

Part of the

transPLANT

European Plant
Genomics Infrastructure

Part 2

3rd transPLANT Training Workshop - October 2014
Exploiting and understanding Solanaceous genomes

Mining structural variation in solanaceous genomes

Dan Bolser
EMBL-EBI

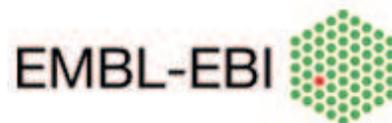




Ensembl Genomes: Overview

~~Versailles, 12th-13th November 2012~~

Wageningen, 13th-14th October 2014



Dan Bolser, EMBL-EBI

dbolser@ebi.ac.uk



The transPLANT project is funded by the European Commission within its 7th Framework Programme under the thematic area "Infrastructures". Contract number 283496.

www.transplantdb.eu



Ensembl Plants: Visualising, mining and analysing crop genomics data

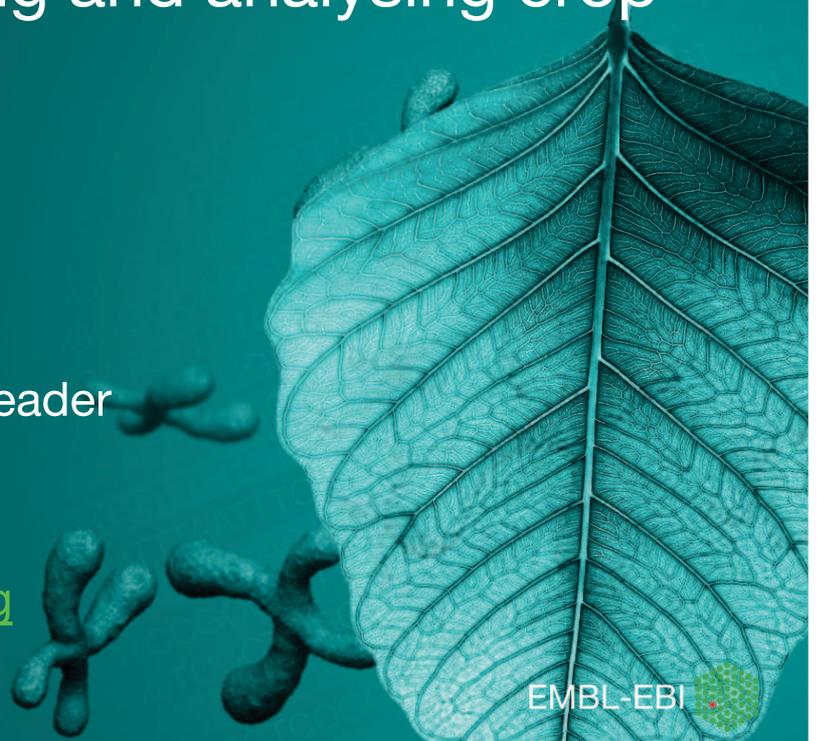
Dan Bolser

Ensembl Plants project leader

EMBL-EBI

<http://plants.ensembl.org>

#EnsemblGenomes



EMBL-EBI

Overview

Background:

- Ensembl Plants

- History
- Data



- Recent updates

- Wheat
- Barley



Visualising, mining and analysing data:

- The Ensembl genome browser



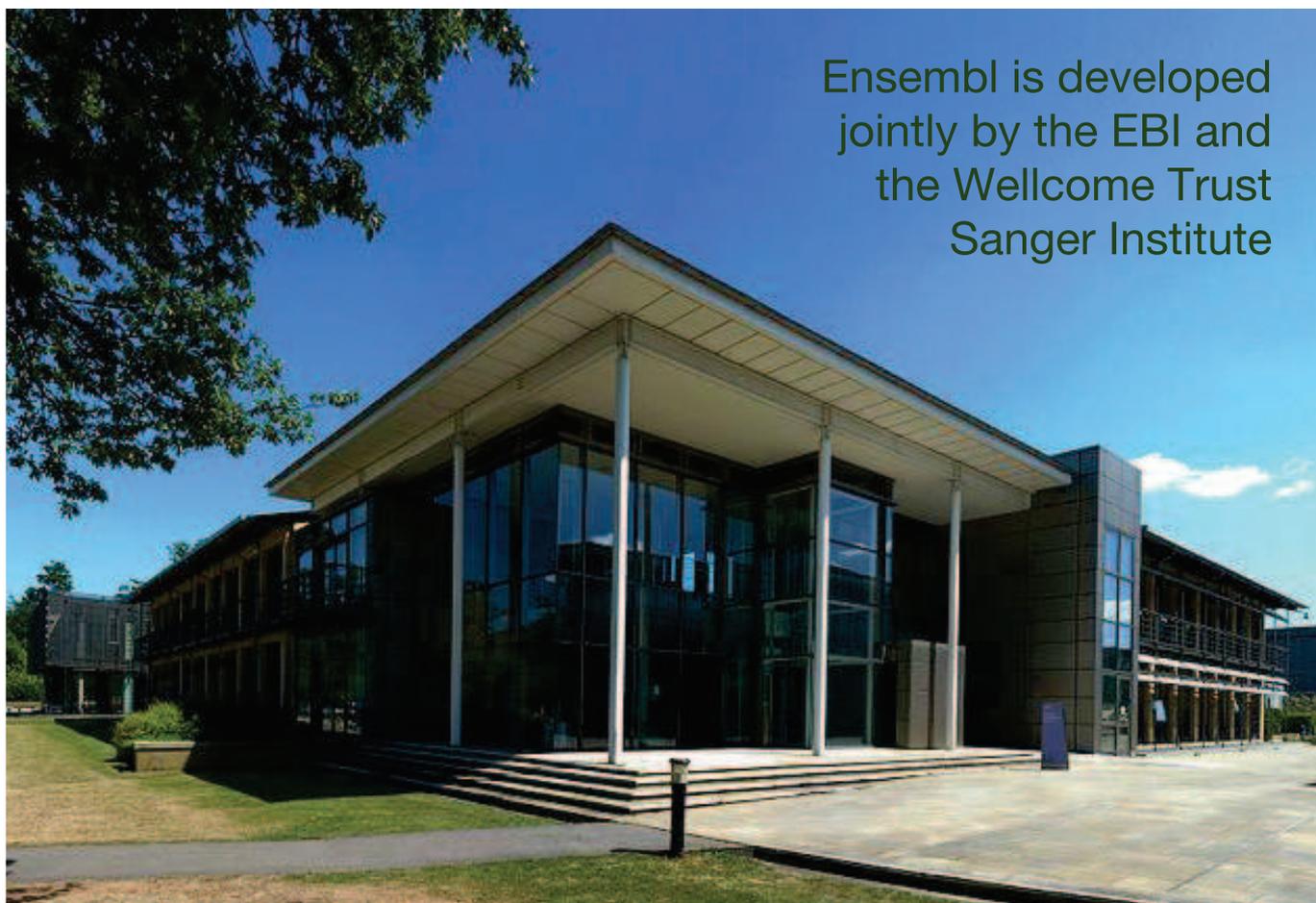
- BioMart



- Tools for processing your own data



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



Ensembl Plants uses Ensembl technology

Ensembl:



- A platform for genome browsing, annotation and analysis developed jointly by the EBI and Wellcome Trust Sanger Institute.
- Has modules for handling:
 - Genomic data, Variations, Comparative genomics, Gene prediction, ...
- Multiple points of access to data:
 - Browser-based application, Perl and REST APIs, direct access (MySQL), BioMart data mining tool, DAS (client and server), FTP...
- 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

Currently 38 genomes in Ensembl Plants

<http://plants.ensembl.org>

	Arabidopsis lyrata GA JGI 81972
	Arabidopsis thaliana VPGA TAIR 3702
	Brassica rapa GA VFCAAS 51351
	Glycine max GA JGI 3847
	Medicago truncatula GA IMGAG 3880
	Populus trichocarpa GA JGI 3694
	Prunus persica GA IPGI 3760
	Solanum lycopersicum PGA ITGSP 4081
	Solanum tuberosum GA PGSC 4113
	Vitis vinifera VPGA Genoscope Vitis vinifera 29760

Brassicales

Fabales

Malpighiales

Rosales

Solanales

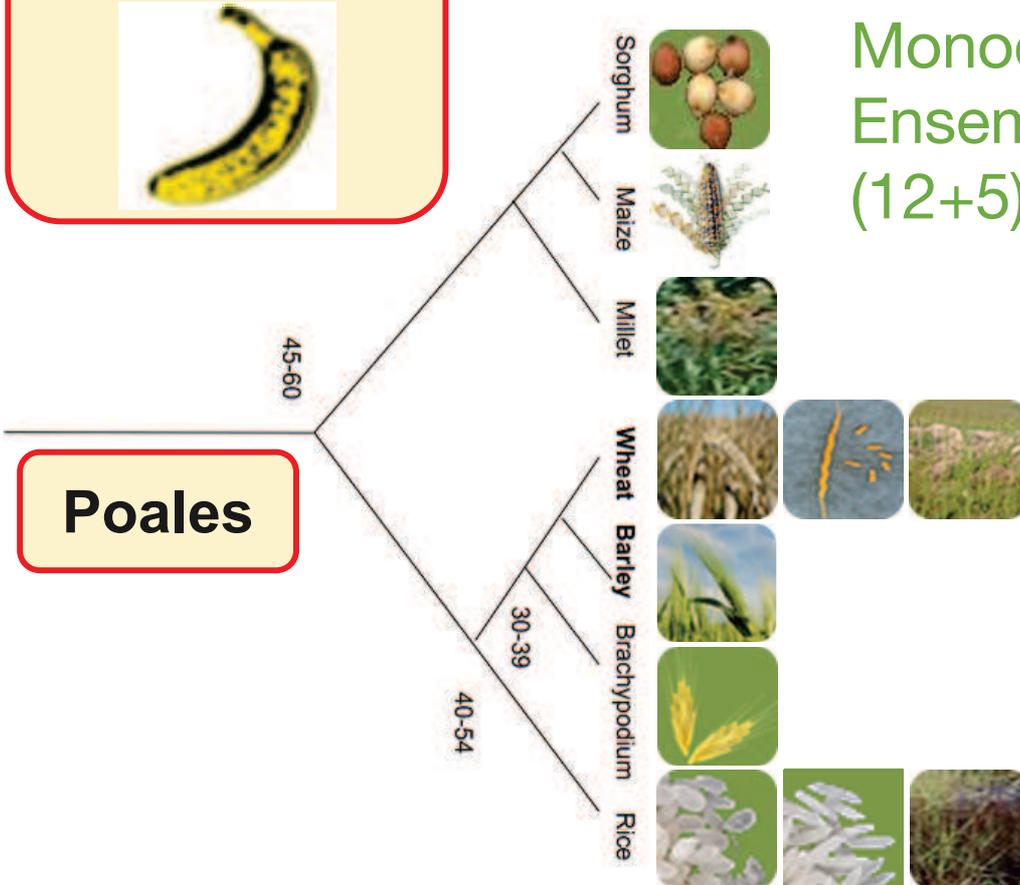
Vitales

Dicots in Ensembl Plants (10)



Zingiberales



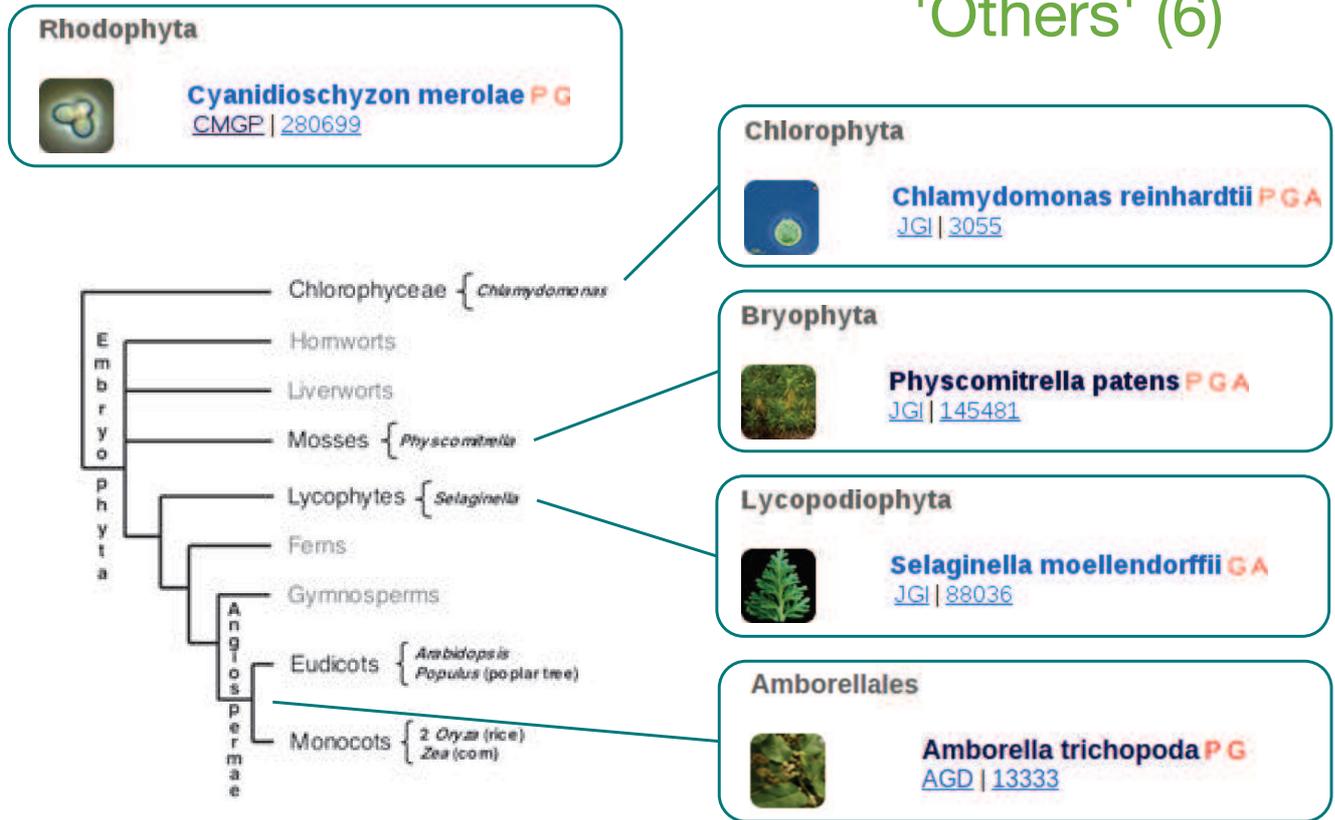


Monocots in Ensembl Plants (12+5)

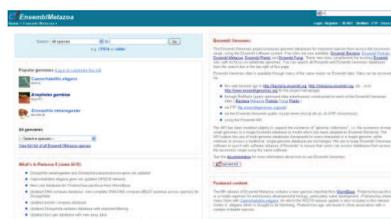
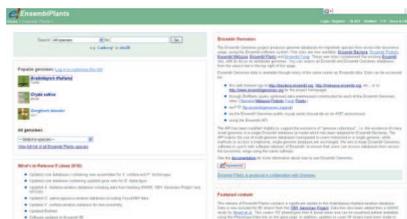
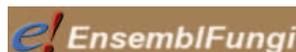
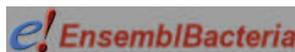
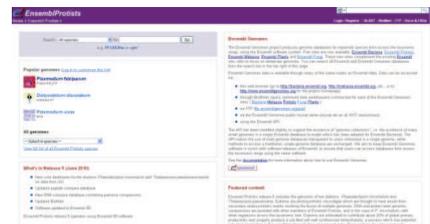
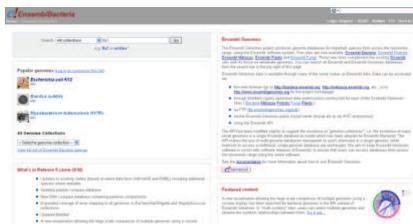
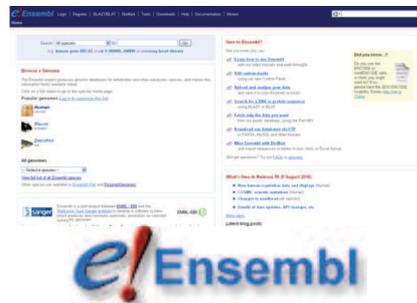
...



'Others' (6)



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



The transPLANT project is funded by the European Commission within its 7th Framework Programme under the thematic area "Infrastructures". Contract number 283496.



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



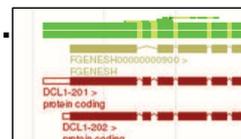
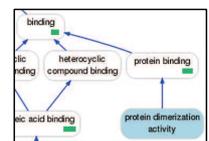
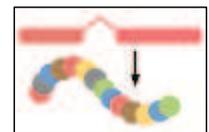
The transPLANT project is funded by the European Commission within its 7th Framework Programme under the thematic area "Infrastructures". Contract number 283496.

www.transplantdb.eu



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



```
ATCGAGCT
ATCCAGCT
ATCGAGAT
```



Walk through 'demo' for Ensembl Plants

(1.5 days in 5 mins)

EMBL-EBI 

 **EnsemblPlants** ▾

[BLAST](#) | [Sequence Search](#) | [BioMart](#) | [Tools](#) | [More ▾](#)

[Login/Register](#)

 Search Ensembl Plants... 

Search: for

e.g. **Carboxy*** or **chx28**

Popular genomes



Arabidopsis thaliana
TAIR10



Oryza sativa Japonica (Rice)
IRGSP-1.0



Triticum aestivum
IWGSP1



Hordeum vulgare
030312v2



Zea mays
AGPV3



Physcomitrella patens
ASM242v1

★ [Log in to customize this list](#)

All genomes

-- Select a species -- ▾

[View full list of all Ensembl Plants species](#)

What's New in Release 22

- New genomes
 - [Amborella trichopoda](#)
 - [Prunus persica](#) (peach)

Did you know...?

The Assembly

Comparing the bread wheat component genomes



Hexaploid bread wheat (*Triticum aestivum*) was formed by two separate hybridization events, bringing together the diploid progenitor genomes into a single species where they have been independently maintained. The A, B and D component genomes have been compared, allowing us to call orthology relationships between them, identifying the so called homoeologous genes. [Click here for example](#). Relationships between the component genomes can also now be browsed in our new region comparison view. [Click here for example](#).

The bread wheat genome in Ensembl Plants is the chromosome survey sequence for *Triticum aestivum* cv. Chinese Spring generated by the [International Wheat Genome Sequencing Consortium](#). The gene models are provided by [MIPS](#) (version 2.1). A total of 99,386 protein coding genes have been predicted.

See also the wheat homepage at 

[Read more](#) about the [assembly](#), [annotation](#) and [analysis](#) of bread wheat provided by [Ensembl Plants...](#)

Ensembl Plants is developed in coordination with other plant genomics and bioinformatics groups via the EBI's role in the [transPLANT](#) consortium. The transPLANT project is funded by the [European Commission](#) within its [7th Framework Programme](#), under the thematic area "Infrastructures", contract number [283496](#).



Wheat genomics resources are developed as part of our involvement in the consortium [Triticeae Genomics For Sustainable Agriculture](#). Barley genomics resources are funded through the [UK barley genome sequencing project](#). Both projects are funded by the BBSRC.



New Search

Search Ensembl Plants

- New Search
- Gene (23)
 - Ensembl Plants (23)
- Gene Tree
 - Ensembl Plants
 - Pan-taxonomic Compara

Configure this page

Manage your data

Export data

Bookmark this page

Share this page

Ensembl Plants is produced in collaboration with Gramene

Search results for 'RPOA'

Showing 1-10 of 12 Genes found in Ensembl Plants (filtered)

Filtered by species: *Triticum aestivum*

Traes_4DL_1638411DC

Description DNA-directed RNA polymerase subunit alpha [Source:UniProtKB/Swiss-Prot;Acc:P12073]
 Gene ID [Traes_4DL_1638411DC](#)
 Species [Triticum aestivum](#)
 Location [IWGSC_CSS_4DL_scaff_14466158:3-1310](#)

Traes_4AS_1638411DC

Description DNA-directed RNA polymerase subunit alpha [Source:UniProtKB/Swiss-Prot;Acc:P12073]
 Gene ID [Traes_4AS_1638411DC](#)
 Species [Triticum aestivum](#)
 Location [IWGSC_CSS_4AS_scaff_5949307:9772-11170](#)

Traes_5AS_6DFE37F47

Description DNA-directed RNA polymerase subunit alpha [Source:UniProtKB/Swiss-Prot;Acc:P12073]
 Gene ID [Traes_5AS_6DFE37F47](#)
 Species [Triticum aestivum](#)
 Location [IWGSC_CSS_5AS_scaff_1552572:38245-39772](#)

Traes_3B_1638411DC

Description DNA-directed RNA polymerase subunit alpha [Source:UniProtKB/Swiss-Prot;Acc:P12073]
 Gene ID [Traes_3B_1638411DC](#)
 Species [Triticum aestivum](#)

Triticum aestivum Location: IWGSC_CSS_2AL_scaff_6420996:5,000-14,997

Location-based displays

- Whole genome
- Chromosome summary
- Region overview
- Region in detail
- Comparative Genomics
 - Alignments (image) (2)
 - Alignments (text) (2)
 - Region Comparison (2)
 - Synteny
- Genetic Variation
 - Resequencing
 - Linkage Data
 - Markers

Configure this page

Manage your data

Export data

Bookmark this page

Share this page

Ensembl Plants is produced in collaboration with Gramene

Scaffold IWGSC_CSS_2AL_scaff_6420996: 1-19,996

Region in detail

Contigs

Protein-coding Gene (...)

Gene Legend

Location: IWGSC_CSS_2AL_scaff_6420996:5000-14997

Triticum UnGene...

T. turgidum CDS, all...

Protein-coding Gene...

Contigs

Protein-coding Gene...

Triticum Genome...

T. turgidum Genome...

All repeats

%GC

Gene Legend

Gene Legend

There are currently 38 tracks turned off.

Ensembl Plants Triticum aestivum version 74.1 (IWGSP1) Scaffold IWGSC_CSS_2AL_scaff_6420996:5,000-14,997

EnsemblPlants BLAST | Sequence Search | BioMart | Tools | Downloads | Help & Documentation Login/Register

Triticum aestivum Location: IWGSC_CSS_5AS_scaff_1552572-39,245-39,772 Gene: RPOA Transcript: RPOA-202

Gene: RPOA Traes_5AS_6DFE37F47

Description: DNA-directed RNA polymerase subunit alpha [Source: UniProtKB/Swiss-Prot P12073](#)

Location: Scaffold IWGSC_CSS_5AS_scaff_1552572_39,245-39,772 reverse strand.

Transcripts: This gene has 1 transcript (splice variant) [Hide transcript table](#)

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype
RPOA-202	Traes_5AS_6DFE37F47.2	1341	Traes_5AS_6DFE37F47.2	355	Protein coding

Summary

Name: RPOA (UniProtKB Gene Name)

Gene type: Protein coding

Prediction Method: Triticum aestivum genes annotated by MIPS

Go to Region in Detail for more tracks and navigation options (e.g. zooming)

Gene Legend: protein coding

Configure this page | Manage your data | Export data | Bookmark this page



EnsemblPlants BLAST | Sequence Search | BioMart | Tools | Downloads | Help | Documentation Login/Register

Triticum aestivum Location: IWGSC_CSS_4DL_scaff_14466158:3-1,310 Gene: Traes_4DL_1638411DC Transcript: RPOA-202

Gene: Traes_4DL_1638411DC

Description: DNA-directed RNA polymerase subunit alpha [Source: UniProtKB/Swiss-Prot P12073](#)

Location: Scaffold IWGSC_CSS_4DL_scaff_14466158:3-1,310 forward strand.

Transcripts: This gene has 1 transcript (splice variant) [Hide transcript table](#)

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype
RPOA-202	Traes_4DL_1638411DC.2	1071	Traes_4DL_1638411DC.2	355	Protein coding

GO: molecular function

Annotated terms | Ancestry chart

Terms:

- Annotated terms
- Generic GO slim terms
- Plant slim terms

Relations:

- is_a
- part_of

```

graph TD
    DNA_directed_RNA_polymerase_activity --> RNA_polymerase_activity
    RNA_polymerase_activity --> nucleotidyltransferase_activity
    nucleotidyltransferase_activity --> transferase_activity_transferring_phosphorus_containing_groups
    transferase_activity_transferring_phosphorus_containing_groups --> transferase_activity
    transferase_activity --> catalytic_activity
    catalytic_activity --> molecular_function
    DNA_binding --> nucleic_acid_binding
    nucleic_acid_binding --> organic_cyclic_compound_binding
    nucleic_acid_binding --> heterocyclic_compound_binding
    organic_cyclic_compound_binding --> binding
    heterocyclic_compound_binding --> binding
    binding --> molecular_function
    protein_dimerization_activity --> protein_binding
    protein_binding --> molecular_function
  
```

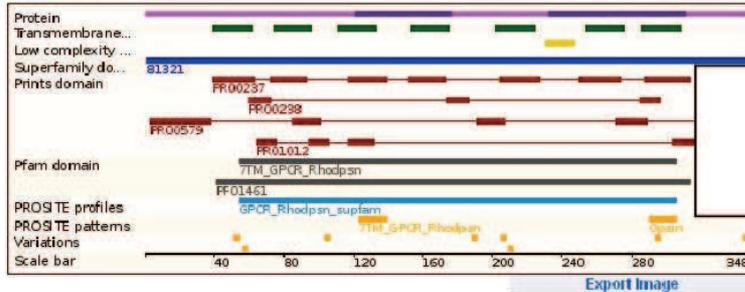
Ensembl Plants is produced in collaboration with Gramene



- Protein
- External References
- General identifiers (32)
- Oligo probes (28)
- Gene ontology (22)
- Genetic Variation
- Population comparison
- Comparison image
- Protein information
- Protein summary**
- Domains & features (70)
- Variations (8)
- External Data
- Personal annotation
- History
- Transcript history
- Protein history

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CCDS
RHO-201	ENS00000296271	2758	ENSP00000296271	348	Protein coding	CCDS3063

« Comparison image Protein summary help Domains & features »



Ensembl protein

Signatures mapped to the sequence



The transPLANT project is funded by the European Commission within its 7th Framework Programme under the thematic area "Infrastructures". Contract number 283496.

www.transplantdb.eu



e!EnsemblPlants
Home > Arabidopsis thaliana
Location: 3:19,431,371-19,434,403 | Gene: PAD4 | Transcript: AT3G52430.1

Gene-based displays

- Gene summary
- Splice variants (1)
- Supporting evidence
- Sequence
- External references (4)
- Regulation
- Plants Compara
 - Genomic alignments (7)
 - Gene Tree (image)
 - Gene Tree (text)
 - Gene Tree (alignment)
 - Orthologues (11)
 - Paralogues (3)
- Pan-taxonomic Compara
 - Gene Tree (image)
 - Gene Tree (text)
 - Gene Tree (alignment)
 - Orthologues (2)
 - Paralogues (3)
 - Protein families (0)
- Genetic Variation
 - Variation Table**
 - Variation Image
- External Data
- Personal annotation
- ID History
- Gene history

Gene: PAD4 (AT3G52430-TAIR-G)

PAD4 (PHYTOALEXIN DEFICIENT 4); lipase/ protein binding / triacylglycerol lipase; Encodes a lipase-like gene that is important for salicylic acid signaling and function in resistance (R) gene-mediated and basal plant disease resistance. PAD4 can interact directly with EDS1, another disease resistance signaling protein. Expressed at elevated level in response to green peach aphid (GPA) feeding, and modulates the GPA feeding-induced leaf senescence through a mechanism that doesn't require camalexin synthesis and salicylic acid (SA) signaling. Source: TAIR PAD4

Location Chromosome 3: 19,431,371-19,434,403 forward strand.

Transcripts There is 1 transcript in this gene: [show transcripts](#)

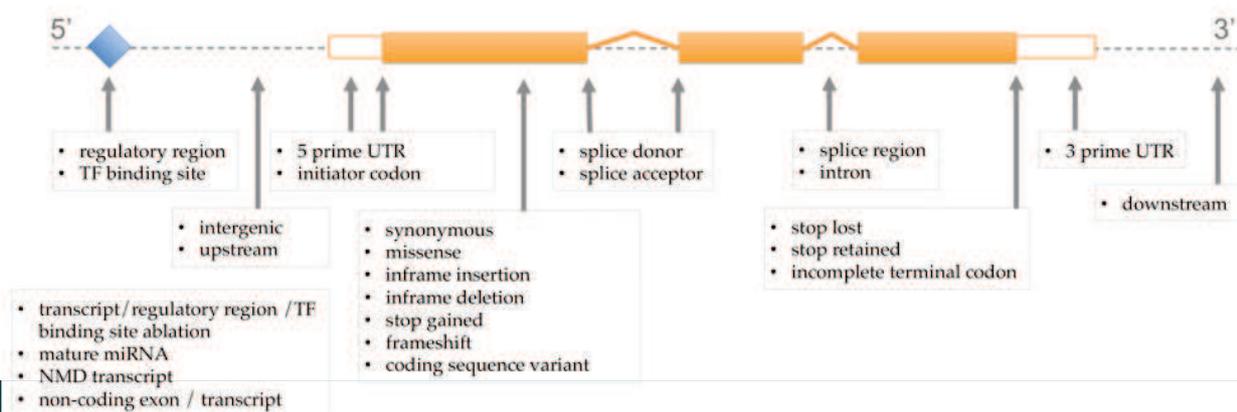
« Variation Table Variation Image help External Data »

Variation legend

- 5' UTR
- Intronic
- Non-synonymous coding
- 3' UTR
- Synonymous coding
- Upstream

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

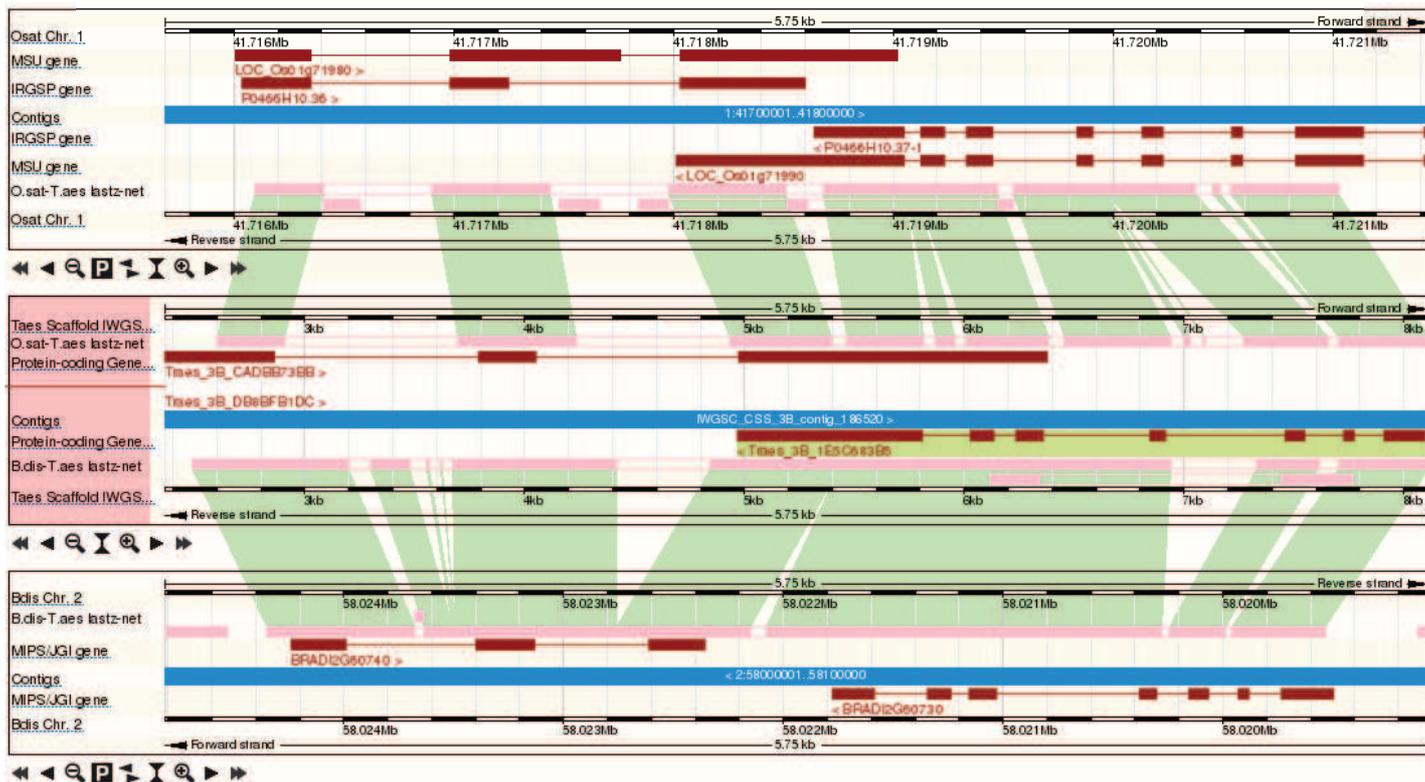


EMBL-EBI

The screenshot shows the Ensembl Plants interface for the gene **KEA1 (AT1G01790-TAIR-G)**. The main content includes:

- Gene summary**: KEA1 (K EFFLUX ANTIPORTER 1); potassium ion transmembrane transporter/ potassium:hydrogen antiporter, K efflux antiporter KEA1. Source: TAIR KEA1.
- Location**: Chromosome 1: 284,781-291,094 forward strand.
- Transcripts**: There is 1 transcript in this gene: [show transcripts](#).
- Gene Tree (image)**: A phylogenetic tree showing relationships between KEA1 orthologs and paralogs across various species, including Arabidopsis thaliana, Oryza sativa, Vitis vinifera, and E. coli. The tree highlights duplication nodes and speciation nodes.
- Orthologues**: A list of orthologous genes from other species, each with a green bar representing the gene structure.
- Legend**:
 - x1 branch length
 - - - x10 branch length
 - - - x100 branch length
 - Gene ID current gene
 - Gene ID within-sp. paralog
 - speciation node
 - duplication node
 - ambiguous node
 - collapsed sub-tree
 - collapsed (current gene)
 - collapsed (paralog)
 - AA alignment match/mismatch
 - AA consensus > 66% (mis)match
 - AA consensus > 33% (mis)match
 - AA alignment gap
- View options**:
 - View current gene only
 - View paralogs of current gene
 - View all duplication nodes

WGA between wheat, rice and brachy



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!

Funding (Ensembl Plants)

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

10/01/2014

pkersey@ebi.ac.uk

EMBL-EBI 

People (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 our HelpDesk:
 - helpdesk@ensemblgenomes.org

10/01/2014

pkersey@ebi.ac.uk

EMBL-EBI 

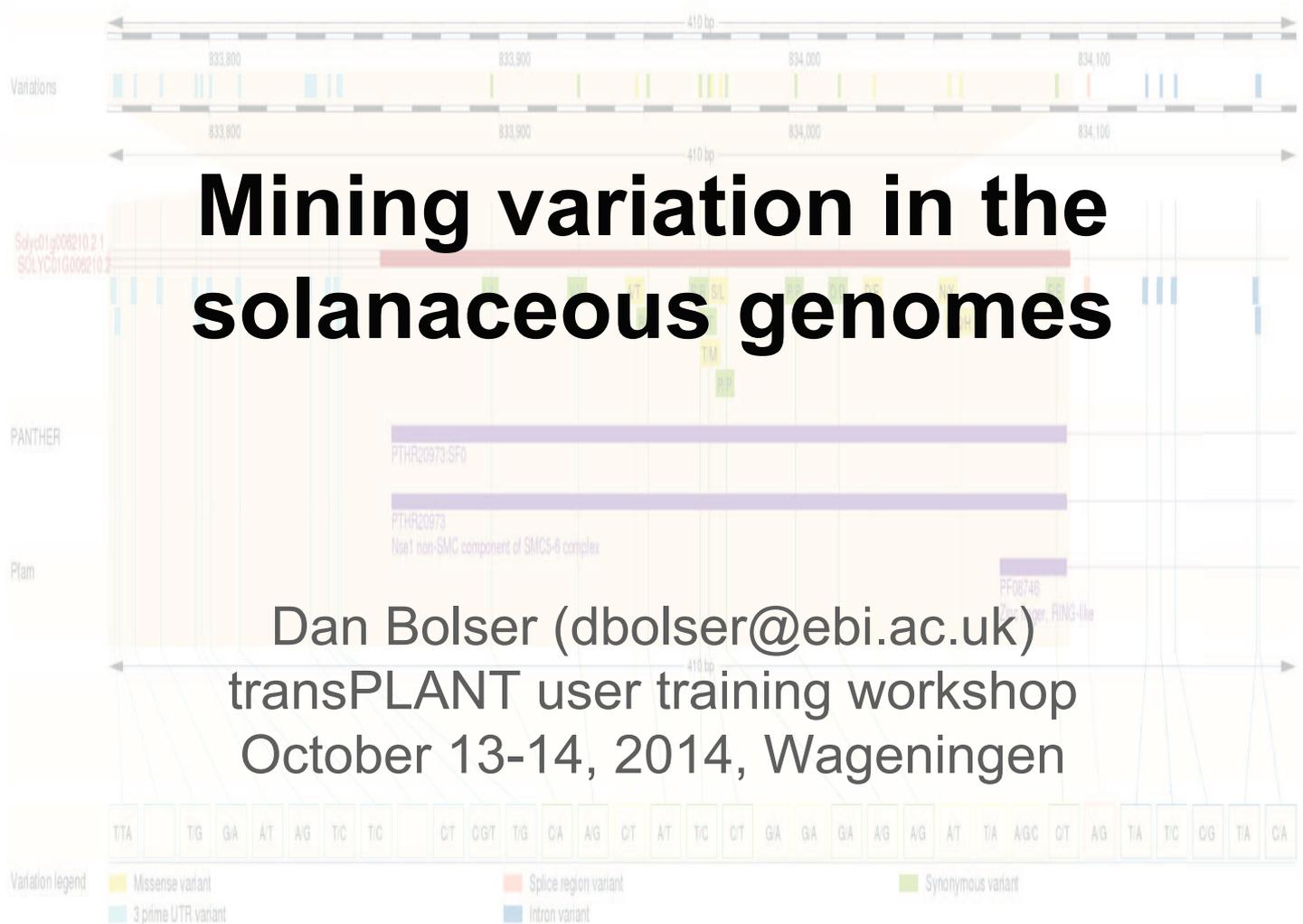
Workshop!



The transPLANT project is funded by the European Commission within its 7th Framework Programme under the thematic area "Infrastructures". Contract number 283496.

www.transplantdb.eu





Overview

1. Quick(!) introduction to Ensembl Plants
 - <http://goo.gl/K3u35V>
2. Hands-on exercise: Exploring variation in Ensembl Plants
 - Slides below
3. The transPLANT variation archive
 - wrap-up

Workshop: Exploring variation in Ensembl Plants

150 Tomato Genome
ReSequencing project

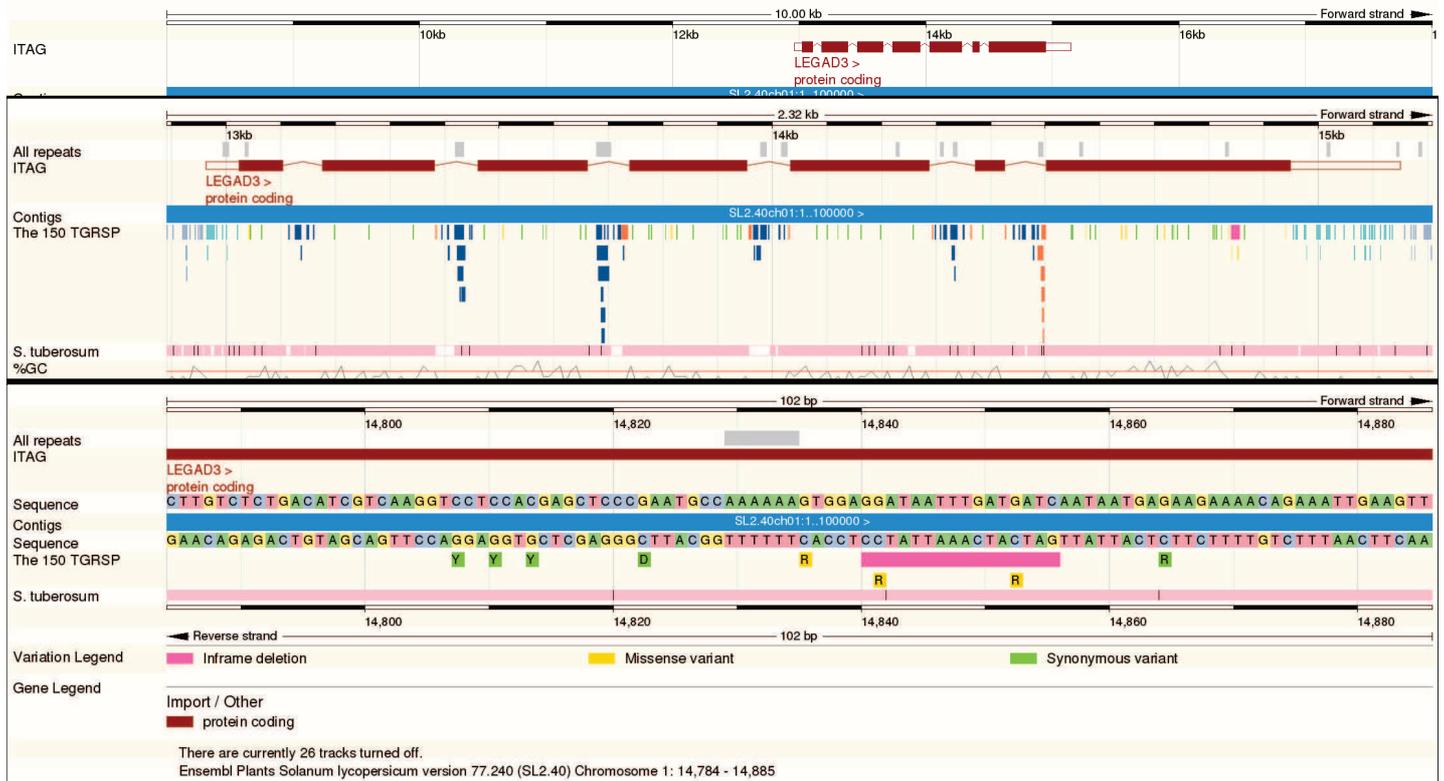
1) Looking at tomato variation data

- Go to <http://test.plants.ensembl.org>
 - Can you find your way to the **tomato** homepage?
- Search for LCY1 (Lycopene beta cyclase) 
- Find variation information over and around the location of this gene...
- Which variation (IDs) have likely functional consequence for this gene?



Can you find primers for this allele?

1) Looking at tomato variation data



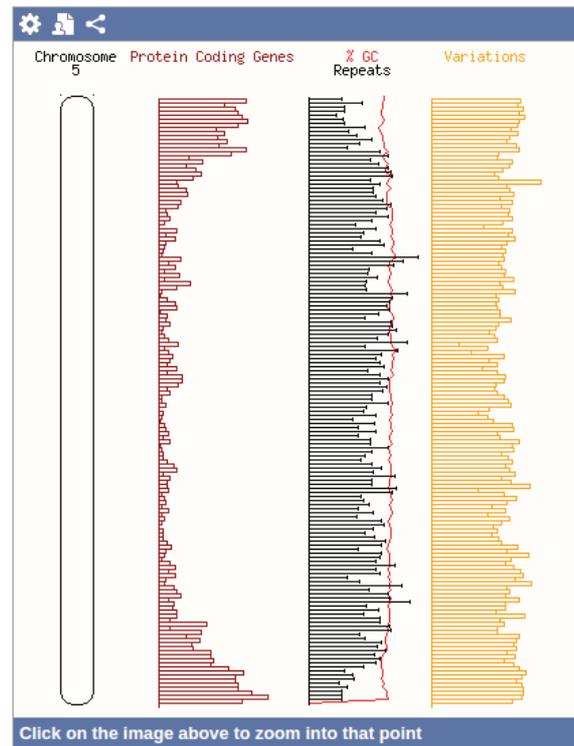
2) Individual genotypes

- Go to <http://plants.ensembl.org>
 - Can you find your way to the *arabidopsis thaliana* homepage?
- Find your way to an (example?) transcript.
- How many individuals have heterozygous variations that fall within this transcript?
 - HINT: Use the “Comparison image”
- What configuration is available for this view?
What would you like to see in tomato?

3) Chromosome summary

- From the location view, what does the “Chromosome summary” look like in arabidopsis?
- How does this differ in tomato?

BONUS Why?



4) Using BioMart

- Can you find your way to ‘BioMart’?
- Filter for some ‘interesting variations’ in a ‘region of interest’...

BONUS Export data and gather some ‘interesting’ statistics.

5) Attaching a VCF

- Locate some VCF to attach to the Ensembl browser
 - Tip, try the ENA's 'Advanced search' for analysis data types.
- 'Attach' the file to the browser
 - Tip, look for 'add your data'
 - Tip, select 'VCF (indexed)'
 - http://www.ebi.ac.uk/~dbolser/RF_001.vcf.gz



Email the help desk to ask why this feature isn't working!

Wrap-up (Aims)

1. Introduce BioLinux, an important tool.
2. Demonstrate the utility of Ensembl Plants (with an emphasis on variation information).
3. Make you aware of the transPLANT variation archive.
4. To be “not a total waste of time”...
5. ~~Provide an idea of the basics of SNP calling by resequencing.~~
6. Give you some reference materials for future use.

RESOURCES

- A SNP calling pipeline overview
 - Call your own SNPs using BioLinux!
 - <http://goo.gl/7C3t5n>
- A SNP calling pipeline tutorial
 - Command by command overview.
 - <http://goo.gl/qolr1g>
- Ensembl exercises
 - Often human centric, but relevant to all species.
 - <http://www.ebi.ac.uk/seqdb/confluence/display/ENSEXT/Ensembl+exercises>
- The transPLANT variation archive
 - <http://www.transplantdb.eu/variation/submit>



RESOURCES

- Ensembl help
 - <http://www.ensembl.org/info/index.html>
- Ensembl Genomes outreach
 - <http://ensemblgenomes.org/info/about/outreach>
- Asking for help
 - <http://www.ensembl.org/Help/Contact>

Ensembl Genomes 2013: scaling up access to genome-wide data

Paul Julian Kersey^{1,*}, James E. Allen¹, Mikkel Christensen¹, Paul Davis¹, Lee J. Falin¹, Christoph Grabmueller¹, Daniel Seth Toney Hughes¹, Jay Humphrey¹, Arnaud Kerhornou¹, Julia Khobova¹, Nicholas Langridge¹, Mark D. McDowall¹, Uma Maheswari¹, Gareth Maslen¹, Michael Nuhn¹, Chuang Kee Ong¹, Michael Paulini¹, Helder Pedro¹, Iliana Toneva¹, Mary Ann Tuli², Brandon Walts¹, Gareth Williams¹, Derek Wilson¹, Ken Youens-Clark³, Marcela K. Monaco³, Joshua Stein³, Xuehong Wei³, Doreen Ware^{3,4}, Daniel M. Bolser¹, Kevin Lee Howe¹, Eugene Kulesha¹, Daniel Lawson¹ and Daniel Michael Staines¹

¹The European Molecular Biology Laboratory, The European Bioinformatics Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK, ²Wellcome Trust Sanger Centre, The Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SA, UK, ³Cold Spring Harbor Laboratory, 1 Bungtown Rd, Cold Spring Harbor, NY 11724, USA and ⁴USDA-ARS, Cornell University, Ithaca, NY, 14853, USA

Received September 13, 2013; Accepted October 1, 2013

ABSTRACT

Ensembl Genomes (<http://www.ensemblgenomes.org>) is an integrating resource for genome-scale data from non-vertebrate species. The project exploits and extends technologies for genome annotation, analysis and dissemination, developed in the context of the vertebrate-focused Ensembl project, and provides a complementary set of resources for non-vertebrate species through a consistent set of programmatic and interactive interfaces. These provide access to data including reference sequence, gene models, transcriptional data, polymorphisms and comparative analysis. This article provides an update to the previous publications about the resource, with a focus on recent developments. These include the addition of important new genomes (and related data sets) including crop plants, vectors of human disease and eukaryotic pathogens. In addition, the resource has scaled up its representation of bacterial genomes, and now includes the genomes of over 9000 bacteria. Specific extensions to the web and programmatic interfaces have been developed to support users in navigating these large data sets. Looking forward, analytic tools to allow targeted selection of data for visualization and download are likely to

become increasingly important in future as the number of available genomes increases within all domains of life, and some of the challenges faced in representing bacterial data are likely to become commonplace for eukaryotes in future.

OVERVIEW AND ACCESS

Ensembl Genomes (<http://www.ensemblgenomes.org>) is organized as five sites, each focused on one of the traditional kingdoms of life: bacteria (specific URL <http://bacteria.ensembl.org>), protists, fungi, plants and (invertebrate) metazoa. Vertebrate metazoa are the focus of the Ensembl project (<http://www.ensembl.org>) (1); Ensembl Genomes provides a complementary set of interfaces for non-vertebrate species. Core data available for all species include genome sequence and annotations of protein-coding and non-coding genes; additional data include transcriptional data, polymorphisms and comparative analysis. Interactive access is provided through a web interface providing genome browsing capabilities: users can scroll through a graphical representation of a DNA molecule at various levels of resolution, seeing the relative locations of features—including conceptual annotations [e.g. genes, single nucleotide polymorphism (SNP) loci], sequence patterns (e.g. repeats) and experimental data (e.g. sequences and external sequence features mapped onto the genome)—supporting the primary

*To whom correspondence should be addressed. Tel: +44 1223 494601; Fax: +44 1223 494468; Email: pkersey@ebi.ac.uk

© The Author(s) 2013. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

annotations. Functional information is provided through direct curation, import from the UniProt Knowledgebase (2) or imputation from protein sequence [using the classification tool InterProScan (3)]. Users can download much of the data available on each page in a variety of formats, and tools exist for upload of (various types of) user data, allowing users to see their own annotation in the context of the reference sequence. DNA- and protein-based sequence search are also available.

The data are stored in a set of MySQL databases using the same schemas as those in use for the Ensembl project. Direct access to these is provided through a public MySQL server (mysql.ebi.ac.uk:4157; user 'anonymous') and additionally through well-developed Application Programming Interfaces (APIs) that provide an object-oriented framework for working with the data. Database dumps and common data sets (e.g. DNA, RNA and protein sequence sets and sequence alignments) can be directly downloaded in bulk via file transfer protocol (<ftp://ftp.ensemblgenomes.org>).

Ensembl Genomes data are also made available through a series of data warehouses, optimized around common (gene and SNP-centric) queries, using the BioMart data warehousing system (4). The BioMart framework provides a series of interfaces, including web-based query building tools, for each of the Ensembl Genomes (eukaryotic) domains (e.g. at <http://plants.ensembl.org/biomart/martview>) and a variety of other interfaces for interactive and programmatic access. BioMarts are not currently available for Ensembl Bacteria.

Ensembl Genomes is released 4–5 times a year, in synchrony with releases of Ensembl, using the same software as the corresponding Ensembl release. The overall suite of Ensembl Genomes interfaces mirrors the interfaces provided for vertebrate genomes in Ensembl, and allows users access to genomic data from across the tree of life in a consistent manner.

A COLLABORATIVE MODEL FOR GENOME-SCALE DATA

The Ensembl Genomes project is driven by a number of domain-specific collaborations, each with a scientific community with its own focus of interest. By working in partnership with us, communities can benefit from a robust infrastructure and the integration of their data within a comprehensive service. These collaborations take a number of forms. In some domains, we work with our partners to develop a community-centric service, aimed at each community's specific needs, but also mirror key data within the central Ensembl Genomes portal. Examples of such collaborations include VectorBase (<http://www.vectorbase.org>) (5), a resource for the genomes of invertebrate pathogens of human diseases; WormBase (<http://www.wormbase.org>) (6), which maintains resources for nematode genomes, especially the model species *Caenorhabditis elegans*; PomBase (<http://www.pombase.org>) (7), the model organism database for the fission yeast *Schizosaccharomyces pombe*; and PhytoPath (<http://www.phytopathdb.org>), a resource for plant pathogens,

with a focus on fungi and oomycetes. In other domains, we collaborate more broadly with other integrative centers, with a goal of developing high-quality networks of inter-linked resources through the sharing of common reference data and standards for interoperability. In the context of Ensembl Plants, for example, we work closely with the Gramene database (<http://www.gramene.org>) (8) and a number of leading European plant genomics and informatics centers through the transPLANT project (<http://www.transplantdb.eu>). In addition, we contribute to many community-driven projects to sequence, assemble and annotate particular genomes, and make the resulting data available through the Ensembl Genomes site.

Ensembl Genomes prioritizes data for incorporation, according to scientific importance. The criteria for priority treatment are first, data relevant to our specific collaborations; second, data from other major experimental species; and third, data from other species that provide local or remote evolutionary context for the priority species, and which are used to strengthen the comparative analysis provided in the site. For the first category of genomes, we actively work with our collaborators to produce the primary community-recognized annotation. For the second category, we supplement the reference annotation (often maintained by model organism databases or other similar resources) with additional high-value data sets. For several species in these two categories, we have constructed variation databases, which store genotypes, loci and phenotypes from large-scale genome-wide array-based and resequencing studies, and have made the data available through specialized graphical views and an SNP-centric BioMart. Variation data are sourced from dbSNP (9) or Database of Genomic Variants archive (10) where available, or otherwise directly from the data producers. For the third category of genomes, annotation is generally incorporated from the original submitters with only limited enhancement (for example, the annotation of non-coding genes, if absent in the original submission).

At the time of writing, there have been 10 releases of Ensembl Genomes since the previous report was published in this journal (11). The current release is release 20, made public in September 2013. In this time, there has been a significant increase in the content of all five Ensembl Genomes sites.

Metazoa

Nineteen new genomes have been added, including the sponge *Amphimedon queenslandica*, the south and central American malarial mosquito *Anopheles darlingi*, the leaf-cutter ant *Atta cephalotes*, the silkworm *Bombyx mori*, the water flea *Daphnia pulex*, the pacific oyster *Crassostrea gigas*, the owl limpet *Lottia gigantea*, the scuttle fly *Megaselia scalaris*, the centipede *Strigamia maritima*, the kissing bug *Rhodnius prolixus*, the red flour beetle *Tribolium castaneum*, the two-spotted spider mite *Tetranychus urticae*, two annelid worms, two butterflies and three nematodes. Additional variation data (12,13) have been introduced for *Anopheles gambiae*, and new DNA-based comparative analysis has been added for nematodes.

Plants

Twenty-two new genomes have been added, including Chinese cabbage (*Brassica rapa*); soy bean (*Glycine max*), barley (*Hordeum vulgare*), banana (*Musa acuminata*), barrel clover (*Medicago trunculata*), the club moss *Selaginella moellendorffii*, foxtail millet (*Setaria italica*), tomato (*Solanum lycopersicum*), potato (*Solanum tuberosum*), two species of rice, two diploid ancestors of hexaploid bread wheat and, as taxonomic outliers, two algal species. In addition, the preliminary genome assemblies, homeologous SNP calls and expressed sequence transcript (EST) sequences available for bread wheat have been aligned to the genomes of barley and *Brachypodium distachyon*, and a sequence search has been implemented against the EST sequences that visualized the results in the context of their alignments to these references. Variation databases have been provided for barley (14), maize (15), rice (16,17) and sorghum (18). The variation data set for *Arabidopsis thaliana* has been expanded to include additional data from the 1001 Genomes Project (19) and other work, including phenotypic data (20). Additional comparative alignments have been produced for cereal genomes.

Fungi

Twenty-four new genomes have been added, including 20 plant pathogens (*Blumeria graminis*, *Botrytis cornerea*, *Fusarium oxysporum*, *Gaeumannomyces graminis*, *Glomerella graminicola*, *Leptosphaeria maculans*, *Melampsora larcini-populina*, *Microbotryum violaceum*, *Nectria haematococca*, *Puccinia triticina*, *Sclerotinia sclerotiorum*, *Sporisorium reilianum*, *Trichoderma reesei*, *Ustilago maydis*, two species of *Gibberella*, two species of *Magnaporthe* and two species of *Pyrenophora*). Other species added include the human pathogen *Cryptococcus neoformans*, the truffle *Tuber melanosporum* and two additional yeast species. RNA-seq alignments (to the genome) have been added for *P. triticina*; EST alignments have been added for *Phaeosphaeria nodorum*, *S. pombe*, *T. melanosporum* and *Zymoseptoria tritici*; and new comparative genomic alignments have been added for certain *Pyrenophora* and yeast species. For phytopathogenic fungal (and protist) species, information about genes impacting on pathogenesis has been imported from the PHI-base database (21), and mutant and overexpression phenotypes are now represented in a color-coded form in the genome browser.

Protists

Eleven new genomes, including those of several important plant and human pathogens, have been added: *Albugo laibachii*, *Entamoeba histolytica*, *Giardia lamblia*, *Guillardia theta*, *Hyaloperonospora arabidopsis*, *Leishmania major*, *Paramecium tetraurelia*, *Pythium ultimum*, *Tetrahymena thermophila*, *Toxoplasma gondii* and *Trypanosoma brucei*. New DNA alignments have been provided for the ciliates, the Peronosporales and the Trypanosomatidae. A variation database has been added for *Phytophthora infestans*.

Bacteria

Ensembl Bacteria has been comprehensively expanded since release 17. Although previously the bacterial division of Ensembl had focused on a small number of selected clades, the division now contains all bacterial genomes that have been completely sequenced, annotated and submitted to the International Nucleotide Sequence databases (European Nucleotide Archive, GenBank and the DNA Database of Japan) (22), a total of 9089 genomes in the latest release. Additional information is incorporated from the UniProtKB, InterPro, information about operons from RegulonDB (23) and about reaction catalysts from Microme (<http://www.micromedb.eu>). To ensure that data within this expanded set remain discoverable, two new species selection mechanisms have been introduced into the portal, one using autocomplete and the other providing a taxonomically structured interface (illustrated in Figure 1). The latter also enables the restriction of (sequence and text) search to user-defined taxonomic segments. Additionally, the Ensembl Perl API has been extended with a new lookup module, allowing users to discover genomes matching their specifications (e.g. full or partial name-match, taxonomic identifier, nucleotide sequence accession) programmatically. Within the browser, an improved representation of transcripts and translations, capable of providing a correct representation of bacterial features (i.e. polycistronic transcripts and alternative translational initiation) has been introduced.

IMPROVED TOOLS FOR DATA ACCESS

A number of improvements to the Ensembl infrastructure have been made during the past year, including the introduction of a scrollable browser and a new RESTful API (a language-agnostic supplement to the existing Perl API), whereas the range of data formats provided (for appropriate data types) via file transfer protocol has been expanded to include Genome File Format and Variant Call Format. A new fast sequence search, based on a back-end provided by the European Nucleotide Archive, has been introduced for all species alongside a Basic Local Alignment Search Tool (BLAST) server. A feature allowing portions of gene trees to be highlighted based on the existence of common annotation has been introduced. Support has been introduced for annotations comprising structured assemblies of ontology terms (e.g. for complex phenotypic description), and a new browser has been implemented for ontological terms, which depicts the ancestry of annotated terms and provides links through to BioMart to allow users to retrieve gene sets annotated with any term in the display. Finally, automatic display of remote files is now supported for any data file using any known synonym to identify the reference sequence on which the data is to be visualized.

COMPARATIVE ANALYSIS

Extensive comparative analyses are performed between the sequences in Ensembl Genomes. Analyses include pairwise alignments between DNA sequences, using the

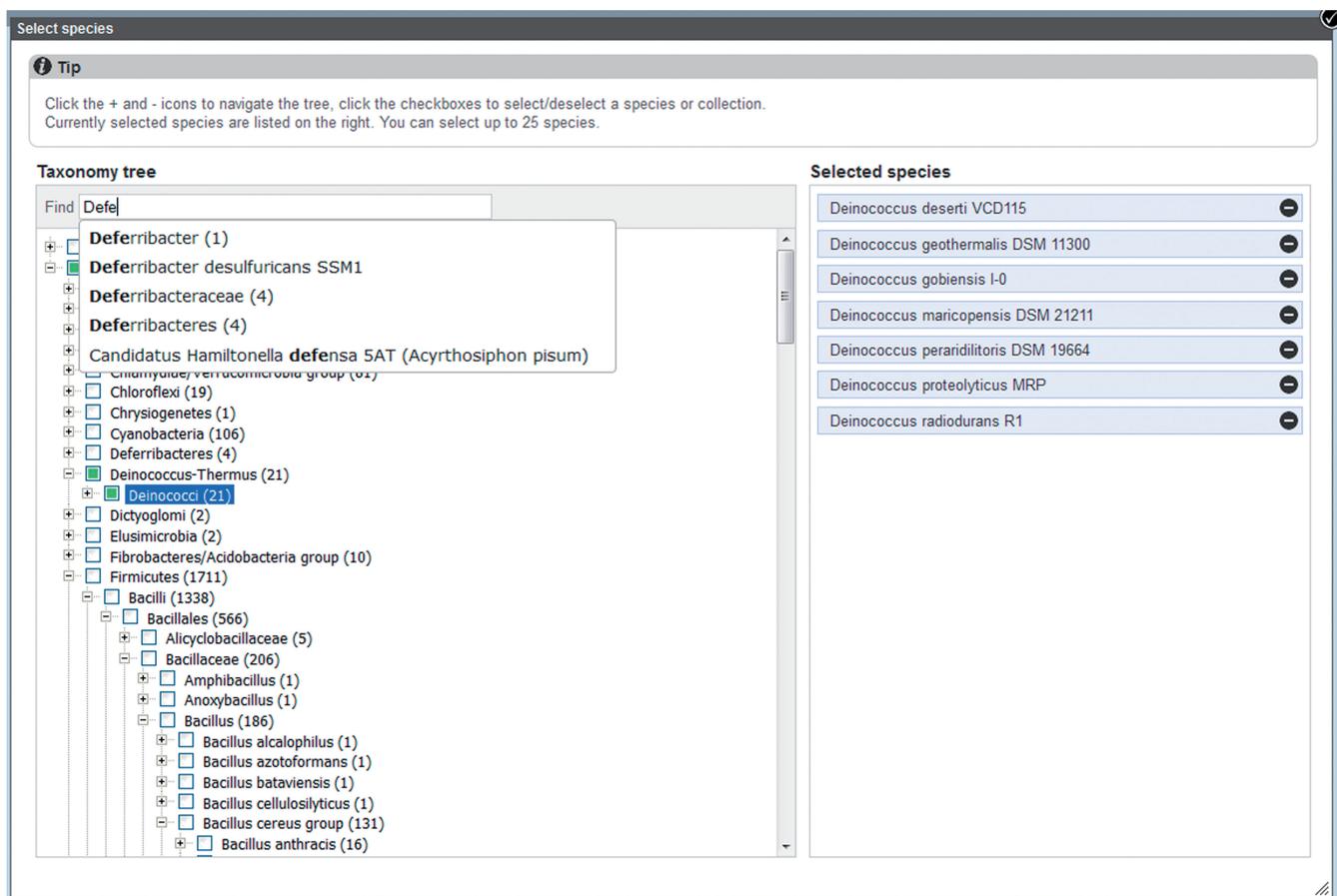


Figure 1. Species selection in Ensembl Bacteria. The figure shows the selection of a basket of genomes for use in a BLAST search. A tree-based navigation system allows the selection of defined portions of the taxonomy for use as library sequences. An autocomplete feature assists the location of particular genomes within the tree.

tools LASTZ (24) and (for more diverged genomes) translated BLAT (25) combined with the use of the chain/net algorithm of Kent et. al (26). The number of these comparisons has increased and we now have 118 pairwise alignments. In Ensembl Plants, pairwise alignments are provided for rice against every other genome, *A. thaliana* against every other genome (except barley) and 14 other pairwise comparisons. In Ensembl Metazoa, comparisons are provided from *Drosophila melanogaster* to 11 other drosophilid species and 4 mosquitoes, for all pairwise combinations of *A. gambiae*, *A. darlingi*, *Aedes aegyptii* and *Culex quinquefasciatus*, from *C. elegans* to 8 other nematodes and from *Brugia malayi* to *Loa loa*. In Ensembl Fungi, all-against-all alignments are available in the *Aspergillus*, *Hypocreales*, *Pucciniales*, *Pyrenophora* clades and for *Saccharomyces cerevisiae* against *Ashyba gossypii*. In Ensembl Protists, DNA alignments are provided for each of three *Phytophthora* species against each other. No DNA-based comparisons are currently provided for bacterial species.

Protein alignments are used to reconstruct evolutionary trees for related genes using the Ensembl Compara Gene Trees pipeline (27). These are run for each eukaryotic domain and additionally for a representative selection of species from across the taxonomic space to identify widely

conserved families and deep homologies between different evolutionary branches. In the current release, the pan-taxonomic database was constructed with the genomes of 12 chordates (11 vertebrates, plus *Ciona intestinalis*), 15 non-chordate metazoans, 7 plants, 7 fungi, 8 protists, 98 bacteria and 25 archaea. Genomes are chosen for inclusion according to a variety of criteria, including mutual taxonomic distance, number of recorded publications, prior inclusion in previous editions of the pan-taxonomic Compara and overlap with the reference proteome sets defined by the UniProt KB. In total, 79 005 gene trees have been constructed for a total of 1 070 325 proteins. Their distribution among the different taxonomic domains is shown in Figure 2. Bacterial proteins (from all included genomes) have additionally been grouped into families using the HAMAP (28) and Panther (29) resources.

CEREALS: SERIOUSLY BIG GENOMES

The genomes of several economically important crop species have not yet been completely sequenced owing to their large size and highly repetitive DNA. However, during the last year, early versions of the diploid barley genome (5 Gb) and the hexaploid bread wheat genome

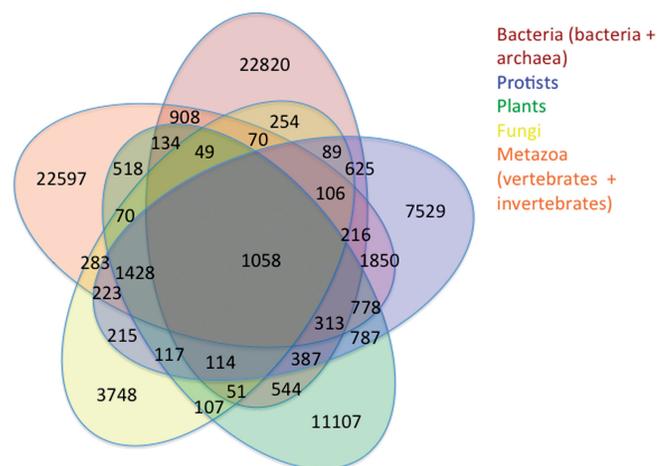


Figure 2. Taxonomic distribution of gene families in the pan-taxonomic comparative analysis in release 19 of Ensembl Genomes. Large numbers of families [defined by clustering according to the Ensembl Gene Trees algorithm (27)] are found only in one domain of life. However, families can be found spanning all combination of domains. The most overrepresented spans (compared with expectations based on the same proportion of families being covering each domain, but assuming the co-coverage of two domains is random) are (i) all five domains and (ii) all four non-bacterial domains; the most underrepresented spans are (i) bacteria and metazoa and (ii) bacteria, metazoa and fungi. For each family of related proteins, a gene tree is constructed and made available for visualization and download, estimating the evolutionary history of that family.

(16Gb) have become available. Neither genome is yet available in a completely assembled form. The current barley genome assembly (14) consists of around 1.9Gb of DNA in 612 267 contigs of over 200 bp, of which ~ 400 Mb of which have been located on chromosome level using markers from extant physical and genetic maps. A total of 24 211 high-confidence protein-coding genes have been called, of which 64% are in anchored locations. The N50 is only 1405 bp, but the N50 of gene-containing scaffolds is much higher (8.4 Kb). Despite the fragmented nature of the genome, barley is represented conventionally in Ensembl Plants, with data shown at all levels from the karyotype through to comparative analysis and variation. In the absence of high-level scaffolding, approximate colocation of contigs to marker sequences located on the physical map is used to provide an approximation of the order and orientation of contigs at each chromosomal locus. Additionally, unanchored contigs have been grouped together in a synthetic 'chromosome' (consisting of the actual contigs with arbitrary gaps between them) to better fit the data model (and critically, to improve analysis times), and contigs of <200 nucleotides have been excluded from the database. In all other respects, the genome can be accessed in the same way as any of the better-assembled genomes in the resource. A typical view of the barley genome in Ensembl Plants is depicted in Figure 3.

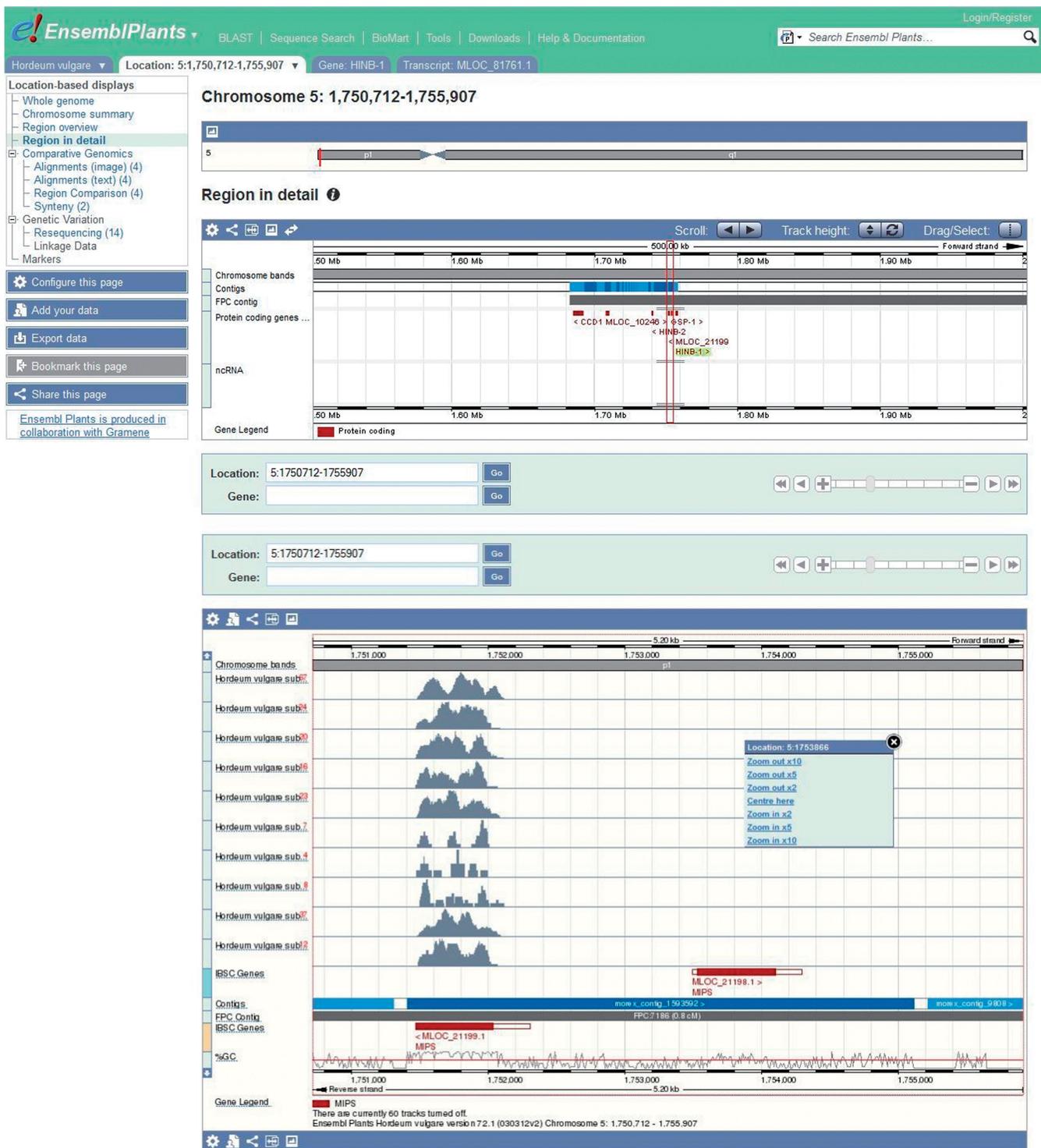
The wheat genome assemblies published in late 2012 (30) are even more fragmentary with many contigs, even those in genic regions, less extensive than the genes themselves; thus, creation of accurate complete gene models is difficult. Therefore, in Ensembl Plants, we have

presented the data in the form of alignments of the genomic contigs onto two better-assembled reference genomes, *Brachypodium* and barley. This enables the wheat sequences, and the homeologous variants (between the three wheat genomes) that have been identified, to be located in the context of full-length gene models predicted in these closely related species. Additionally, a set of 1.3 million wheat ESTs has been mapped onto the *Brachypodium* and barley genomes; and a sequence search facility provided against the wheat EST set that returns an alignment of the query sequence against the top-matching ESTs and additionally shows where those ESTs align to the genomes of these two related species.

PERSPECTIVE AND PRIORITIES

Over the next 2 years, we anticipate that increasingly complete versions of the hexaploid bread wheat genome will emerge from the efforts of the International Wheat Genome Sequencing Consortium (<http://www.wheatgenome.org/>), allowing for a transition from the current limited representation of the genome to a more complete representation. As such, wheat will be the first polyploid genome to be fully represented in the Ensembl system. Although the size and repetitive nature of the genome is a challenge in terms of assembly and annotation, the Ensembl database schema and interface are flexible and should accommodate data from polyploid species with only minor modification. Each of the three genomes will be separately analyzed in the Ensembl comparative analysis pipelines, placing each genome separately within each gene tree and identifying homeologous regions of DNA sequence that will be displayed in an integrated stacked visualization. The existing representation of homeologous variants will be extended to show their functional consequence on gene models.

A more significant challenge lies in data discovery, as the number of available genomes and data sets continues to rise: how can users discover whether information that might be of interest to them exists in the system? We anticipate an increased need for data analysis tools not just as an end in themselves but also as a route of data access. Users will not necessarily start their analyses knowing which genomes or genes they wish to work on; instead, they might wish to ask questions of the complete data set to identify genomes that differ in terms of their gene content or genes whose presence/absence/copy number difference differentiates two genomes (for example, to determine the difference between pathogenic and non-pathogenic strains of related organisms or between competent and non-competent vectors). Likewise, variant features will also be studied in the context of their presence or absence in certain individuals/populations. Supporting these use cases will require rapid gene classification (including common occurrences of novel gene families identified in multiple species) and a high-performance data warehouse to support analysis of the data and help users identify the features of interest within the total data set. Another increasingly important use case is likely to be



Downloaded from <http://nar.oxfordjournals.org/> at Wageningen UR Library on October 11, 2014

Figure 3. The barley genome represented in Ensembl Plants. The figure shows resequencing alignments from a number of cultivars against the reference cultivar Morex genome assembly and annotation for a sequenced contig given approximate chromosomal location through integration with the genetic map.

for the dynamic display of data on demand from archived analyses (e.g. sequence alignments, variant calls) selected on the basis of associated experimental metadata. Developing such tools will be a priority for Ensembl Genomes over the next years.

ACKNOWLEDGEMENTS

The authors would also like to acknowledge the contributions of all our collaborators and of all colleagues working on the Ensembl project.

FUNDING

UK Biosciences and Biotechnology Research Council [BB/I0010077/1, BB/H531519/1, BB/F19793/1, BB/J017299/1, BB/J00328X/1, BB/I008071/1 to P.K.]; Wellcome Trust [090548/B/09/Z to P.K.]; Bill and Melinda Gates Foundation [OPPGD1491 to P.K.]; U.S. National Science Foundation [41686 IPGA Gramene to D.W.]; 7th Framework Programme of the European Union [contract numbers 228421, INFRAVEC; 222886-2, Microme; and 284496, transPLANT to P.K.]. Funding for open access charge: The European Molecular Biology Laboratory.

Conflict of interest statement. None declared.

REFERENCES

1. Flicek,P., Ahmed,I., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fairley,S. *et al.* (2013) Ensembl 2013. *Nucleic Acids Res.*, **41**, D48–D55.
2. The UniProt Consortium. (2012) Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res.*, **41**, D43–D47.
3. Hunter,S., Jones,P., Mitchell,A., Apweiler,R., Attwood,T.K., Bateman,A., Bernard,T., Binns,D., Bork,P., Burge,S. *et al.* (2011) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*, **40**, D306–D312.
4. Kasprzyk,A. (2011) BioMart: driving a paradigm change in biological data management. *Database*, **2011**, bar049–bar049.
5. Megy,K., Emrich,S.J., Lawson,D., Dialynas,E., Hughes,D.S., Koscielny,G., Louis,C., Maccallum,R.M., Redmond,S.N. *et al.* VectorBase: improvements to a bioinformatics resource for invertebrate vector genomics. *Nucleic Acids Res.*, **40**, D729–D734.
6. Yook,K., Harris,T.W., Bieri,T., Cabunoc,A., Chan,J., Chen,W.J., Davis,P., de la Cruz,N., Duong,A. and Fang,R. (2012) WormBase 2012: more genomes, more data, new website. *Nucleic Acids Res.*, **40**, D735–D741.
7. Wood,V., Harris,M.A., McDowall,M.D., Rutherford,K., Vaughan,B.W., Staines,D.M., Aslett,M., Lock,A., Bähler,J. and Kersey,P.J. (2011) PomBase: a comprehensive online resource for fission yeast. *Nucleic Acids Res.*, **40**, D695–D699.
8. Youens-Clark,K., Buckler,E., Casstevens,T., Chen,C., DeClerck,G., Derwent,P., Dharmawardhana,P., Jaiswal,P., Kersey,P. and Karthikeyan,A.S. (2011) Gramene database in 2010: updates and extensions. *Nucleic Acids Res.*, **39**, D1085.
9. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
10. Lappalainen,I., Lopez,J., Skipper,L., Hefferon,T., Spalding,J.D., Garner,J., Chen,C., Maguire,M., Corbett,M., Zhou,G. *et al.* (2013) DbVar and DGVa: public archives for genomic structural variation. *Nucleic Acids Res.*, **41**, D936–D941.
11. Kersey,P.J., Staines,D.M., Lawson,D., Kulesha,E., Derwent,P., Humphrey,J.C., Hughes,D.S., Keenan,S., Kerhornou,A. and Koscielny,G. (2012) Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. *Nucleic Acids Res.*, **40**, D91–D97.
12. Weetman,D., Wilding,C.S., Steen,K., Morgan,J.C., Simard,F. and Donnelly,M.J. (2010) Association mapping of insecticide resistance in wild anopheles gambiae populations: major variants identified in a low-linkage disequilibrium genome. *PLoS One*, **5**, e13140.
13. Weetman,D., Wilding,C.S., Steen,K., Pinto,J. and Donnelly,M.J. (2011) Gene flow-dependent genomic divergence between anopheles gambiae M and S forms. *Mol. Biol. Evol.*, **29**, 279–291.
14. International Barley Genome Sequencing Consortium, Mayer,K.F., Waugh,R., Brown,J.W., Schulman,A., Langridge,P., Platzer,M., Fincher,G.B., Muehlbauer,G.J., Sato,K. *et al.* (2012) A physical, genetic and functional sequence assembly of the barley genome. *Nature*, **491**, 711–716.
15. Chia,J.-M., Song,C., Bradbury,P.J., Costich,D., de Leon,N., Doebley,J., Elshire,R.J., Gaut,B., Geller,L., Glaubitz,J.C. *et al.* (2012) Maize HapMap2 identifies extant variation from a genome in flux. *Nat. Genet.*, **44**, 803–807.
16. McNally,K.L., Childs,K.L., Bohnert,R., Davidson,R.M., Zhao,K., Ulat,V.J., Zeller,G., Clark,R.M., Hoen,D.R., Bureau,T.E. *et al.* (2012) Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proc. Natl Acad. Sci. USA*, **106**, 12273–12278.
17. Zhao,K., Wright,M., Kimball,J., Eizenga,G., McClung,A., Kovach,M., Tyagi,W., Ali,M.L., Tung,C.-W., Reynolds,A. *et al.* (2010) Genomic diversity and introgression in *O. sativa* reveal the impact of domestication and breeding on the rice genome. *PLoS One*, **5**, e10780.
18. Zheng,L.-Y., Guo,X.-S., He,B., Sun,L.-J., Peng,Y., Dong,S.-S., Liu,T.-F., Jiang,S., Ramachandran,S., Liu,C.-M. *et al.* (2011) Genome-wide patterns of genetic variation in sweet and grain sorghum (*Sorghum bicolor*). *Genome Biol.*, **12**, R114.
19. Weigel,D. and Mott,R. (2009) The 1001 genomes project for *Arabidopsis thaliana*. *Genome Biol.*, **10**, 107.
20. Atwell,S., Huang,Y.S., Vilhjálmsson,B.J., Willems,G., Horton,M., Li,Y., Meng,D., Platt,A., Tarone,A.M., Hu,T.T. *et al.* (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature*, **465**, 627–631.
21. Winnenburg,R., Urban,M., Beacham,A., Baldwin,T.K., Holland,S., Lindeberg,M., Hansen,H., Rawlings,C., Hammond-Kosack,K.E. and Köhler,J. (2008) PHI-base update: additions to the pathogen host interaction database. *Nucleic Acids Res.*, **36**, D572–D576.
22. Karsch-Mizrachi,I., Nakamura,Y., Cochrane,G. and on behalf of the International Nucleotide Sequence Database Collaboration (2011) The international nucleotide sequence database collaboration. *Nucleic Acids Res.*, **40**, D33–D37.
23. Salgado,H., Peralta-Gil,M., Gama-Castro,S., Santos-Zavaleta,A., Muñoz-Rascado,L., García-Sotelo,J.S., Weiss,V., Solano-Lira,H., Martínez-Flores,I., Medina-Rivera,A. *et al.* (2013) RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res.*, **41**, D203–D213.
24. Harris,R. (2007) *Improved Pairwise Alignment of Genomic DNA*, Ph.D. Thesis, The Pennsylvania State University.
25. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
26. Kent,W.J., Baertsch,R., Hinrichs,A., Miller,W. and Haussler,D. (2003) Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl Acad. Sci. USA*, **100**, 11484–11489.
27. Vilella,A.J., Severin,J., Ureta-Vidal,A., Heng,L., Durbin,R. and Birney,E. (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, **19**, 327–335.
28. Pedruzzi,I., Rivoire,C., Auchincloss,A.H., Coudert,E., Keller,G., de Castro,E., Baratin,D., Cuche,B.A., Bougueleret,L., Poux,S. *et al.* (2012) HAMAP in 2013, new developments in the protein family classification and annotation system. *Nucleic Acids Res.*, **41**, D584–D589.
29. Mi,H., Muruganujan,A. and Thomas,P.D. (2013) PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.*, **41**, D377–D386.
30. Brechley,R., Spannagl,M., Pfeifer,M., Barker,G.L.A., D'Amore,R., Allen,A.M., McKenzie,N., Kramer,M., Kerhornou,A., Bolser,D. *et al.* (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, **491**, 705–710.