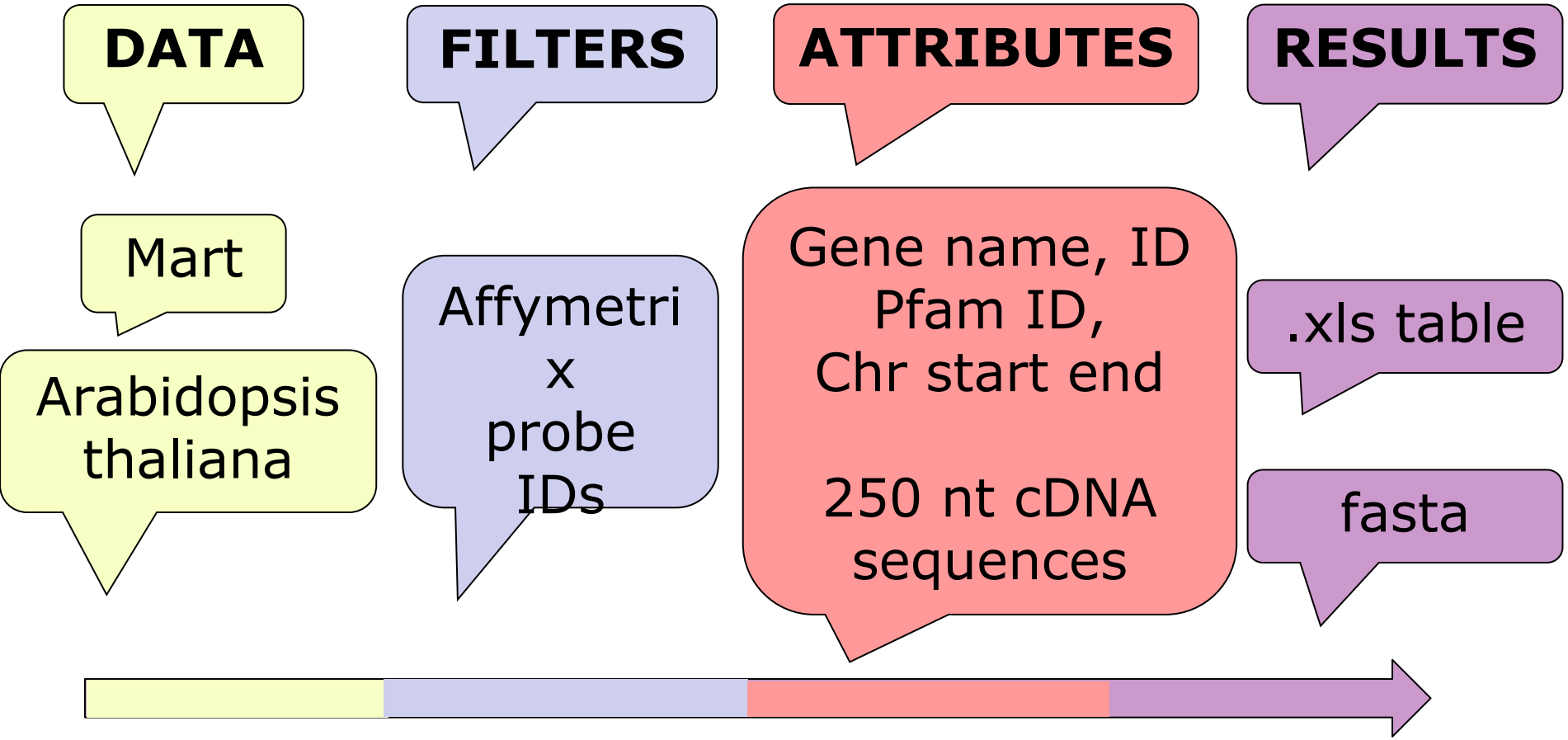
259842_at, 251193_at, 259303_at, 252534_at, 266957_at, 257891_at, 263593_at, 266372_at, 265342_at, 254011_at, 260623_at, 262238_at, 264118_at, 256910_at, 263846_at, 249996_at, 248094_at, 267361_at, 246275_at, 258034_at, 248622_at, 263483_at, 254250_at, 257964_at

Tutorial: BioMart

For this list of probes (Affymetrix array Arabidopsis), can I use BioMart to retrieve

- a table with names and IDs, Pfam IDs, and location of the genes the probes map to?
- a fasta (sequence) file with 250 nt upstream of the cDNA sequences of those genes?

The four-step principle



Live demo

BioMart started at Ensembl

EMBL-EBI

Ontario Institute for Cancer Research

BioMart Central Portal

Home

Identifier Search

Go

Examples: KRAS, ENSG00000146648

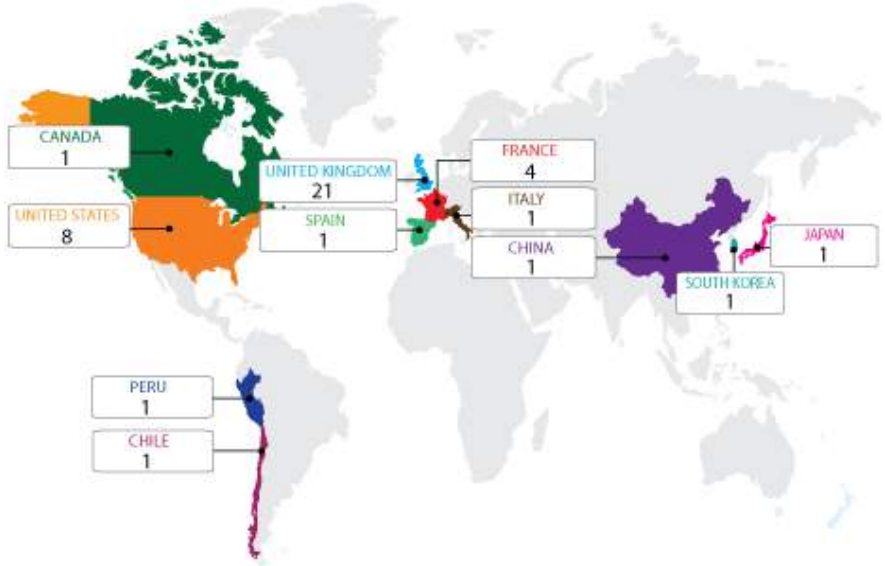
Tools

Gene retrieval | Variant retrieval | Sequence retrieval | ID converter

Ensembl
COSMIC
VEGA
MGI
EnsemblGenomes
Gramene
Rfam
PRIDE
UniProt
InterPro
Reactome
and many more…

BioMart Central Portal

Databases: **41**



EMBL-EBI

Find BioMart

http://plants.ensembl.org/biomart/martview/

central.biomart.org

More sophisticated platforms

- BioMart 

queries: MartService

  www.biomart.org/martservice.html

- APIs: PERL, Java, Web Services
- Third party softwares

taverna.org.uk

biomaRt

bioconductor.org

galaxyproject.org

Ensembl BioMarts

DATABASE
The Journal of Biological Databases and Curation

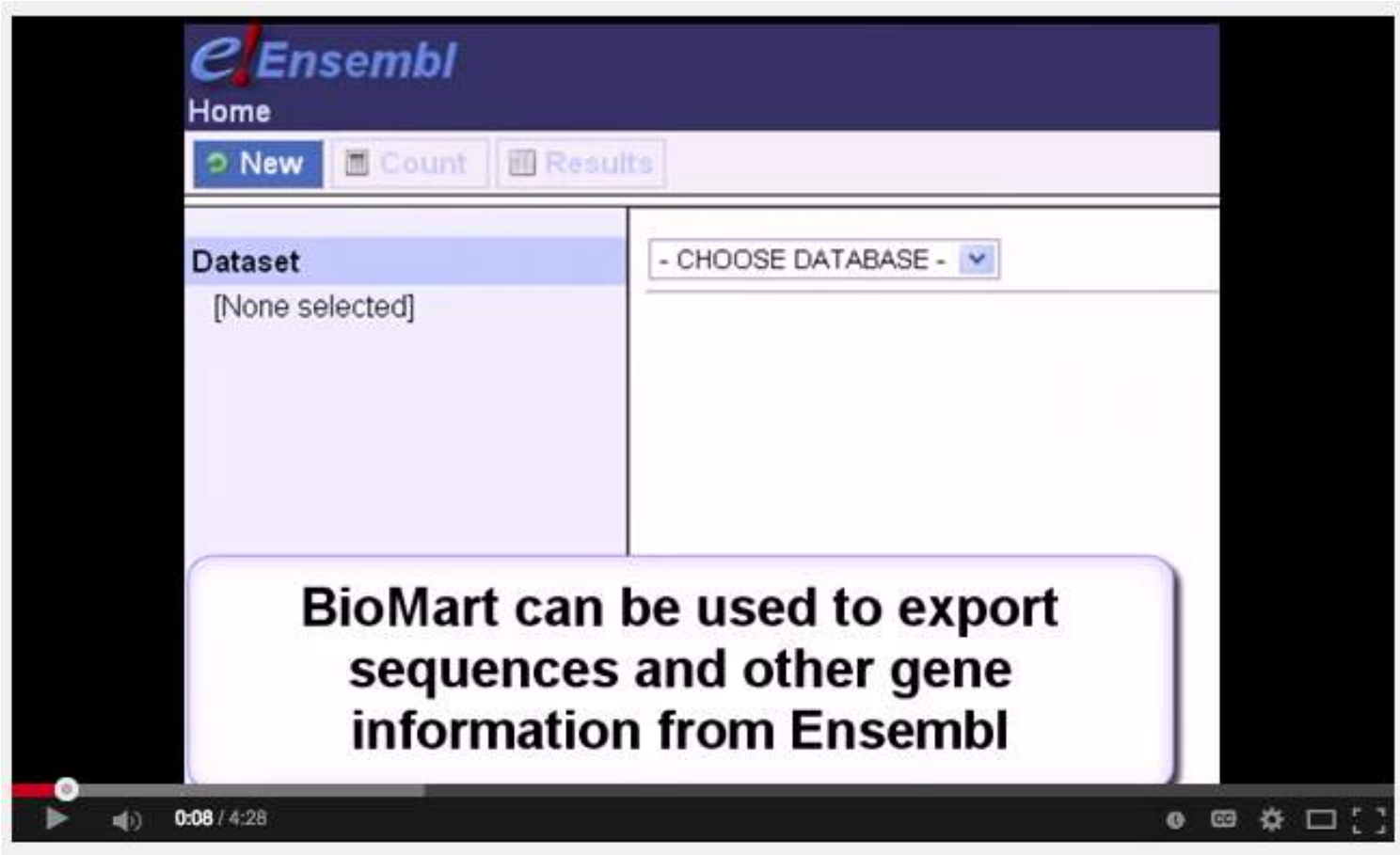
Database, Vol. 2011, Article ID bar030, doi:10.1093/database/bar030

Original article

Ensembl BioMarts: a hub for data retrieval across taxonomic space

Rhoda J. Kinsella[1,*], Andreas Kähäri[1], Syed Haider[2], Jorge Zamora[1], Glenn Proctor[1], Giulietta Spudich[1], Jeff Almeida-King[1], Daniel Staines[1], Paul Derwent[1], Arnaud Kerhornou[1], Paul Kersey[1] and Paul Flicek[1,*]

BioMart YouTube video



http://tinyurl.com/video-biomart

Connect with Ensembl Genomes

helpdesk@ensemblgenomes.org

twitter.com/ensemblgenomes

EXERCISE

Mining variation data with BioMart

Exercise 5 – Retrieve a list of SNPs from tomato (S. lycopersicum)

The region between coordinates 21,394,819 and 21,397,868 on chromosome 6 in tomato contains a gene involved in oxidation-reduction process (GO: 0055114).

Can you use BioMart to retrieve all the SNPs that cause a change at the amino acid level of this gene (those SNPs are known as missense variants) including their IDs and possible alleles?

Exercise 6 – Open-ended BioMart

- Filter for some ‘interesting variations’ in a ‘region of interest’...

Export data and gather some ‘interesting’ statistics.

ANSWERS

Answer 1 – Exploring a SNP in Arabidopsis

A. Search for the ATCDSP32 gene the Arabidopsis page in Ensembl Plants. On the left hand side menu of the Gene tab, click on Genetic variation to find out the number of variants that cause a change at the amino acid level of the ATCDSP32 protein. Those are know as missense variants. There are 16 of them.

B. Next to the 'Missense variant' under the column Type, click on 'Show'. Look for the residue 60, where a change from A to T is reported. The ID of this variant is ENSVATH05153232, located at position 28549171 on chromosome 1. The two possible alleles at this locus are C and T.

C. Click on the link ENSVATH05153232. Then click on 'Flanking sequence' in the left hand side menu. Now click on 'Download sequence' and select the Rich Text Format (RTF). If you want to change how much flanking sequence is displayed on the browser, go back to the Flanking sequence page, click in the Configuration page and change the length of the sequence. The default settings is 400 bp.

D. Click on 'Genes and regulation' to find out this SNP does not cause a change at the amino acid level for any other genes or transcripts in that genome. The other consequence types for this SNP are Downstream gene variant and Upstream gene variant.

E. Click on 'Genotype frequency' in the left hand side menu to find out the most frequent genotype at this locus in the '1001 Population' is CC (homozygous for allele C, the reference allele).

Answer 2 – Variation data in the tomato (S. lycopersicum) genome

A. Go to the tomato page in the Ensembl Plants Genome Browser and search for Solyc02g085360.2 and click on the Location link in the results page. Configure the page and look for the data track named ‘The 150 TGRSP variations’.

B. Zoom in around the last exon of this gene by drawing a box in the respective region. Please note the gene is on the reverse strand, so the last exon will be on the left hand side of that image. The types of variants seen in that region are 3 primer UTR variants, missense variants, synonymous variants and a few others. Look for the Variation legend of the different colour codes. The location of the only inframe deletion mapped in the region is 2:42861257-42861272. Just click on the pink box that denotes inframe deletion for a pop-up box with additional information on that variant including location.

C. The splice region variant is coded in orange. Its reported alleles are G/A, G being the reference allele, and A the alternative allele.

Answer 3 - Missense variants in the bread wheat genome

A. Search for Traes_2AL_5C7E76139.1 in bread wheat, click on the Gene ID in the results page and then on the transcript ID in the Transcript tab. Now click on ‘Variations’ to find out the two types are ‘missense variants’ and ‘synonymous variants’.

B. One missense variants are predicted to be deleterious according to SIFT. The possible amino acid residues are D (Aspartic acid) and E (Glutamic acid).

C. This missense variant has been reported by two different sources (inter-homoeologous variants from Ensembl Genomes and variants from CerealsDB) and gets two IDs, i.e. EPITAEV06993600 and BA00258972 as it’s a polymorphism within the different components of the hexaploid genome and between different populations of bread wheat.

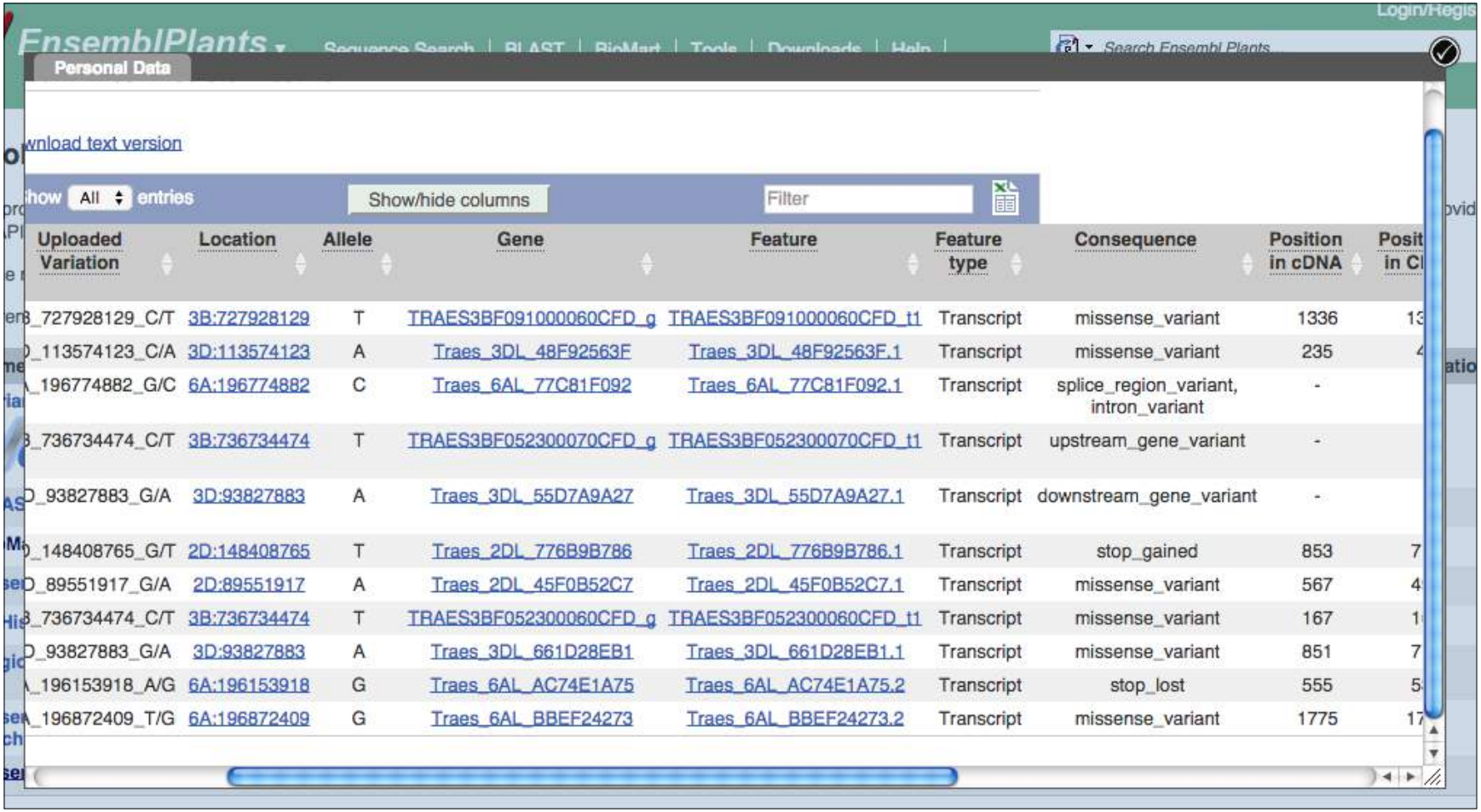
Answer 4: The VEP tool, Variant Effect Predictor

Go the plants.ensembl.org and click on Tools on the top banner of the page. Click on VEP and select the species T. aestivum. Type in the following input data:

Scroll down in that page and click Next. Click on ‘HTML’ and you will see a table like this:

```
2D  89551917 89551917 G/A
2D 148408765 148408765 G/T
3D 113574123 113574123 C/A
3D 93827883 93827883 G/A
3B 727928129 727928129 C/T
3B 736734474 736734474 C/T
6A 196872409 196872409 T/G
6A 196153918 196153918 A/G
6A 196774882 196774882 G/C
```

Answer 4: The VEP tool, Variant Effect Predictor



Uploaded Variation	Location	Allele	Gene	Feature	Feature type	Consequence	Position in cDNA	Posit in C
3_727928129_C/T	3B:727928129	T	TRAES3BF091000060CFD_g	TRAES3BF091000060CFD_t1	Transcript	missense_variant	1336	13
D_113574123_C/A	3D:113574123	A	Traes_3DL_48F92563F	Traes_3DL_48F92563F.1	Transcript	missense_variant	235	4
A_196774882_G/C	6A:196774882	C	Traes_6AL_77C81F092	Traes_6AL_77C81F092.1	Transcript	splice_region_variant, intron_variant	-	
3_736734474_C/T	3B:736734474	T	TRAES3BF052300070CFD_g	TRAES3BF052300070CFD_t1	Transcript	upstream_gene_variant	-	
D_93827883_G/A	3D:93827883	A	Traes_3DL_55D7A9A27	Traes_3DL_55D7A9A27.1	Transcript	downstream_gene_variant	-	
D_148408765_G/T	2D:148408765	T	Traes_2DL_776B9B786	Traes_2DL_776B9B786.1	Transcript	stop_gained	853	7
D_89551917_G/A	2D:89551917	A	Traes_2DL_45F0B52C7	Traes_2DL_45F0B52C7.1	Transcript	missense_variant	567	4
3_736734474_C/T	3B:736734474	T	TRAES3BF052300060CFD_g	TRAES3BF052300060CFD_t1	Transcript	missense_variant	167	1
D_93827883_G/A	3D:93827883	A	Traes_3DL_661D28EB1	Traes_3DL_661D28EB1.1	Transcript	missense_variant	851	7
A_196153918_A/G	6A:196153918	G	Traes_6AL_AC74E1A75	Traes_6AL_AC74E1A75.2	Transcript	stop_lost	555	5
A_196872409_T/G	6A:196872409	G	Traes_6AL_BBEF24273	Traes_6AL_BBEF24273.2	Transcript	missense_variant	1775	17

Answer 4: The VEP tool, Variant Effect Predictor

A. Several genes and transcripts have been annotated in the region where these variants map to, e.g.

   Traes_2DL_776B9B786 / Traes_2DL_776B9B786.1

   Traes_3DL_48F92563F / Traes_3DL_48F92563F.1

   Traes_6AL_AC74E1A75 / Traes_6AL_AC74E1A75.2

B. The consequences for these variants are missense, splice region, stop loss, downstream gene, among others. The missense variants do cause a change at the amino acid level.