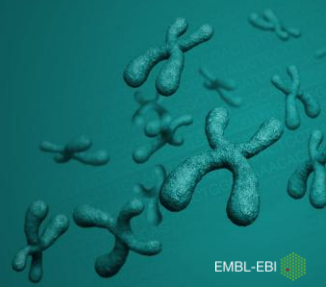
Wheat Genomes in Ensembl Plants

Paul Kersey



EMBL-EBI

Ensembl empowered

- A platform for genome annotation, analysis and browsing, developed jointly by the EBI and Wellcome Trust Sanger Institute
- Modules for handling variation data, comparative genomics, gene prediction etc.
- Multiple points of access to data: web-based application, Perl and REST-ful APIs, public mySQL databases, BioMart data mining tool, DAS client and server, FTP
- Upload your own data and compare to reference annotation
- Originally developed for vertebrate genomes, subsequently extended to non-vertebrate species ("Ensembl Genomes")

EMBL-EBI

vertebrates (ensembl.org)

bacteria.ensembl.org

fungi.ensembl.org

protists.ensembl.org

plants.ensembl.org

metazoa.ensembl.org

empowered

EMBL-EBI

Data

- Genomic sequence
- Gene / transcript / protein models
- External references
- Mapped cDNAs, proteins, microarray probes, BAC clones, repeats, markers etc. etc.
- Variation data:
  - sequence variants
  - structural variants
- Comparative data:
  - gene trees, orthologues, paralogues
  - pairwise whole genomic alignments, syntenic regions

EMBL-EBI

Wheat data in Ensembl Plants (release 21)

- Version1.0 of the chromosome survey sequence assembly generated by the International Wheat Genome Sequencing Consortium.
- Version 2.0 of the IWGSC gene models called on the survey sequence
  - Version 2.1 due in next release (March 2014)
- Whole genome alignments against Brachypodium and rice
- Alignments to wheat Unigenes and RNA-seq data

EMBL-EBI

Chromosome survey sequence: plans for the next release (release 22, due March 2014)

- Whole genome alignments
  - Against barley, *Triticum uratu* and *Aegilops taucshcii*
  - Wheat AvB, AvD, BvD
- Alignment to wheat full length ESTs
- Additional RNA-seq alignments

EMBL-EBI

Genome assemblies from Brenchley *et al*.

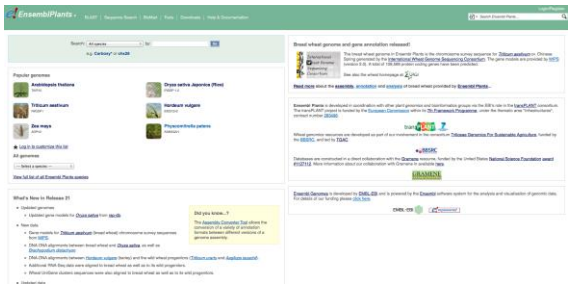
- The wheat genome assemblies generated by Brenchley et al. (PMID: 23192148) have also been aligned to the survey sequence, Brachypodium and barley
  - Homeolegous variants inferred between the 3 wheat genomes (from this work) are also displayed in the context of the gene models from Brachypodium and barley

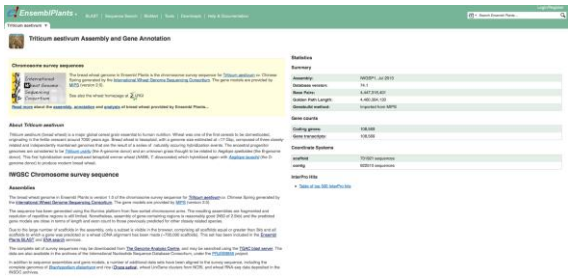
Diploid progenitor genomes

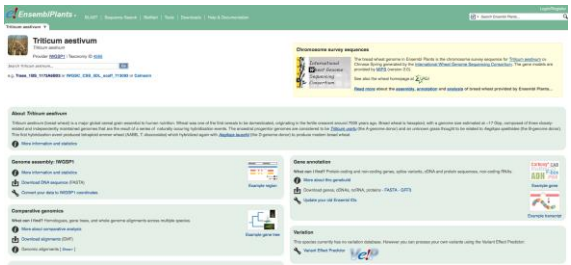
- *Aegilops tauschii* and *Triticum uratu* are also included in Ensembl Plants
- These genomes have been aligned to rice, and to relevant RNA-seq reads.
- These genomes and other more distant relatives (e.g. barley, Brachypodium, and rice) are all included in gene-centric comparative analyses

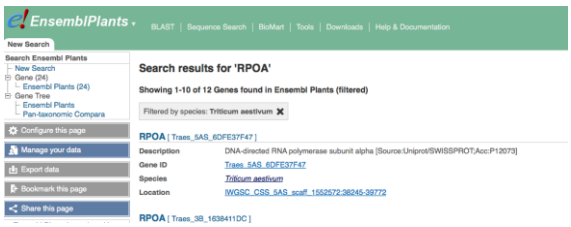
Sequence

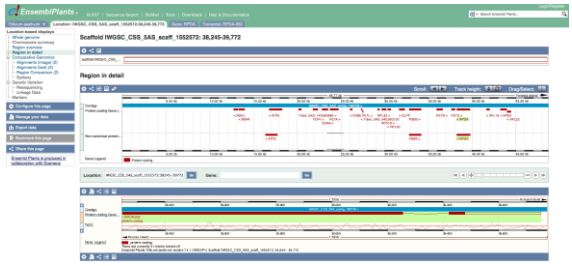
- The chromosome survey sequence (and its annotations) are available to search via BLAST and other search alignment algorithms.
- EST-based search (with onward genomic mapping) also available

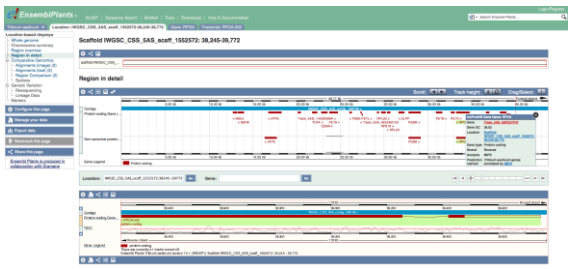


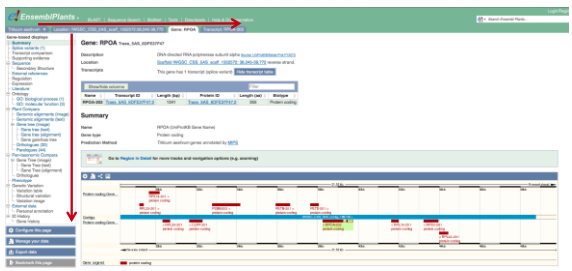


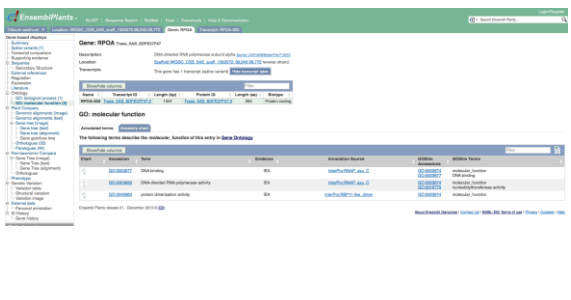


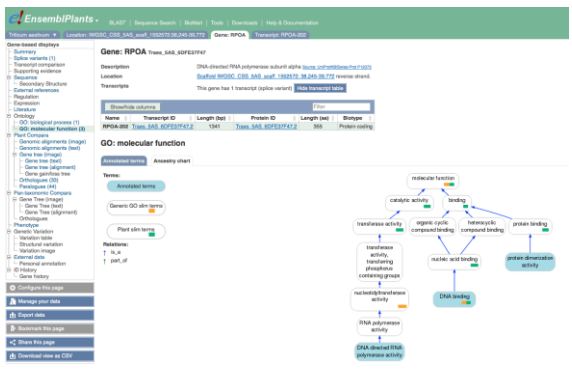

EnsemblPlants
BLAST | Sequence Search | BioMart | Tools | Downloads | Help & Documentation
New Search
Search Ensembl Plants
New Search
Gene (24)
Ensembl Plants (24)
Gene Tree
Ensembl Plants
Pan-taxonomic Compara
Configure this page
Manage your data
Export data
Bookmark this page
Share this page
Search results for 'RPOA'
Showing 1-10 of 12 Genes found in Ensembl Plants (filtered)
Filtered by species: Triticum aestivum
RPOA [Traes_5AS_6DFE37F47]
Description
DNA-directed RNA polymerase subunit alpha [Source:Uniprot/SWISSPROT;Acc:P12073]
Gene ID
Traes_5AS_6DFE37F47
Species
Triticum aestivum
Location
IWGSC_CSS_5AS_scaff_1552572:38245-39772
RPOA [Traes_3B_1638411DC]

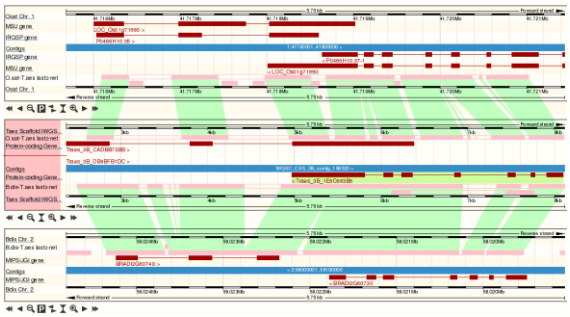

Scaffold IWGSC_CSS_5AS_scaff_1552572: 38,245-39,772
Region in detail

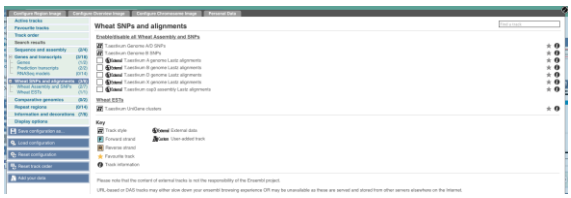

Scaffold IWGSC_CSS_5AS_scaff_1552572: 38,245-39,772
Region in detail

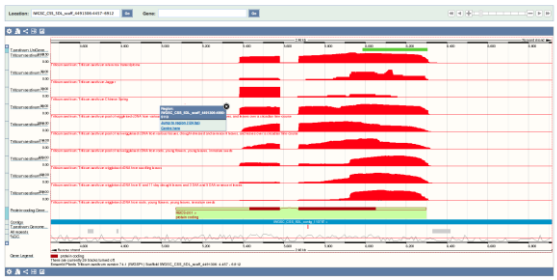

Gene: RPOA
Summary


Gene: RPOA
GO: molecular function


Gene: RPOA
GO: molecular function
molecular function
catalytic activity
binding
transferase activity
organic cyclic compound binding
heterocyclic compound binding
protein binding
nucleic acid binding
DNA binding
RNA polymerase activity
DNA-directed RNA polymerase activity



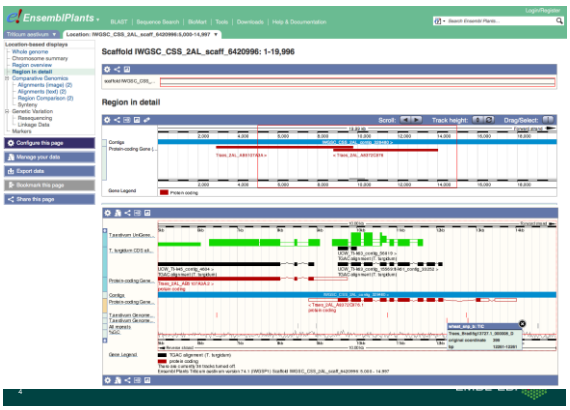




Inter-homeologeous variants

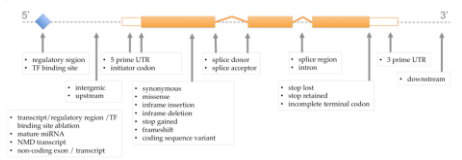
- Data from Brenchley et al. mapped onto chromosome survey sequence
- Data from chromosome survey sequence itself due in next release

	Inter-homeologue variants (total)	Inter-homeologue variants (mapped)	% variants mapped
A/D	102,593	101,570	99
B	22,959	29,638	99
Total	132,552	131,208	99



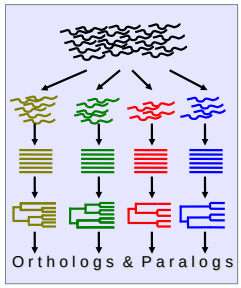
Variant Effect Predictor (VEP)

- 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
- http://www.ensembl.org/info/docs/variation/vep/index.html



EMBL-EBI

Gene tree pipeline



Take canonical protein for each gene belonging to one Ensembl Genomes clade

Cluster: WU-BLASTP + Smith-Waterman all-versus-all, hcluster_sg

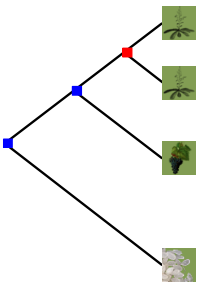
Align: multiple aligners consensified by M-Coffee

Build trees: PhyML-WAG + PhyML-HKY + NJ-p + NJ-dN + NJ-dS + species tree → TreeBeST-merge

Infer orthologues and paralogues

EMBL-EBI

Orthologues and paralogues



Paralogues:

Any gene pairwise relationship where the ancestor node is a duplication event

Orthologues:

Any gene pairwise relationship where the ancestor node is a speciation event

EMBL-EBI

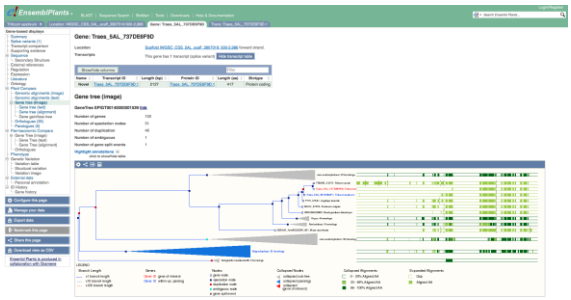
Orthology / paralogy types

- ortholog_one2one
- ortholog_one2many
- ortholog_many2many
- apparent_ortholog_one2one
- possible_ortholog (weakly supported duplication node)

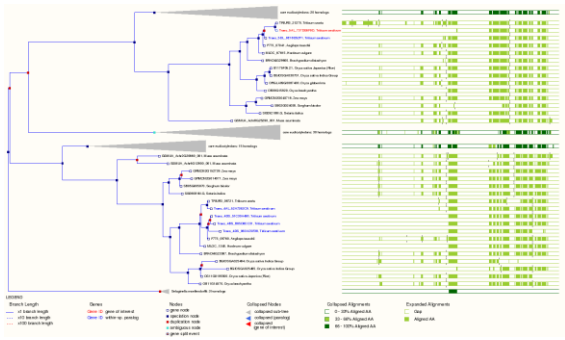
- within_species_paralog
- other_paralog (too distant to be in the same tree)

- contiguous_gene_split (artefact)
- putative_gene_split (artefact)
- 94,280 wheat genes in inferred orthology relationships with 270216 genes from other creals

EMBL-EBI



EMBL-EBI



EMBL-EBI

Visualise your own data

Upload data
- Data saved by Ensembl
- 5 MB limit (and therefore not possible for large file formats)

Attach remote file
- URL-based (HTTP or FTP)
- No size limit

- Uploaded / attached tracks can be saved (to account)
- Uploaded / attached tracks can be shared with other users
- Only trivial security, do not use for sensitive data!

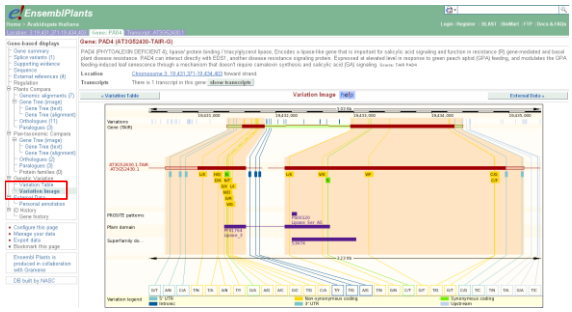
- http://www.ensembl.org/info/website/upload/index.html

Possible formats

- BAM sequence alignments (no upload)
- BED genes / features
- BedGraph continuous-valued data
- BigBed genes / features (no upload)
- BigWig continuous-valued data (no upload)
- TrackHub collection of tracks
- Gbrowse genes / features
- GFF/GTF genes / features
- PSL sequence alignments
- VCF variants (no upload)
- WIG continuous-valued data

Next release

- Improved representation of the polyploid genome
  - Preconfigured A v B v D compartive views
  - Separation of 3 wheat genomes in the gene tree analysis
- Representation of inter-homologue data as a full variation database
  - Data structure can also accommodate inter-varietal SNPs, links to phenotypes, haplotype structure etc.



Funding

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

People

- James Allen, Irina Armean, Dan Bolser, Mikkel Christensen, Paul Davies, Lee Falin, Christoph Grabmueller, Kevin Howe, Malcolm Hinsley, Jay Humphrey, Arnaud Kerhornou, 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

- The Ensembl teams @ EBI (Paul Flicek) and WTSI (Steve Searle)