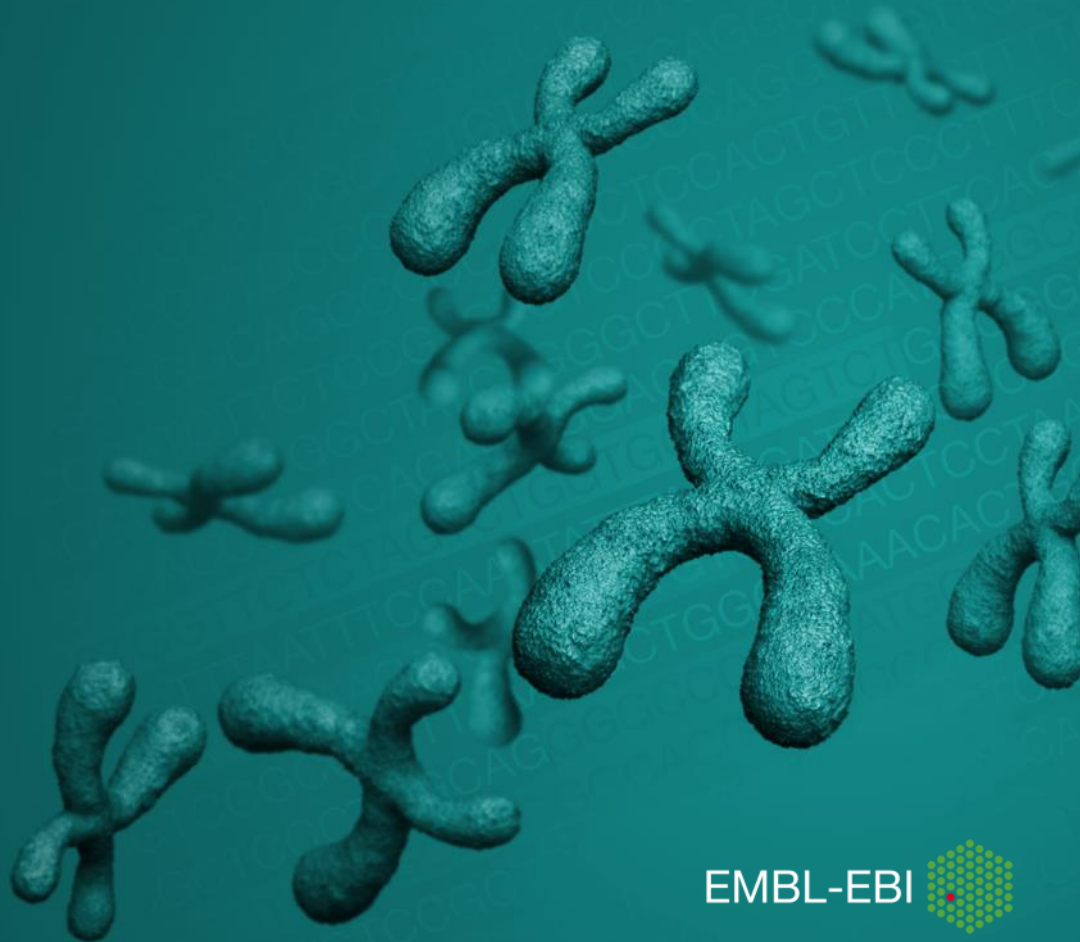


# Adventures in Cereal Genomics

## And Future Directions for Genomic Infrastructure

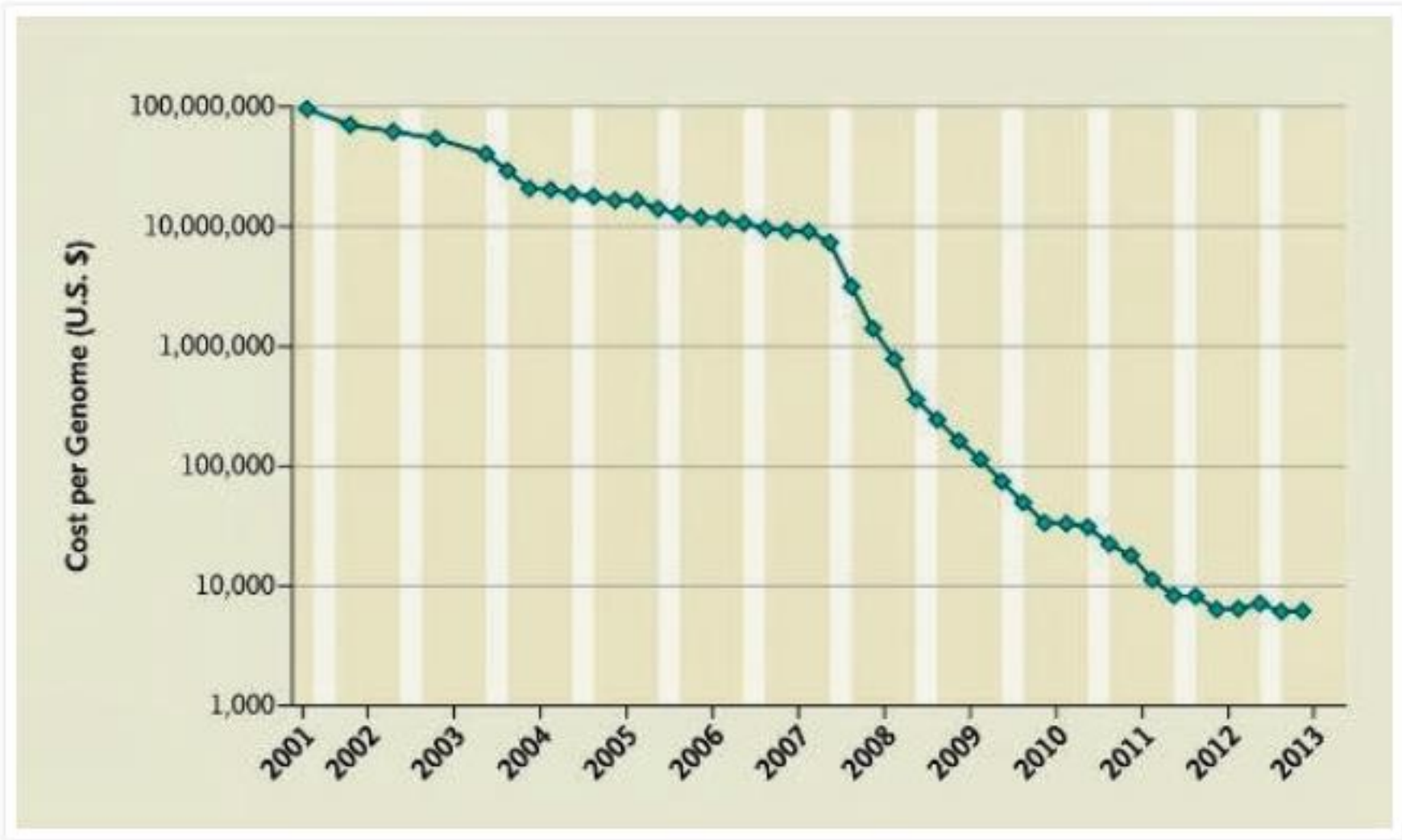
Paul Kersey



# A brief history of genome sequencing

- 1995 *Haemophilus influenzae* 1.8 Mb
  - 1996 *Saccharomyces cerevisiae* 12 Mb
  - 1999 *Drosophila melanogaster* 140 Mb
  - 2001 *Homo sapiens* 3.1 Gb
- 
- Sequencing technology is continuously improving, but (massively parallel) “next generation” techniques really were game-changers

# Cost of Sequencing a Human Genome 2001-2013



# A brief history of genome sequencing

- 2008-2015 1000 genomes project (2500 human genomes)
- 2008-2015 1001 genomes project (1,001 Arabidopsis genomes)
- 2015-2019 Genomics England (100,000 human genomes)

# What can we do with thousands of genome sequences?

- Statistical association of traits with markers
- Increased marker resolution to find causative variants
- Understand population structure and evolutionary processes
  - Track epidemics
- Assay for known variation
  - Environmental distribution
  - Tool for managing crosses
- More genomes...
  - More statistical power, find rarer causative alleles

# Thousands of genomes – a tool for breeding

- Characterize germplasm of land races and wild relatives
- Understand what's actually present in an existing line
- Find alleles associated with traits
  - Combine genotyping with various (laboratory, greenhouse, field) phenotyping mechanisms, themselves increasingly automated and high-throughput
- Manage crosses

# The EBI mission

- **EMBL-EBI provides freely available data from life science experiments, performs basic research in computational biology and offers an extensive user training programme, supporting researchers in academia and industry.**



# Data resources at EMBL-EBI

## Genes, genomes & variation

European Nucleotide Archive

Ensembl  
Ensembl Genomes

European Genome-phenome Archive  
Metagenomics portal

1000 Genomes

## Gene, protein & metabolite expression

ArrayExpress  
Expression Atlas

Metabolights  
PRIDE

## Literature & ontologies

Europe PubMed Central  
Gene Ontology  
Experimental Factor Ontology

## Protein sequences, families & motifs

InterPro

Pfam

UniProt

## Molecular structures

Protein Data Bank in Europe  
Electron Microscopy Data Bank

## Chemical biology

ChEMBL

ChEBI

## Reactions, interactions & pathways

IntAct

Reactome

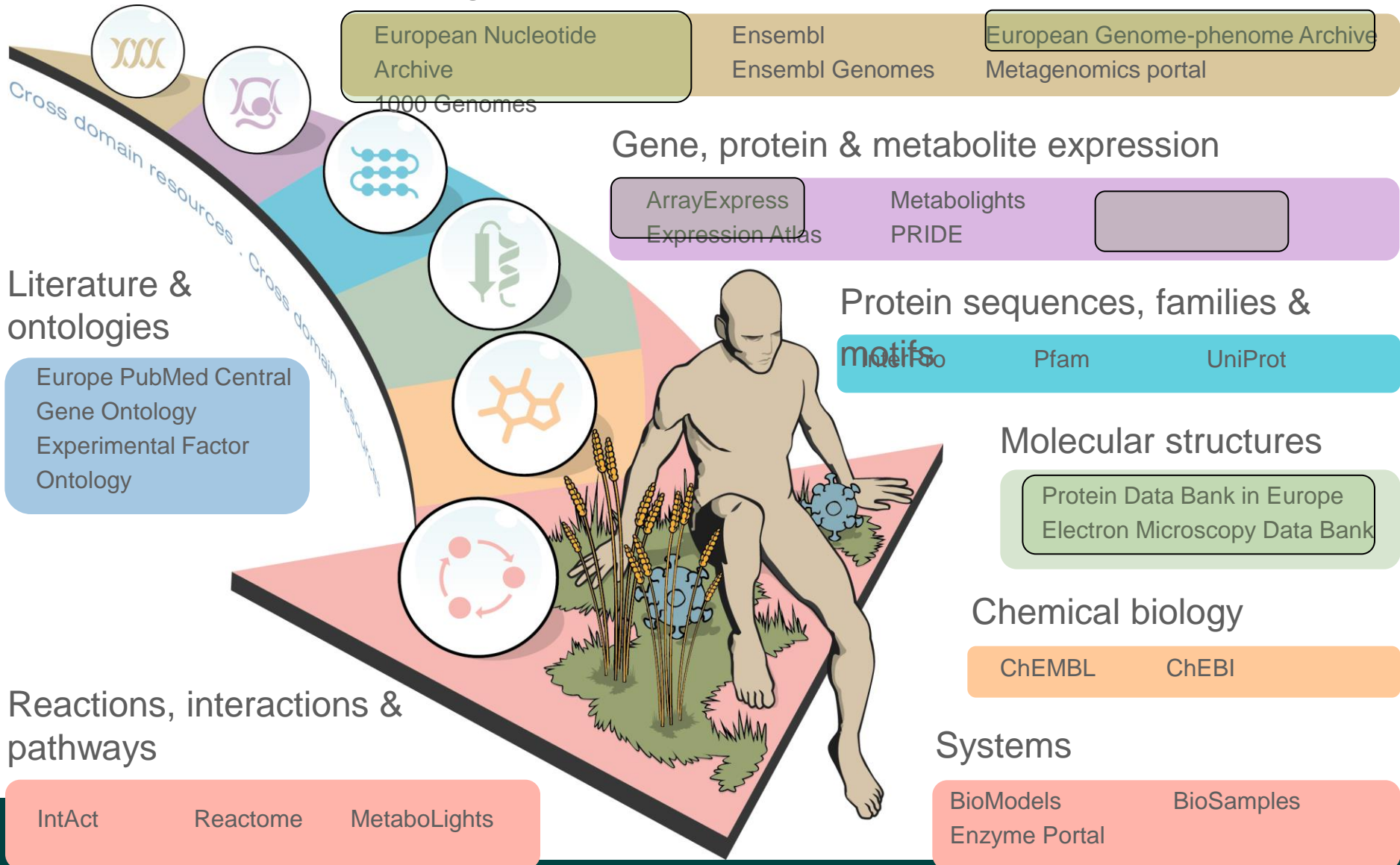
MetaboLights

## Systems

BioModels

Enzyme Portal

BioSamples







- A modular suite of software for genome analysis and visualisation developed jointly by the Wellcome Trust Sanger Institute and the European Bioinformatics Institute
- Now used for genomes from across the taxonomic space
- Offers a standard set of interfaces to a wide range of genome-scale data, including:
  - Web-based GUI
  - Public mySQL server
  - Perl and REST-ful API
  - FTP
  - Data mining tool (constructed using BioMart) framework with its own set of interfaces: web GUI, web services, command line and local client

# vertebrates (ensembl.org)

bacteria.ensembl.org

fungi.ensembl.org

protists.ensembl.org

plants.ensembl.org

metazoa.ensembl.org


# Talk outline


- Introduction to Ensembl Plants
- Cereal genomes in Ensembl
- Future directions for genomic infrastructure


**Ensembl Genomes: Extending Ensembl across the taxonomic space.**


▶  Strigamia maritima and Megaselia scalaris genomes NEW!

In Release 18 we have added [Strigamia maritima](#), a coastal European centipede, and [Megaselia scalaris](#), a scuttle fly of the Phoridae family, also known as the coffin fly or humpbacked fly.

▶  Brugia malayi genome NEW!

▶  Barley synteny NEW!


▶  Wheat sequence search

▶  GO annotation extensions

Ensembl Genomes is developed by [EMBL-EBI](#) and is powered by [Ensembl](#) software system for the analysis and visualisation of genomic data. For details of our funding please [click here](#).



**EnsemblBacteria**

 Over 6000 bacterial genome sequences have been annotated and deposited in the public archives of the members of the [International Nucleotide Sequence Database Collaboration](#). This site provides access to complete, annotated bacterial genomes (present in the [European Nucleotide Archive](#)) through the Ensembl graphical user interface (genome browser). More details about the integration are provided [here](#)

Programmatic access is available through the Ensembl Perl and REST-ful APIs and through publicly accessible mysql databases, along with full data dumps (including DNA sequence and protein sequence in FASTA format, annotations in GTF format, and mysql dump files). Due to the large number of these databases, there is some modification to the APIs, and database and FTP site structure, compared to that used for other branches of the taxonomy (e.g. the storage of many genomes in one database; the provision of lookup services to identify genomes by INSDC identifiers, taxonomy identifiers, or partial names. [Full details are available here](#).

BioMart access is not available, but we are working on providing new, more powerful data mining tools to allow users to exploit these genomes. A selection of over 100 key bacterial genomes has been included in the pan-taxonomic [Compara](#), and genes from all genomes are classified into families using HAMAP and PANTHER ([more details](#))

**What's New in Release 18 (April 2013)**

The eighteenth release of Ensembl Genomes features updates to version 71 of the Ensembl software across all divisions, and a number of new genomes (including 6,305 genomes in the latest version of Ensembl Bacteria) bringing the total number of genomes to 6431 ([full list](#)). Detailed notes can be found [here](#). See the individual homepages for [Bacteria](#), [Protists](#), [Fungi](#), [Plants](#) and [Metazoa](#) for more information.

**Have a question?**

New Frequently Asked Questions (FAQs) are now available for all domains of Ensembl Genomes. Have a question? Check if it's been asked before! If there is a FAQ

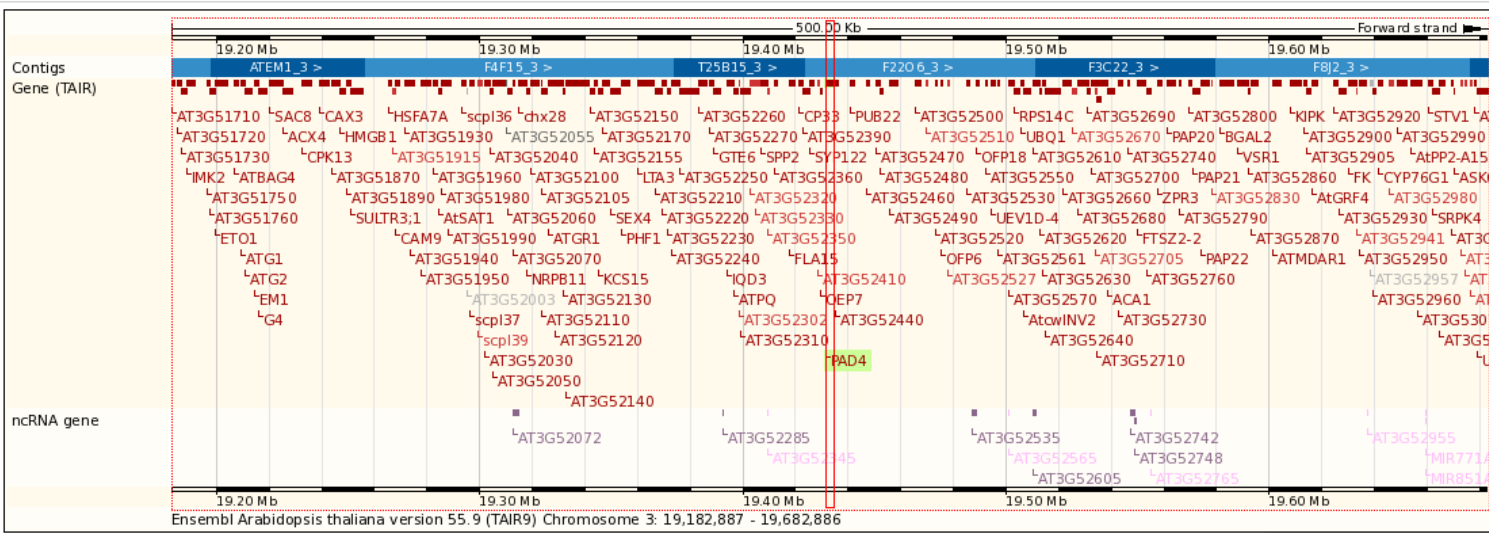
- Location-based displays**
- Whole genome
  - Chromosome summary
  - Region overview
  - Region in detail**
  - Comparative Genomics
    - Genomic alignments (7)
    - Synteny (0)
  - Genetic Variation
    - Resequencing (0)
    - Linkage Data
  - Markers
  - Other genome browsers
- Configure this page  
 Manage your data  
 Export data  
 Bookmark this page
- Ensembl Plants is produced in collaboration with Gramene
- DB built by NASC

**Chromosome 3: 19,431,371-19,434,403**

chromosome 3

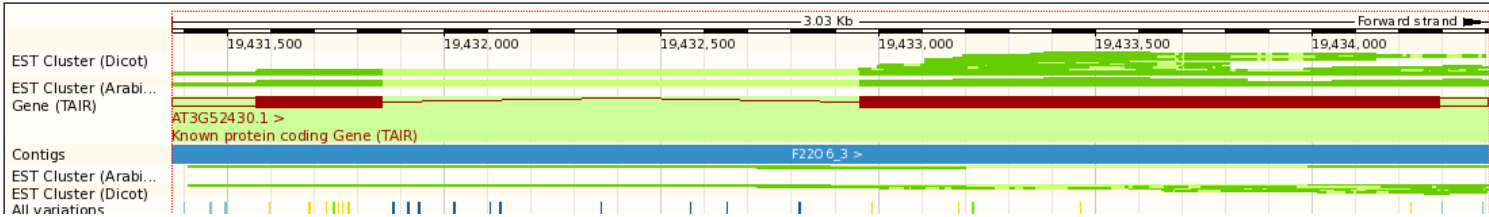
[Export image](#)

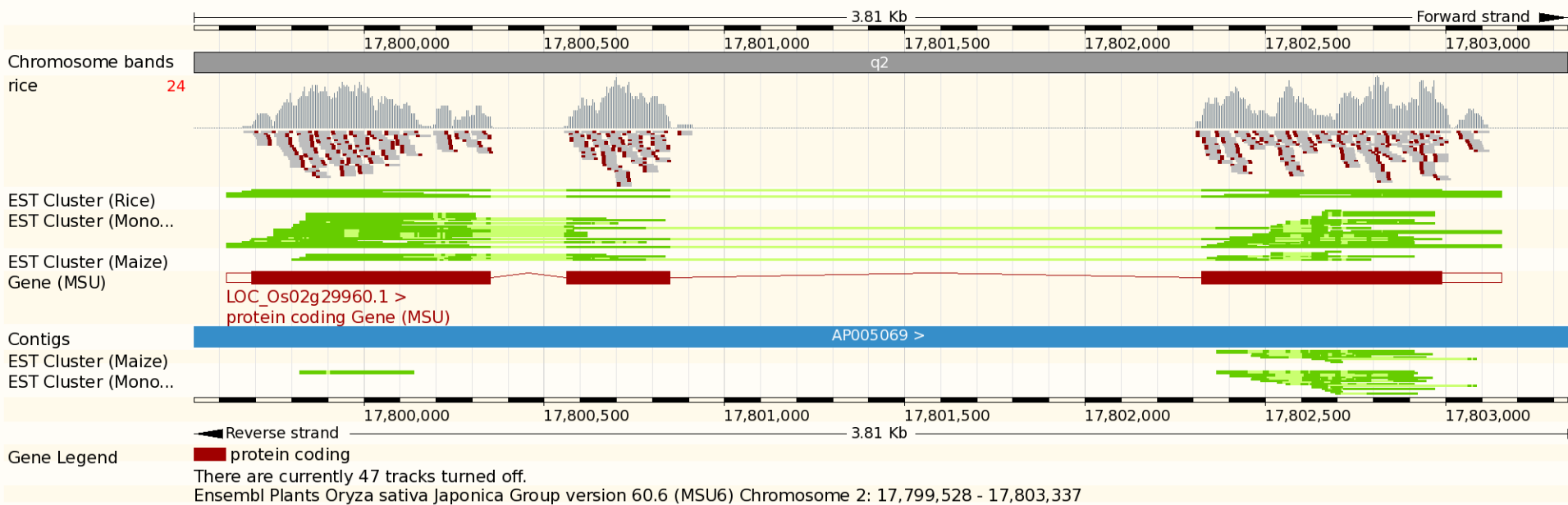
« [Region overview](#) **Region in detail** [help](#) [Genomic alignments »](#)



Location:  :  -  [Go>](#)

Navigation icons: << < > >> << >> << >> << >>





**Gene-based displays**

- Summary
- Splice variants
- Transcript comparison
- Supporting evidence
- Gene alleles
- [-] Sequence
  - Secondary Structure
- External references
- Regulation
- Literature
- Ontology
- [-] Plant Compara
  - Genomic alignments
  - Gene tree
  - Gene gain/loss tree
  - Orthologues
  - Paralogues
- [-] Pan-taxonomic Compara
  - Gene Tree
  - Orthologues
- Phenotype
- Genetic Variation
  - Variation table
  - Structural variation
  - Variation image
- [-] External data
  - **Gene expression**
  - Personal annotation
- [-] ID History
  - Gene history

Configure this page

Add your data

Export data

Share this page

Bookmark this page

Ensembl Plants is produced in collaboration with Gramene

## Gene: OS12G0515800

**Description** Expressed protein; cDNA clone:J013098I09, full insert sequence [Source:UniProtKB/TrEMBL;Acc:Q2QPV9]

**Location** [Chromosome 12: 20,042,293-20,043,043](#) forward strand.

**About this gene** This gene has 1 transcript ([splice variant](#)), [12 orthologues](#) and [2 paralogues](#).

**Transcripts**

[Hide transcript table](#)

Show/hide columns (1 hidden)		Filter				
Name	Transcript ID	bp	Protein	Biotype	UniProt	Flags
Novel	<a href="#">OS12T0515800-01</a>	504	<a href="#">78aa</a>	Protein coding	<a href="#">Q2QPV9</a>	

## Gene expression



Showing 2 of 2 experiments found: ?

Experiment	anther	callus	early inflores...	embryo	emerging info...	endosperm	leaf	pistil	post-flowering...	pre-flowering...	root	seed	seed 10 days...	seed 5 days...	shoot
	<a href="#">Tissues - 9 Davidson</a>		NA							NA	NA	NA	NA		
<a href="#">Tissues - 7 Sakai</a>	NA		NA	NA	NA	NA		NA					NA	NA	

Up to 50 of top genes displayed on page. Download results to see the rest.

[Display levels](#)

This expression view is provided by [Expression Atlas](#). Please direct any queries or feedback to [arrayexpress-atlas@ebi.ac.uk](mailto:arrayexpress-atlas@ebi.ac.uk)

Ensembl Plants release 28 - August 2015 © [EBI](#)

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

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

Ensembl Plants is produced in collaboration with Gramene

DB built by NASC

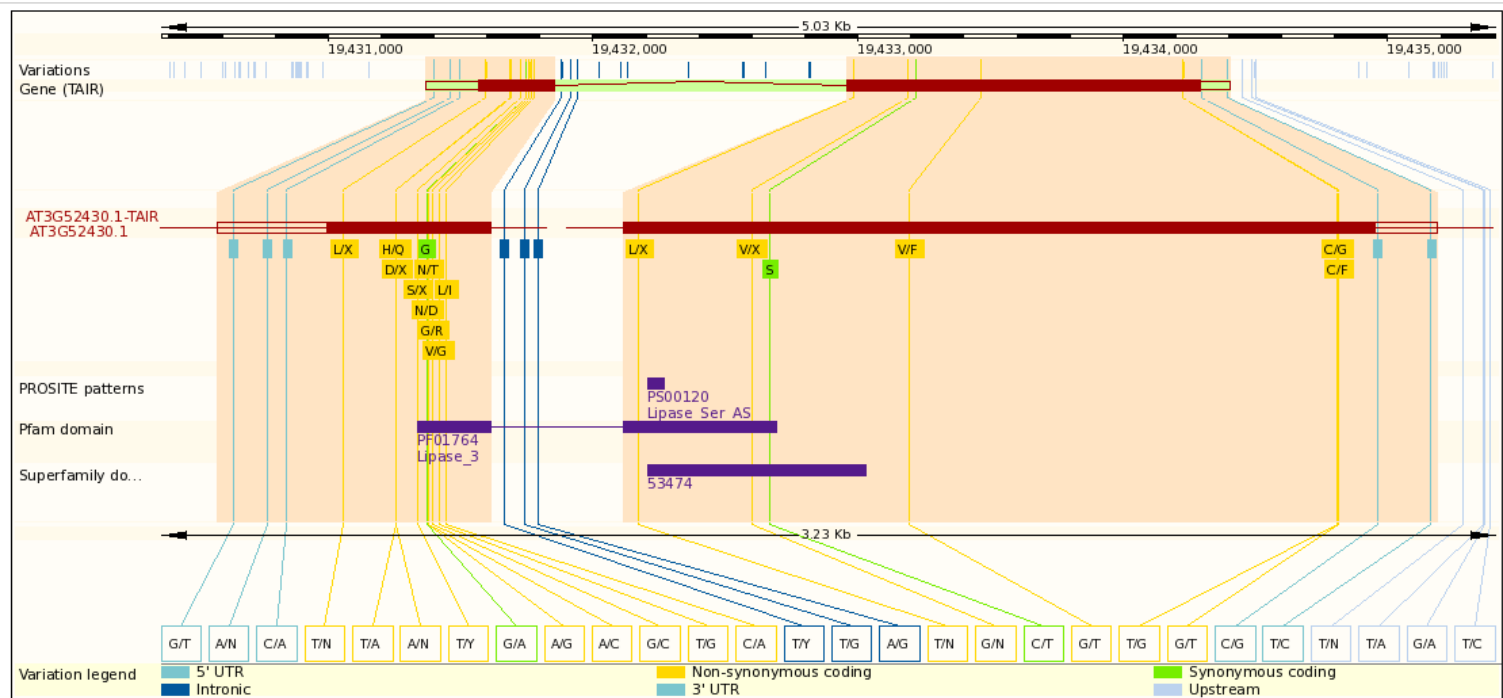
Gene: PAD4 (AT3G52430-TAIR-G)

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.

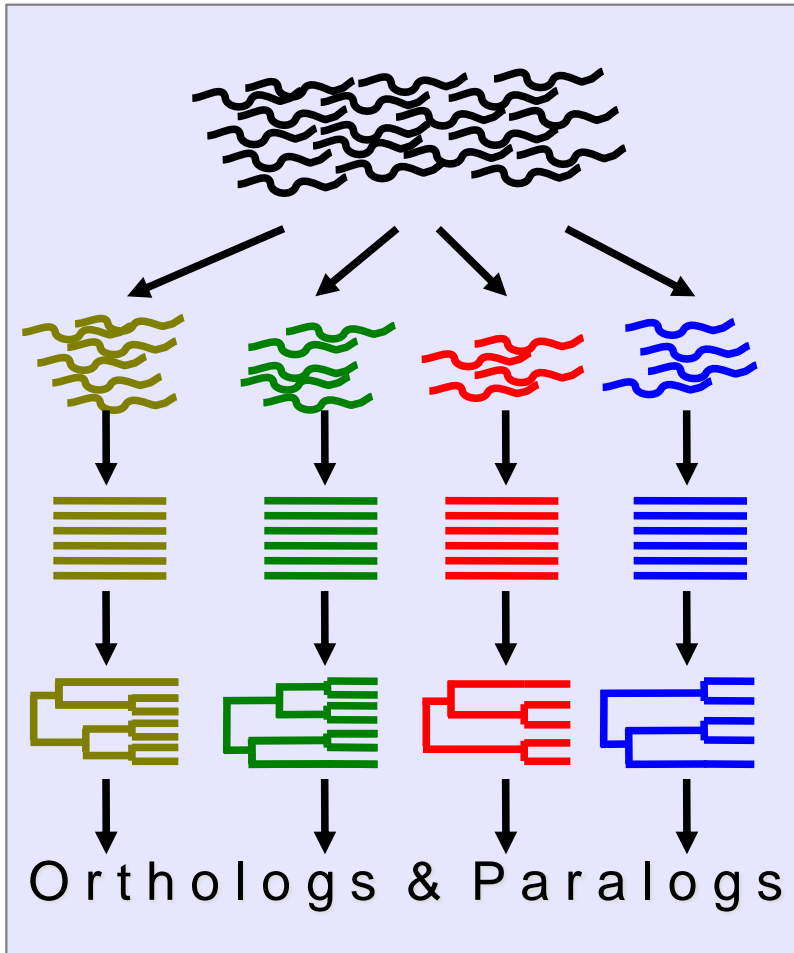
Transcripts There is 1 transcript in this gene: [show transcripts](#)

« Variation Table Variation Image help External Data »





# 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

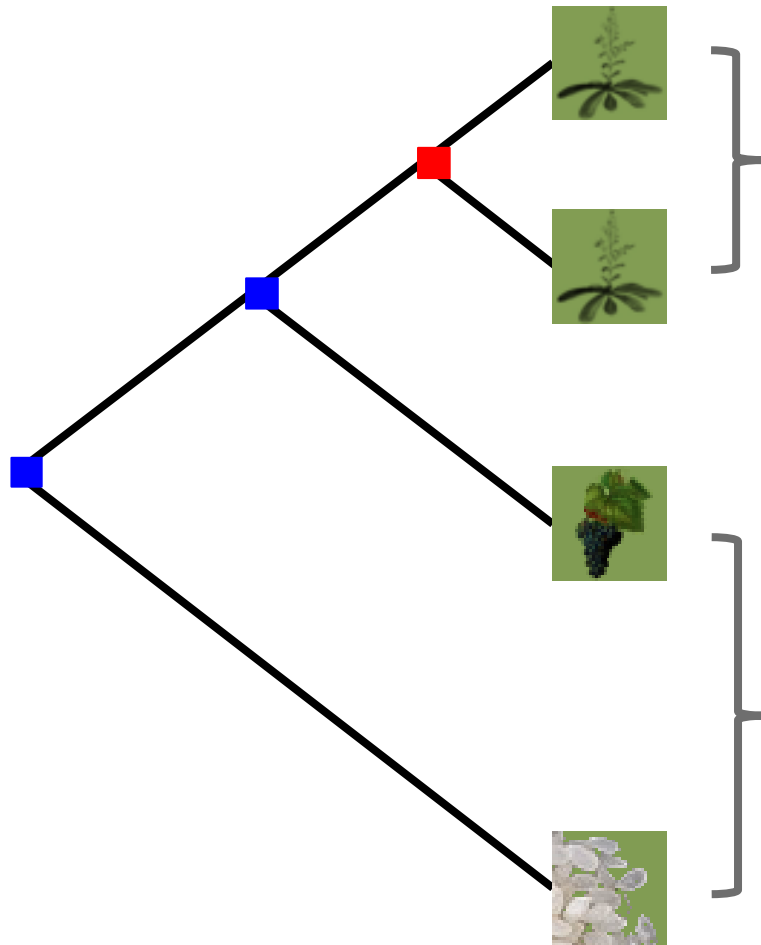
**Align:** multiple aligners consensified by M-Coffee

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

**Infer orthologues and paralogues**

# Orthologues and paralogues

---



Paralogues:

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

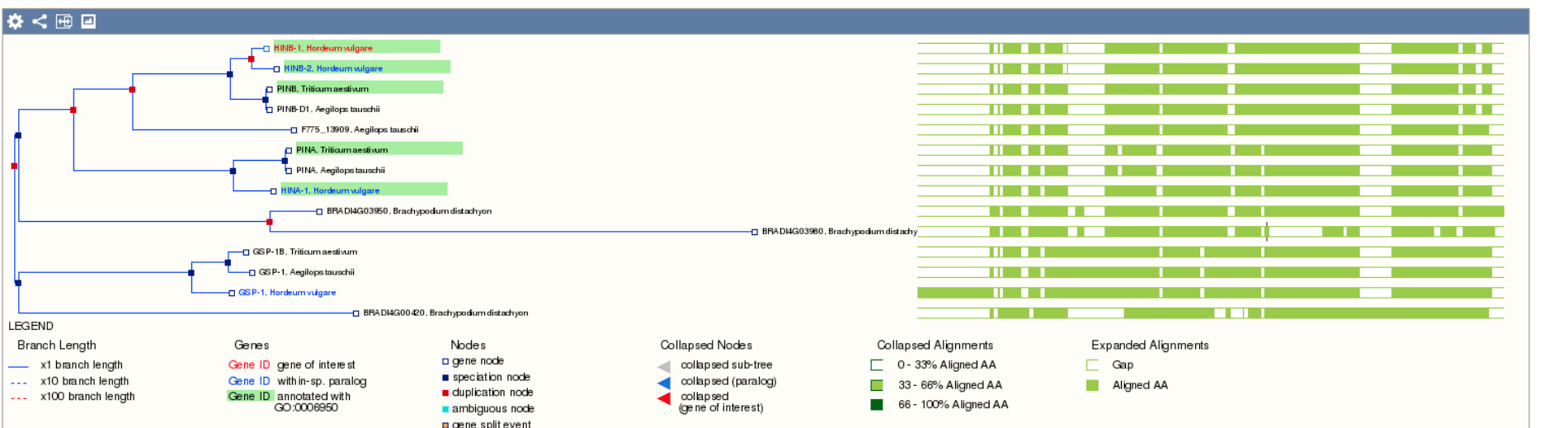
Orthologues:

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

Show 10 entries Filter

highlight	Accession	Description
<input type="radio"/> 7 members	<a href="#">GO:0008150</a>	biological process
<input type="radio"/> 5 members	<a href="#">GO:0003674</a>	molecular function
<input type="radio"/> 5 members	<a href="#">GO:0004857</a>	enzyme inhibitor activity
<input type="radio"/> 5 members	<a href="#">GO:0004866</a>	endopeptidase inhibitor activity
<input type="radio"/> 5 members	<a href="#">GO:0004867</a>	serine-type endopeptidase inhibitor activity
<input type="radio"/> 5 members	<a href="#">GO:0005575</a>	cellular component
<input type="radio"/> 5 members	<a href="#">GO:0005576</a>	extracellular region
<input type="radio"/> 5 members	<a href="#">GO:0005615</a>	extracellular space
<input checked="" type="radio"/> 5 members	<a href="#">GO:0006950</a>	response to stress
<input type="radio"/> 5 members	<a href="#">GO:0006952</a>	defense response

Showing 1 to 10 of 48 entries << < 1 2 3 4 5 > >>



# Pairwise whole genome alignments & synteny

---

- Only for certain combinations of species
- Generated using (B)LASTz-net

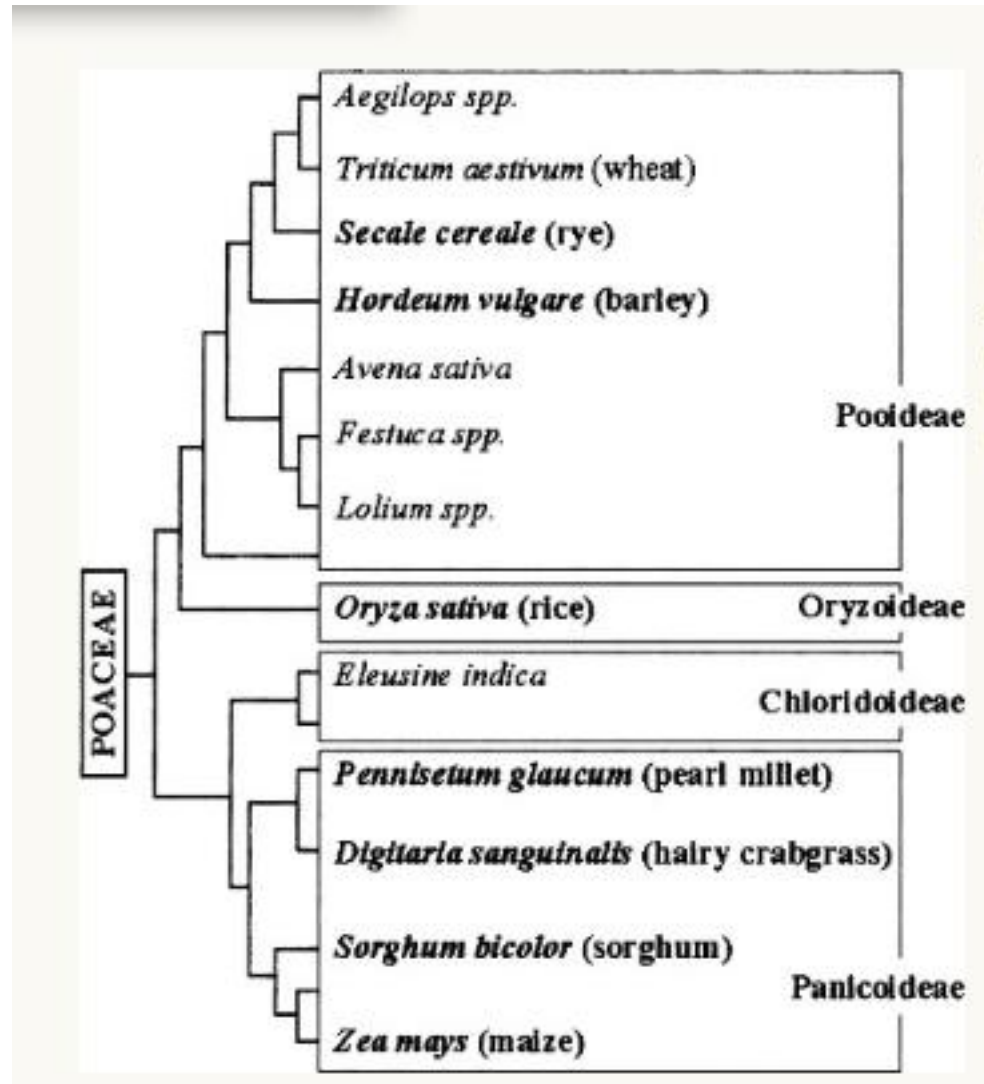
## Synteny

- Organisms of relatively recent divergence show similar blocks of genes in the same relative positions in the genome
- Shows how the genome is “cut and pasted” in the course of evolution
- Calculated using pairwise whole genome alignments
- Only for certain combinations of species

# Poaceae, true grasses

- With more than 10,000 domesticated and wild species, the Poaceae represent the fifth-largest plant family
- Grasslands are estimated to compose 20% of the vegetation cover of the Earth
- Domestication of poaceous cereal crops such as maize (corn), wheat, rice, barley, and millet lies at the foundation of sedentary living and civilization around the world, and the Poaceae still constitute the most economically important plant family in modern times, providing forage, building materials (bamboo, thatch) and fuel (ethanol), as well as food.

# Grasses



[Journal of Experimental Botany Volume 56, Issue 416](#)

# Barley and Wheat



- 2 of the most important cereal crops (ranked 5<sup>th</sup>/1<sup>st</sup> for global food production)
- Barley can survive in a wider climactic range
- Estimated evolutionary distance is 13 Mya (Human, Orang Utan 15 Mya)

**BREAD FREE** 12.10.13

David  
Perlmutter, MD

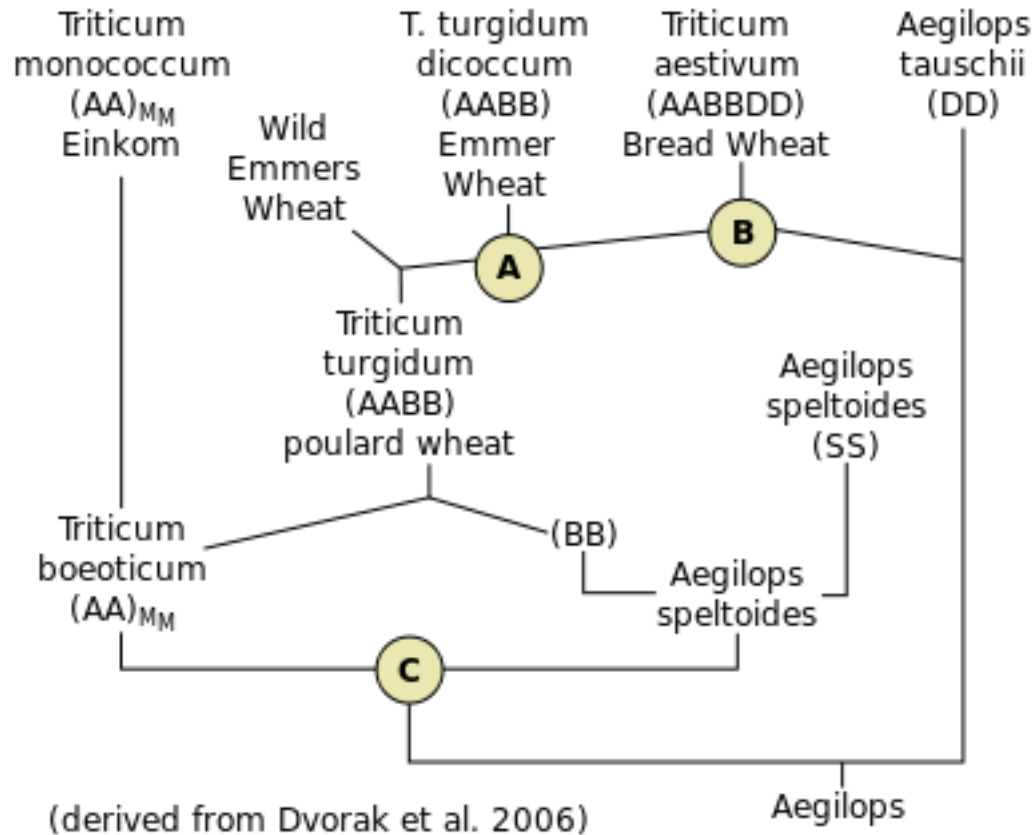


# Wheat Threatens All Humans, New Research Shows

Bread lovers beware! Gluten-free diets may not just be a trendy choice but something everyone should follow. New research reveals that proteins in wheat may be detrimental to all humans.



# Evolution of hexaploid bread wheat



Event A - Giazentep/Euphrates, SE Turkey ~8000 BC.  
 Event B - Southeastern Armenia 6500 BC  
 Event C - 4 million years ago

# Large cereal genomes are approaching completion



**nature** International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

Archive > Volume 491 > Issue 7426 > Articles > Article

NATURE | ARTICLE OPEN

日本語要約

## Analysis of the bread wheat genome using whole-genome shotgun sequencing

Rachel Brenchley, Manuel Spannagl, Matthias Pfeifer, Gary L. A. Barker, Rosalinda D'Amore, Alexandra M. Allen, Neil McKenzie, Melissa Kramer, Arnaud Kerhornou, Dan Bolser, Suzanne Kay, Darren Waite, Martin Trick, Ian Bancroft, Yong Gu, Naxin Huo, Ming-Cheng Luo, Sunish Sehgal, Bikram Gill, Sharyar Kianian, Olin Anderson, Paul Kersey, Jan Dvorak, W. Richard McCombie, Anthony Hall *et al.*

Affiliations | Contributions | Corresponding authors

Nature 491, 705–710 (29 November 2012) | doi:10.1038/nature11650  
Received 04 March 2012 | Accepted 01 October 2012 | Published online 28 November 2012



**Science** AAAS.ORG | FEEDBACK | HELP | LIBRARIANS

All Science Journals Enter Search Term

WELLCOME TRUST GANGER INSTITUTE ALERTS | ACCESS RIGHTS

NEWS SCIENCE JOURNALS CAREERS MULTIMEDIA COLLECTIONS

**Science** The World's Leading Journal of Original Scientific Research, Global News, and Commentary.

Science Home Current Issue Previous Issues Science Express Science Products My Science About the Journal

Home > Science Magazine > 18 July 2014 > Choulet *et al.*, 345 (6194):

Article Views

Abstract

Full Text

Full Text (PDF)

Figures Only

Supplementary Materials

Article Tools

Leave a comment (0)

Save to My Folders

Download Citation

Science 18 July 2014:  
Vol. 345 no. 6194  
DOI: 10.1126/science.1249721

RESEARCH ARTICLE

## Structural and functional partitioning of bread wheat chromosome 3B

Frédéric Choulet<sup>1,2,4</sup>, Adriana Alberti<sup>3</sup>, Sébastien Theil<sup>1,2</sup>, Natasha Glover<sup>1,2</sup>, Valérie Barbe<sup>3</sup>, Josquin Daron<sup>1,2</sup>, Lise Pingault<sup>1,2</sup>, Pierre Sourdille<sup>1,2</sup>, Arnaud Couloux<sup>3</sup>, Etienne Paux<sup>1,2</sup>, Philippe Leroy<sup>1,2</sup>, Sophie Mangenot<sup>3</sup>, Nicolas Guilhot<sup>1,2</sup>, Jacques Le Gouis<sup>1,2</sup>, Francois Balfourier<sup>1,2</sup>, Michael Alaux<sup>4</sup>, Véronique Jamilloux<sup>4</sup>, Julie Poulain<sup>3</sup>, Céline Durand<sup>3</sup>, Arnaud Bellec<sup>5</sup>, Christine Gaspin<sup>6</sup>, Jan Safar<sup>7</sup>, Jaroslav Dolezel<sup>7</sup>, Jane Rogers<sup>8</sup>, Klaas Vandepoel<sup>9</sup>, Jean-Marc Aury<sup>3</sup>, Klaus Mayer<sup>10</sup>, Hélène Berges<sup>5</sup>, Hadi Quesneville<sup>4</sup>, Patrick Wincker<sup>3,11,12</sup>, Catherine Feuillet<sup>1,2</sup>

Author Affiliations

Chapman *et al.* *Genome Biology* (2015) 16:26  
DOI 10.1186/s13059-015-0582-8

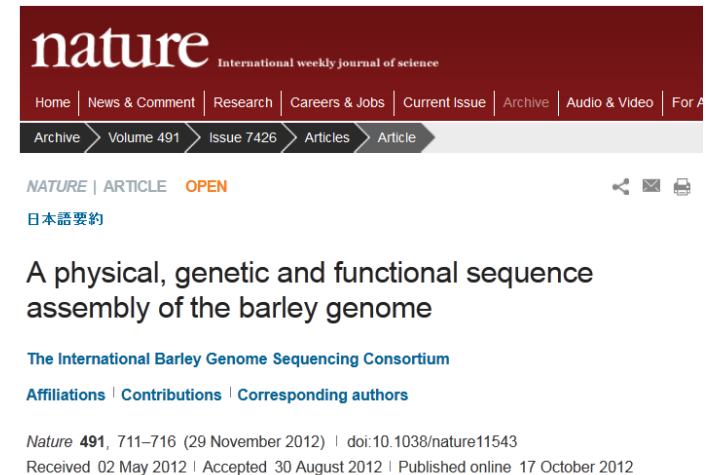


METHOD

Open Access

## A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome

Jarrold A Chapman<sup>1†</sup>, Martin Mascher<sup>2†</sup>, Aydin Buluç<sup>3</sup>, Kerrie Barry<sup>1</sup>, Evangelos Georganas<sup>3,4</sup>, Adam Session<sup>5</sup>, Veronika Strnadova<sup>6</sup>, Jerry Jenkins<sup>1,7</sup>, Sunish Sehgal<sup>8,11</sup>, Leonid Olikier<sup>3</sup>, Jeremy Schmutz<sup>1,7</sup>, Katherine A Yelick<sup>3,4</sup>, Uwe Scholz<sup>2</sup>, Robbie Waugh<sup>9</sup>, Jesse A Poland<sup>8</sup>, Gary J Muehlbauer<sup>10</sup>, Nils Stein<sup>2</sup> and Daniel S Rokhsar<sup>1,5\*</sup>



**nature** International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

Archive > Volume 491 > Issue 7426 > Articles > Article

NATURE | ARTICLE OPEN

日本語要約

## A physical, genetic and functional sequence assembly of the barley genome

The International Barley Genome Sequencing Consortium

Affiliations | Contributions | Corresponding authors

Nature 491, 711–716 (29 November 2012) | doi:10.1038/nature11543  
Received 02 May 2012 | Accepted 30 August 2012 | Published online 17 October 2012

# Cereal genomes: how large?

Species	Genome size (n)
<i>Arabidopsis thaliana</i>	120 Mb
<i>Oryza sativa</i>	374 Mb
<i>Setaria italica</i>	405 Mb
<i>Sorghum bicolor</i>	784 Mb
<i>Zea mays</i>	2.07 Gb
<i>Brachypodium distachyon</i>	2.71 Gb
<i>Aegilops tauschii</i>	3.31 Gb
<i>Homo sapiens</i>	3.33 Gb
<i>Triticum urartu</i>	3.74 Gb
<i>Hordeum vulgare</i>	4.70 Gb
<i>Triticum aestivum (3n)</i>	~ 16 Gb

# Cereal genomes: assembly fragmentation

Species	Contig count	Scaffold count
<i>Arabidopsis thaliana</i>	24	7
<i>Oryza sativa</i>	972	61
<i>Homo sapiens</i>	992	297
<i>Brachypodium distachyon</i>	1,754	83
<i>Setaria italica</i>	6,791	336
<i>Sorghum bicolor</i>	12,873	3,304
<i>Zea mays</i>	127,422	523
<i>Triticum aestivum</i> **	1,043,087	731,921
<i>Aegilops tauschii</i>	1,107,056	492,892
<i>Hordeum vulgare</i>	1,380,863	118,942
<i>Triticum urartu</i>	1,455,436	499,221

\*\*  
Scaffolds >  
3 Kb or  
containing  
a gene  
model from  
the  
IWGSC  
survey  
sequence

Triticum aestivum Location: 5D:17,856,312-17,892,564

Location-based displays

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

Configure this page

Add your data

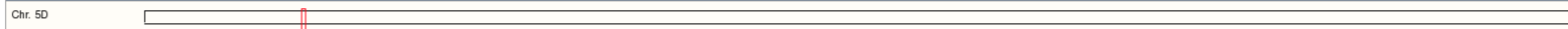
Export data

Share this page

Bookmark this page

Ensembl Plants is produced in collaboration with Gramene

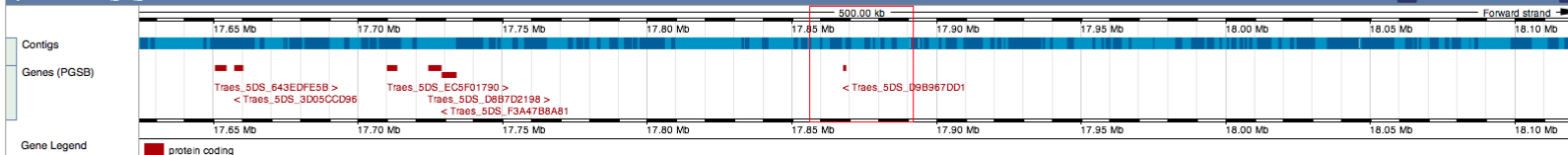
### Chromosome 5D: 17,856,312-17,892,564



#### Non-standard assembly

This assembly comprises sets of contigs co-located by linkage with markers on physical and/or genetic maps. Internally, each set (or "bin") is arbitrarily ordered. The scope of each bin is indicated in a track in the panel below. [More information](#)

### Region in detail

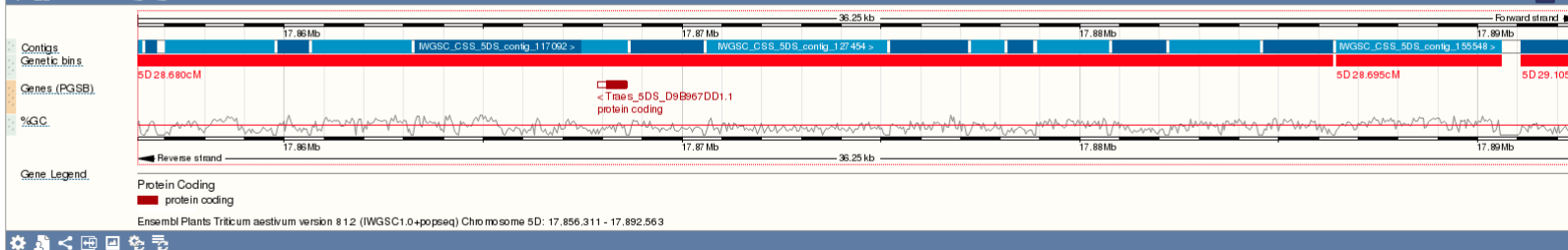


Location: 5D:17856311-17892563

Go

Gene:

Go



# The bread wheat genome: a challenge at many levels

- Large numbers of contigs create computational problems (most of our data processing pipelines have not been constructed to deal with such large numbers)
- A very large number of transcript models have been predicted from some genes in the IWGSC assembly (up to 78 transcripts)

# The bread wheat genome: a challenge at many levels

- There are many types of variation in a hexaploid genome:
  - Classical polymorphisms (which may be heterozygous or homozygous in sequenced individuals)
  - Inter-homoeologous variants (which may not necessarily be polymorphic)
- Functional annotation and gene naming are problems
- The likely continuation of the rapid evolution of the “best” wheat genome sequence over the next few years
  - And the need to accommodate hybrid assemblies

# The reference genome is imperfect, but it is improving

- 454 sequencing by Brenchley *et al.*, 2012
- Contigs from the IWGSC Chromosome Survey Sequence
  - Genome annotation (gene models) produced by PGSB group at the Helmholtz Institute, Munich
- Approximate co-location of contigs through the use of a high-density genetic map approach (POPSEQ) (Chapman *et al.*, Genome Biology 2015)
- Contigs and assembly for chromosome 3B produced using a BAC by BAC approach (Choulet *et al.*, Science, 2014).
  - Genome annotation (for 3B only) produced by the same

project





**12<sup>th</sup> International Wheat Genetics Symposium**  
**8-13 September 2013**  
**Yokohama, Japan**

**Summary Table 1. Symbols with ‘known function’**

The term set(s) indicates that the loci have been grouped into one or (more than one) orthologous (=‘homoeologous’) sets.

Symbol		Trait
<i>Aadh-1,2</i>	sets	Aromatic alcohol dehydrogenase
<i>a-Amy1,2</i>	sets	Alpha-amylase
<i>Aco-1,2</i>	sets	Aconitase
<i>AcpH-1</i>	sets	Acid phosphatase
<i>Adh-1</i>	sets	Alcohol dehydrogenase (Aliphatic)
<i>Adk-1</i>	sets	Adenylate kinase

**Summary Table 2. (Cont.) :** Chromosomal locations of wheat genes that are known to be members of orthologous sets of Triticeae genes.

GENOME A		GENOME B		GENOME D	
Chromosome		Chromosome		Chromosome	
Arm	Gene	Arm	Gene	Arm	Gene
	<i>Gli-A2</i>		<i>Gli-B2</i>		<i>Gli-D2</i>
	<i>Got-A1</i>		<i>Got-B1</i>		<i>Got-D1</i>
			<i>Gpc-B1b</i>		
			<i>Nor-B2</i>		
6AL	<i>Aadh-A2</i>	6BL	<i>Aadh-B2</i>	6DL	<i>Aadh-D2</i>
	<i>a-Amy-A1</i>		<i>a-Amy-B1</i>		<i>a-Amy-D1</i>
	<i>Aco-A1</i>		<i>Aco-B1</i>		<i>Aco-D1</i>
	<i>AhasL-A1</i>		<i>AhasL-B1</i>		<i>AhasL-D1</i>
	<i>Dip-A1</i>		<i>Dip-B1</i>		<i>Dip-D1</i>
	<i>Est-A4</i>		<i>Est-B4</i>		<i>Est-D4</i>
	<i>Got-A2</i>		<i>Got-B2</i>		<i>Got-D2</i>

# What do the genes look like?

	Ensembl Plants (average)	Bread wheat	Barley
Gene count	33037 (STD 9915)	33645 (average per genome)	24,286
Average exon count	5 (STD 1.4)	4	5
Average gene length	3162 (STD 1038)	2557	3006

# Functional annotation and gene naming

- Official wheat gene names are annotated to genetic markers, not to genomic sequence: hard to utilise this data
- Normal automatic annotation by InterProScan and GO interference provides basic functional annotation

# Cereal genomes: conserved families

Species	Gene count	% of 50% conserved families present	% of 90% conserved families present
<i>Homo sapiens</i>	20,805	99.3	99.92
<i>Hordeum vulgare</i>	24,211	95.41	98.94
<i>Brachypodium distachyon</i>	26,552	98.66	100
<i>Arabidopsis thaliana</i>	27,416	96.37	100
<i>Triticum urartu</i>	33,424	90.15	97.44
<i>Aegilops tauschii</i>	33,849	89.55	97.44
<i>Sorghum bicolor</i>	34,496	99.10	99.89
<i>Setaria italica</i>	35,471	97.57	99.39
<i>Oryza sativa</i>	35,679	98.28	99.94
<i>Zea mays</i>	39,479	98	99.44
<i>Triticum aestivum</i>	98,799	98.36	99.94

# Cereal genomes: functional annotation

Species	InterPro Coverage, %	GO coverage in 3 domains, %	“Deep” GO coverage in 3 domains, %
<i>Homo sapiens</i>	88.71	69.53	46.48
<i>Brachypodium distachyon</i>	86.44	20.75	11.54
<i>Arabidopsis thaliana</i>	85.26	42.72	21.27
<i>Hordeum vulgare</i>	84.02	23.2	12.6
<i>Triticum aestivum</i>	83.83	10.21	5.01
<i>Sorghum bicolor</i>	88.22	23.86	12.45
<i>Aegilops tauschii</i>	82	15.89	9.02
<i>Triticum urartu</i>	75.41	15.30	8.62
<i>Setaria italica</i>	73.86	19.92	10.31
<i>Zea mays</i>	73.55	19.69	10.47
<i>Oryza sativa</i>	67.98	22.61	12

# Defining homoeology

- Whole genome alignments have been performed between the three wheat genomes
- The Ensembl Gene Trees pipeline, which uses protein sequence and species history to infer an evolutionary history for each gene family, is run treating each of the three wheat genomes as a separate species
- Orthoeologous genes are inferred using the gene trees, and used to link to the corresponding DNA-based alignments

# Use orthology calls to determine inter-homoeologous variants

- Find differences between A, B and D (can only do this where a 1:1 relationship exists (synteny is unknown in the absence of chromosome-level assemblies))
- Can parse single nucleotide and other changes out of gaps/mismatches in alignments
- Calling all differences between the longest alignment connecting each pair of 1:1 homoeologues
  - including extensions beyond the genic regions

- Gene-based displays
  - Summary
  - Splice variants
  - Transcript comparison
  - Supporting evidence
  - Gene alleles
  - Sequence
    - Secondary Structure
  - External references
  - Regulation
  - Literature
  - Ontology
    - GO: molecular function
  - Plant Compara
    - Genomic alignments
    - Gene tree
    - Gene gain/loss tree
    - Orthologues
    - Paralogues
  - Homoeologues**
  - Pan-taxonomic Compara
    - Gene Tree
    - Orthologues
  - Phenotype
    - Genetic Variation
      - Variation table
      - Structural variation
      - Variation image
  - External data
    - Gene expression
    - Personal annotation
  - ID History
    - Gene history

## Gene: Traes\_4AL\_676E30BB2

**Location** [Scaffold IWGSC\\_CSS\\_4AL\\_scaff\\_7111379:2,045-11,868](#) reverse strand.  
**About this gene** This gene has 1 transcript ([splice variant](#)), [41 orthologues](#) and [3 paralogues](#).  
**Transcripts** [Hide transcript table](#)

Show/hide columns (1 hidden)		Filter				
Name	Transcript ID	bp	Protein	Biotype	UniProt	Flags
Novel	<a href="#">Traes_4AL_676E30BB2.1</a>	1956	<a href="#">608aa</a>	Protein coding	<a href="#">WSDPW1</a>	

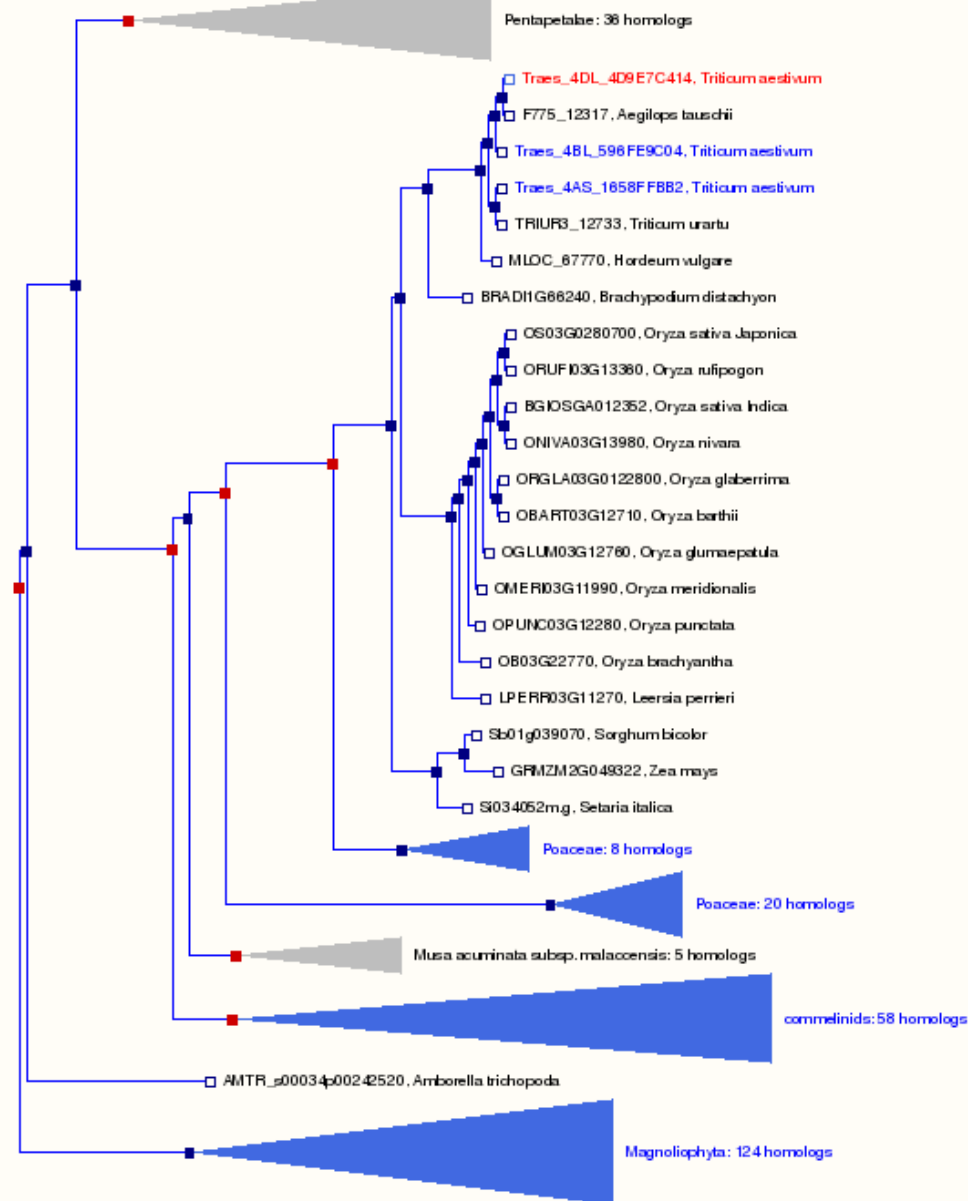
## Homoeologues

[View protein alignments of all homoeologues](#) | [View genomic alignments of all homoeologues](#) | [Download all protein sequences](#) | [Download all DNA sequences](#)

Show/hide columns		Filter					
Species	Type	dN/dS	Ensembl Identifier & gene name	Compare	Location	Target %Id	Query %Id
Triticum aestivum	1-to-1	0.35294	<a href="#">Traes_4BS_6DA9858E8</a> No description	<ul style="list-style-type: none"> <li>• Region Comparison</li> <li>• Alignment (protein)</li> <li>• Alignment (cDNA)</li> <li>• Gene Tree (image)</li> </ul>	<a href="#">4B:113369236-113372168:-1</a>	93	43
Triticum aestivum	1-to-1	1.02778	<a href="#">Traes_4DL_B10644C53</a> No description	<ul style="list-style-type: none"> <li>• Region Comparison</li> <li>• Alignment (protein)</li> <li>• Alignment (cDNA)</li> <li>• Gene Tree (image)</li> </ul>	<a href="#">4D:73931761-73938238:-1</a>	82	81

- Configure this page
  - Manage your data
  - Export data
  - Share this page
  - Bookmark this page
- [Ensembl Plants is produced in collaboration with Gramene](#)





#### LEGEND

Branch Length

- x1 branch length
- - - x10 branch length
- - - x100 branch length

Genes

- Gene ID gene of interest
- Gene ID with in-sp. paralog

Nodes

- gene node
- speciation node
- duplication node
- ambiguous node
- gene split event

Collapsed Nodes

- ▲ collapsed sub-tree
- ▲ collapsed (paralog)
- ▲ collapsed (gene of interest)

Collapsed Alignments

- 0 - 33% Aligned AA
- 33 - 66% Aligned AA
- 66 - 100% Aligned AA

Expanded Alignments

- Gap
- Aligned AA

**Gene-based displays**

- Summary
- Splice variants
- Transcript comparison
- Supporting evidence
- Gene alleles
- Sequence
  - Secondary Structure
- External references
- Regulation
- Literature
- Ontology
- Plant Compara
  - Genomic alignments
  - Gene tree
  - Gene gain/loss tree
  - Orthologues
  - Paralogues
  - Homoeologues
- Pan-taxonomic Compara
  - Gene Tree
  - Orthologues
- Phenotype
- Genetic Variation
  - Variation table
  - Structural variation
  - Variation image
- External data
  - Gene expression
  - Personal annotation
- ID History
  - Gene history

- Configure this page
- Add your data
- Export data
- Share this page
- Bookmark this page

## Gene: Traes\_4DL\_4D9E7C414

**Location** [Chromosome 4D: 79,406,465-79,412,011](#) forward strand.

**About this gene** This gene has 1 transcript ([splice variant](#)), [62 orthologues](#) and [4 paralogues](#).

**Transcripts**

[Hide transcript table](#)

Show/hide columns (1 hidden)		Filter				
Name	Transcript ID	bp	Protein	Biotype	UniProt	Flags
Novel	<a href="#">Traes_4DL_4D9E7C414.1</a>	3442	<a href="#">1009aa</a>	Protein coding	<a href="#">W5EHC7</a>	

## Homoeologue alignment

[Download homology](#)

**Type: 1-to-1 homoeologues**

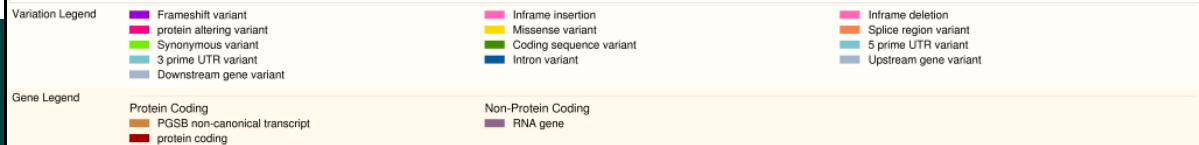
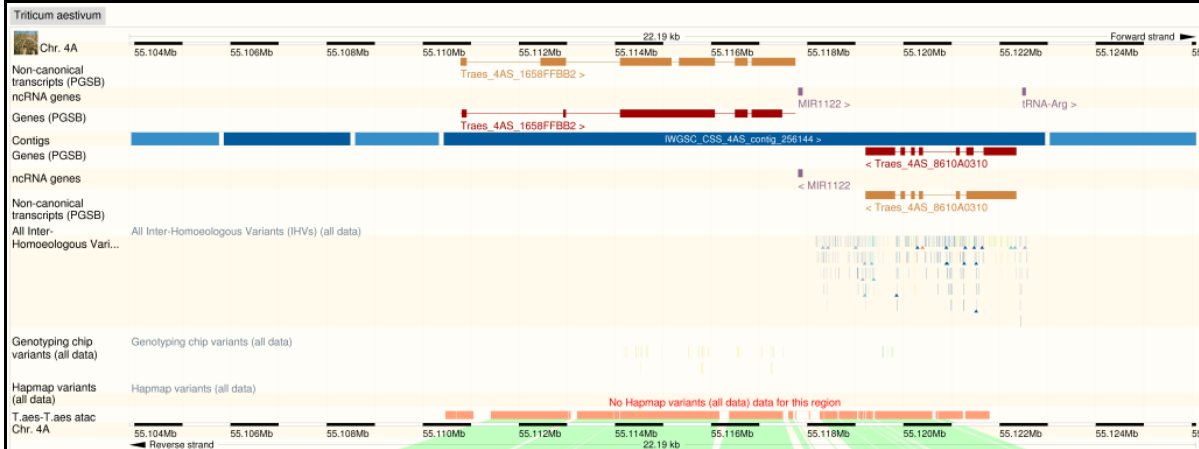
Species	Gene ID	Peptide ID
Triticum aestivum	Traes_4DL_4D9E7C414	Traes_4DL_4D9E7C414.1
Triticum aestivum	<a href="#">Traes_4AS_1658FFBB2</a>	<a href="#">Traes_4AS_1658FFBB2.1</a>

Alignment details			
Alignment length	1015	gaps	11
identical residues	958	similar residues	972

CLUSTAL W(1.81) multiple sequence alignment

```

Traes_4DL_4D9E7C414.1/1-1009  -----MKFLLLAMISYCFQCQFSLLYKLFVTVSFLFKPGAGSSRKQASNSGILSHTLLDK
Traes_4AS_1658FFBB2.1/1-1010 MASVGRSRSRRRRGEVGGPFWDAAAPSGDYSA----DHHGAGSSRKQASNSGILSHTLDFE
                                :   . : . * : :   . :   : *****:
Traes_4DL_4D9E7C414.1/1-1009 EIRKSKPRQSSCVPMKKLIDEEFKSDVNRHTSPGAVGRLMGLDLSLPTSSGTHSQHRSSR
Traes_4AS_1658FFBB2.1/1-1010 EIRKSKPRQSSCVPMKKLIDEEFKSDVNRHTSPGAVGRLMGLDLSLPTSSGTHSQHRSSR
*****
    
```



# The “best” wheat genome is likely to change further in future

- “Lift over” functionality can re-map features upon re-scaffolding of existing contigs
  - New primary sequence is still problematic, will require complete re-running of all alignment programs

# Ensembl Assembly Converter Tool

## Tools

We provide a number of ready-made tools for processing your data. At the moment, small datasets can be uploaded to our servers and processed online; for larger datasets, we provide an API script that can be downloaded (you will also need to [install our Perl API](#) to use these).

In the near future we aim to offer an intermediate service, whereby medium-to-large data sets can be submitted to a queue, similar to BLAST.

Currently available:

Name	Description	Online tool	Download code	Documentation
<a href="#">Variant Effect Predictor</a>				
<a href="#">BLAST/BLAT</a>				
<a href="#">BioMart</a>				
<a href="#">Assembly converter</a>				
<a href="#">ID History converter</a>				
<a href="#">Ensembl Genomes Virtual</a>				
<a href="#">Ensembl Genomes REST</a>				
<a href="#">Ensembl Plants release 28 - A</a>				

### Personal Data

- [Login](#)
- [Register](#)
- [Lost Password](#)
- Custom Data
  - [Add your data](#)
  - [Attach DAS](#)
  - [Manage Data](#)
  - [Features on Karyotype](#)
- Manage Configurations
  - [Configurations for this page](#)
  - [All configurations](#)
  - [Configuration sets](#)
- Online Tools
  - [Variant Effect Predictor](#)
  - [Assembly Converter](#)
  - [ID History Converter](#)
  - [Data Slicer](#)
- [Help](#)

### Tips

Map your data to the current assembly. The tool accepts a [list of simple coordinates](#), or files in these formats: [GFF](#), [GTF](#), [BED](#), [PSL](#).

N.B. Export is currently in GFF only

For large data sets, you may find it more efficient to use our [ready-made converter script](#).

Species:

Assembly/coordinates to convert:

Upload file

Name for this data (optional):

Paste data:

Upload file:  no file selected

or provide file URL:

# Additional alignment data for bread wheat

- Repbase repeats
- Triticeae repeats from TREP
- Wheat RNA-Seq, ESTs, and UniGene datasets have been aligned to the *Triticum aestivum* genome:
- 454 RNA-seq data for the following INSDC studies: SRP02455 (Akhunova *et al.*), ERP001415 (Brenchley *et al.*), SRP004502
- Sequences from TriFLDB
- **Transcriptome assembly from diploid einkorn wheat *Triticum monococcum* (Fox *et al.*)**

# Polymorphism data for bread wheat

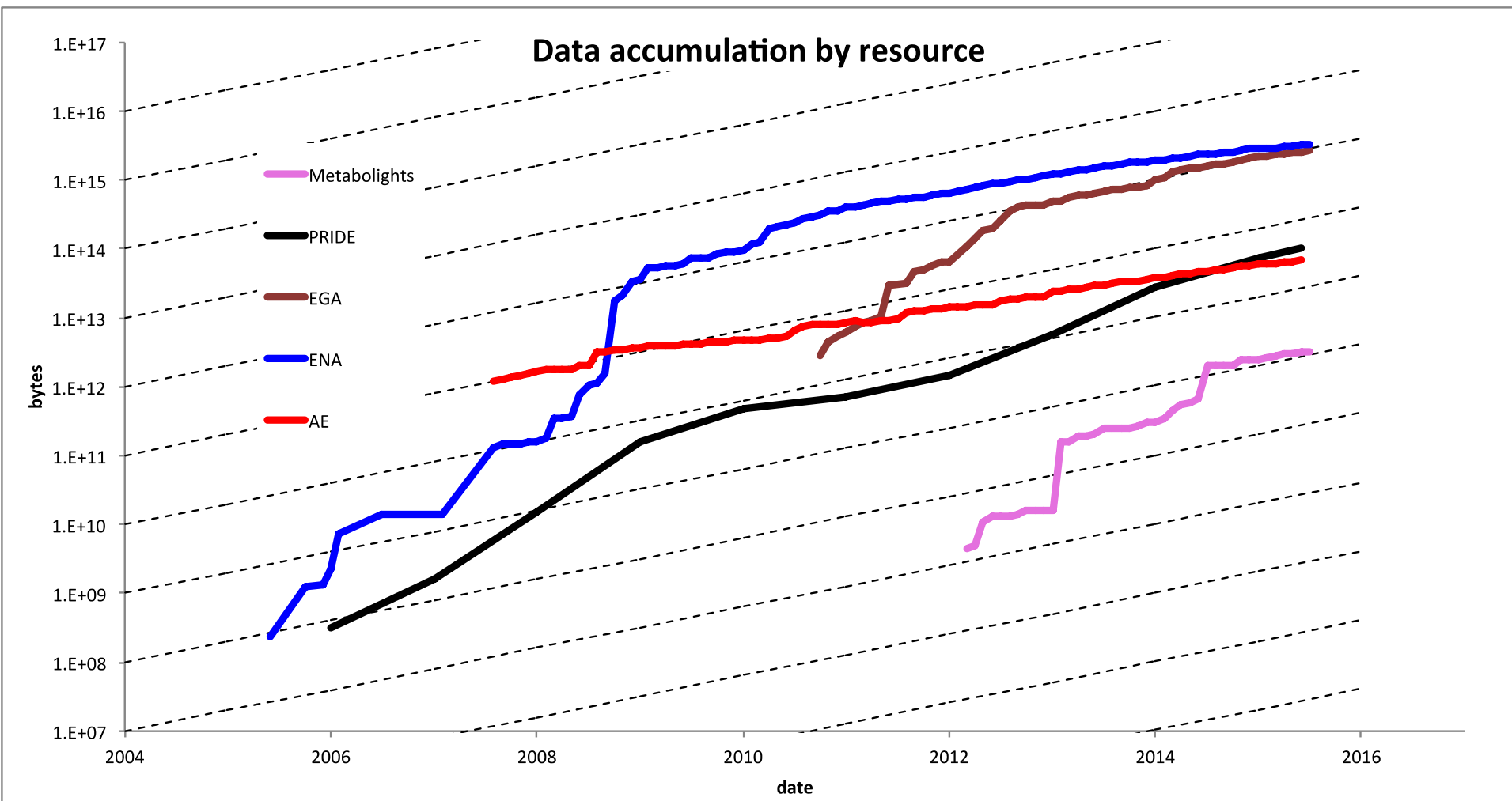
- ~900,000 SNPs provided by CerealsDB, as follows:
  - The Axiom 820K SNP Array contains 820,000 SNPs of which ~684,000 have been mapped.
  - The iSelect 80K Array contains over 80,000 SNP loci of which ~58,000 have been mapped.
  - The KASP probeset contains ~3,900 SNP loci of which ~3,100 have been loaded in Ensembl Plants
- The Wheat HapMap project data set containing 1.57 million SNPs and 161,719 small indels

# Bread wheat whole genome alignment

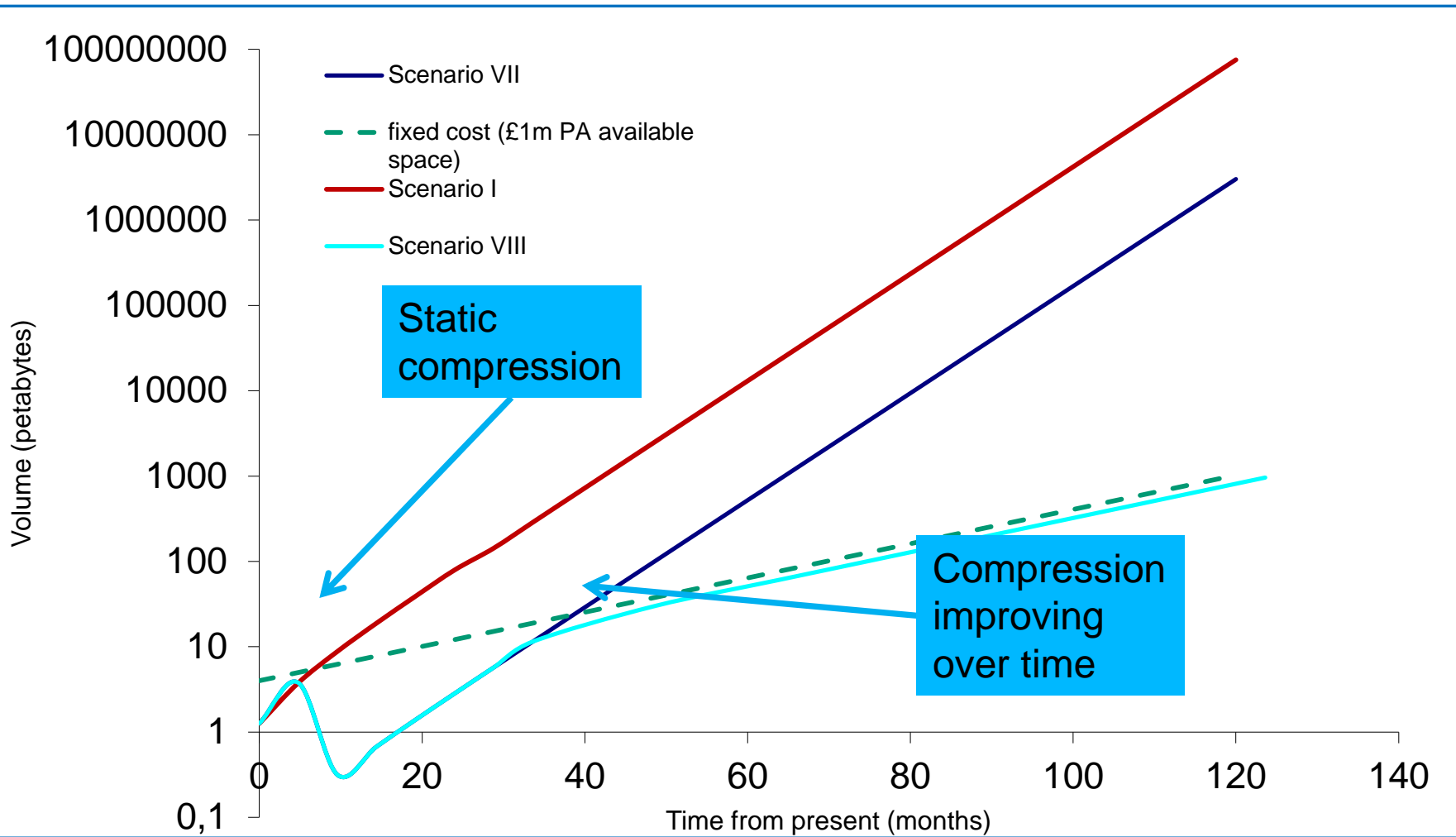
- DNA-DNA pairwise alignments with lastZ
  - *Brachypodium distachyon*: 617,996,145 Mb (14% of bread wheat) in 1,310,922 blocks
  - *Hordeum vulgare*: 423,284,874 Mb (9% of bread wheat) in 2,902,234 blocks
  - *Oryza sativa* Japonica: 312,857,683 Mb out of 4,460,951,632 (7% of bread wheat) in 718,036 blocks
- Additional alignments to the precursor genomes *A. tauschii* and *T. uratu*.



# Growing data



# Compression options



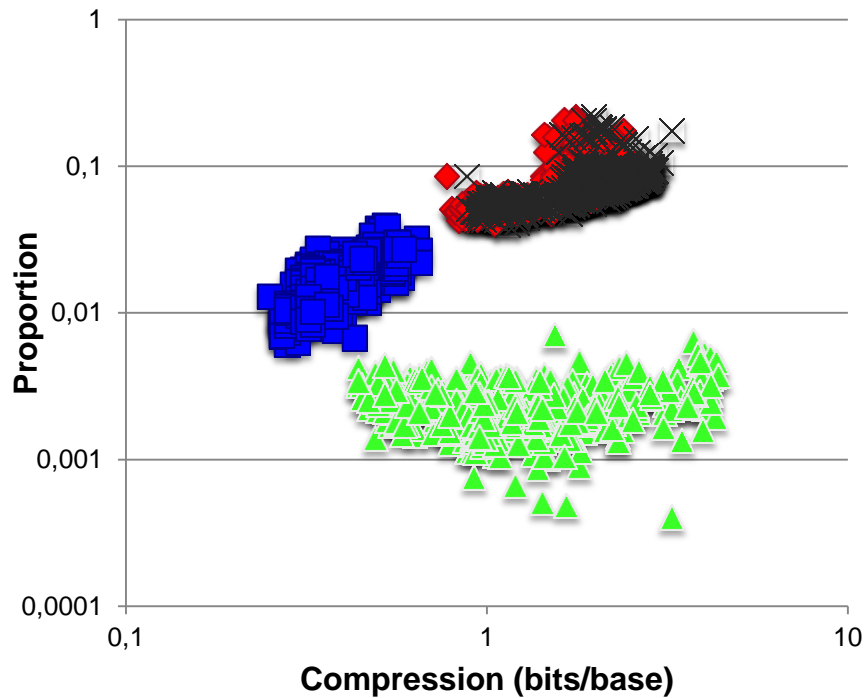
# Reference-based compression

- Assemble and map if no reference exists
- 0.02-0.66 bits/base pair (bzip: 1 bit/base pair)
- Controlled loss of precision: score quality scores at variant locations and elsewhere according to a user-set “quality budget”
- Increase in performance as read length/knowledge of sequence space improves
- Makes continued universal archiving at fixed disc cost possible
- Main cost is staff, not disc

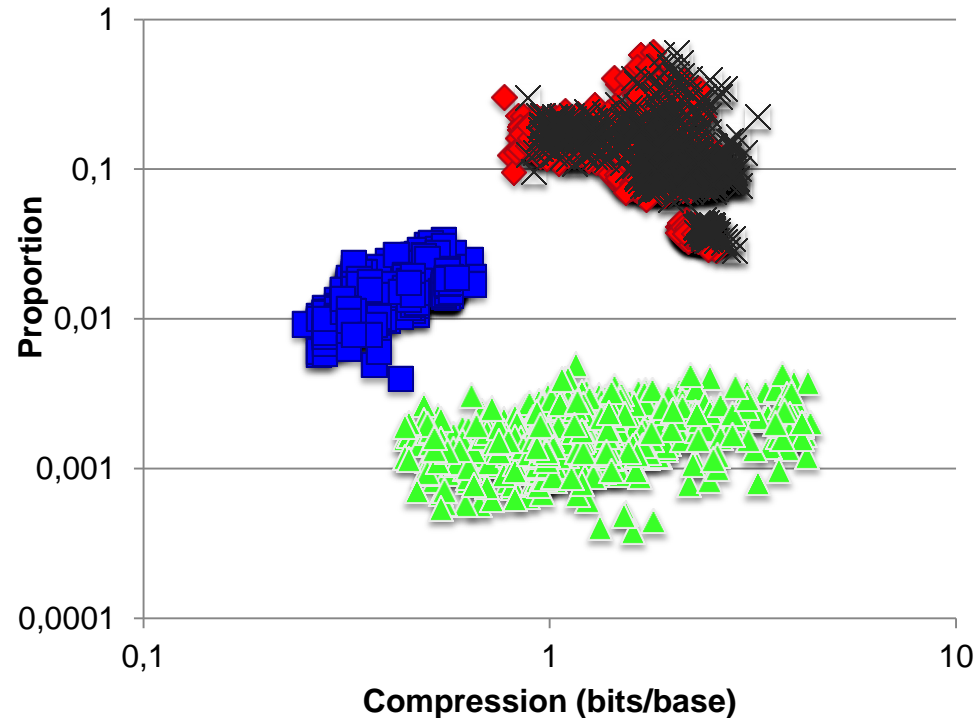
# Lossy models for per-base quality compression

- ◆ quantisation 4-level
- substitutions and insertions
- ▲ all
- × quantisation 8-level

## False negative



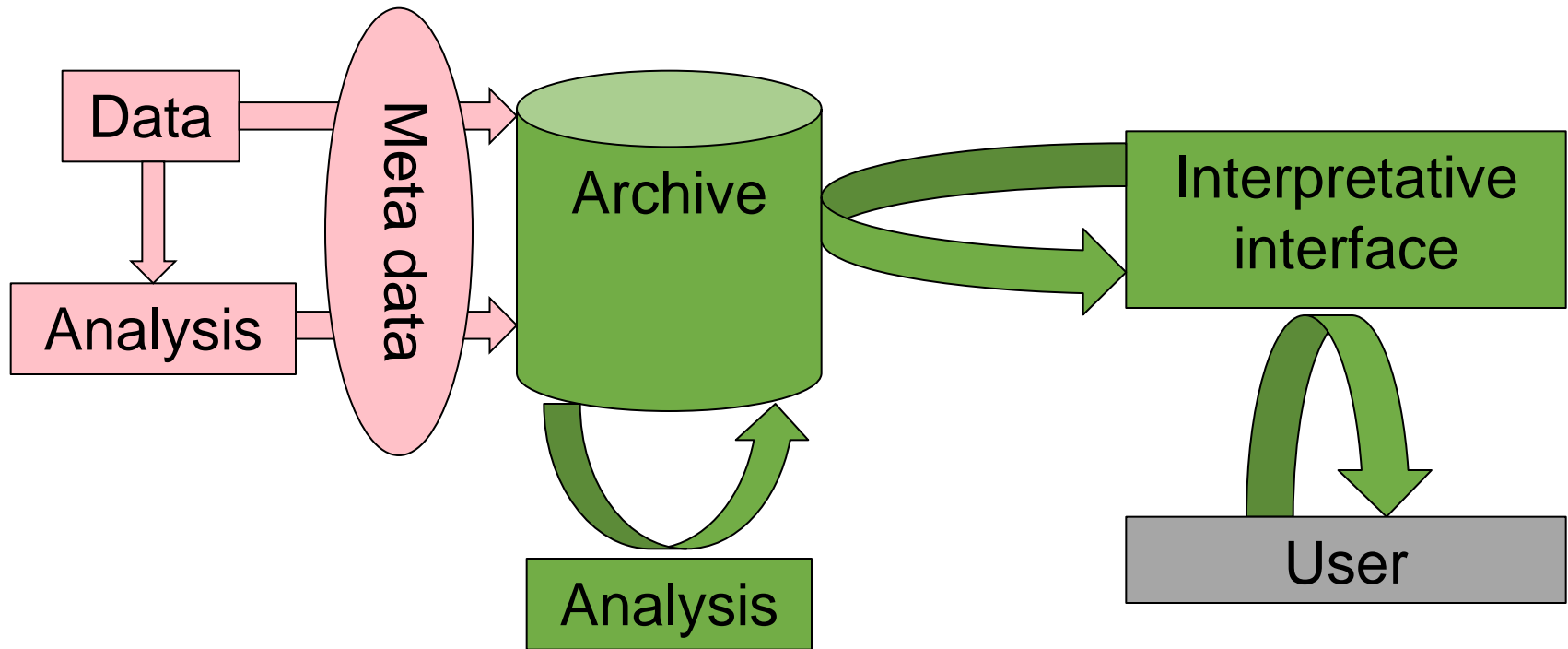
## False positive



# A T-shaped model

- Deep coverage of certain model species or other species where we are working closely in collaboration with the community
  - e.g. *Schizosaccharomyces pombe*
- Broad coverage (all genome sequences submitted to the archival databases of the International Nucleotide Sequence Database Consortium (ENA, GenBank, DDBJ))
  - Ensembl provides a view on the archives
  - Expansion to all submitted bacteria (2012)
  - Expansion to all submitted protists and fungi (2015)

# A Scalable Model for Archival and Integrative Services



# BioSamples - database of sample descriptions

The BioSamples database aggregates sample information for reference samples (e.g. Coriell Cell lines) and samples for which data exist in one of the EBI's assay databases such as [ArrayExpress](#), the [European Nucleotide Archive](#) or [PRIDE](#). It provides links to assays an specific samples, and accepts direct submissions of sample information.

## Info

- [Help pages](#) about how to search BioSamples, how to submit data, and FAQ.
- [Programmatic access](#) to query and download data using web services.
- [Latest news](#) about BioSamples database.
- The BioSamples database now offers access to [RDF](#), and a supporting [SPARQL endpoint](#) as part of the [EBI's RDF platform](#).
- Contact us by emailing [biosamples@ebi.ac.uk](mailto:biosamples@ebi.ac.uk)

## External links

- [Experimental Factor Ontology](#)
- [Human Induced Pluripotent Stem Cells Initiative \(HipSci\)](#)
- [1000 Genomes](#)
- [Encyclopedia of DNA Elements \(ENCODE\)](#)
- [Catalogue Of Somatic Mutations In Cancer \(COSMIC\)](#)

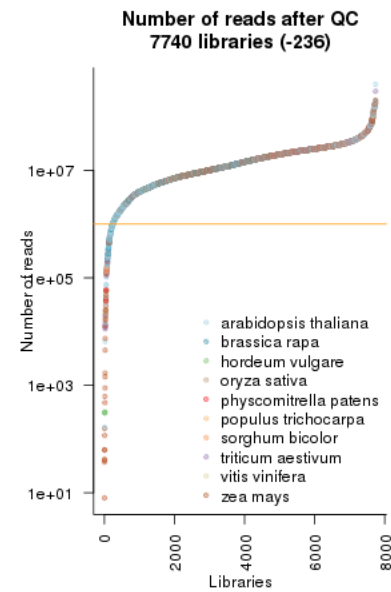
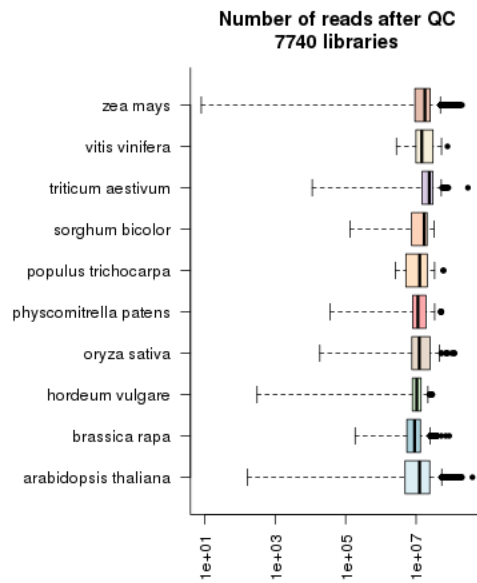
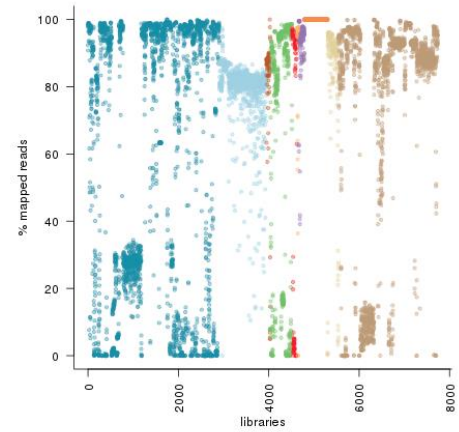
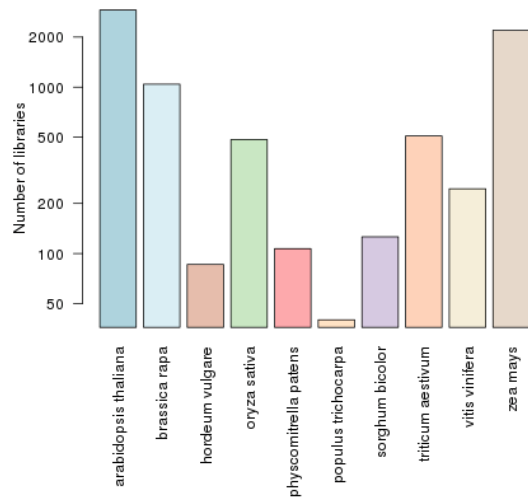
## Data Content

- 4,296,238 Samples
- 51,359 Groups









## NEWS HIGHLIGHTS

News from the transPLANT project and partners. See also the [transPLANT newsletter](#).

### Mining Plant Variation Data

Registration now closed!



Workshop on plant variation data, re-sequencing projects and GWAS analyses within transPLANT, hosted at The European Bioinformatics Institute, Cambridge, UK from 1st July to 3rd July. [Click here to apply](#).

### BSC releases COMPSs version 1.2

The Grid Computing and Clusters group of Barcelona Supercomputing Center is proud to announce the release of COMPSs version 1.2. A framework for easily implement distributed applications.



COMPSs

### Filling the gap between sequence and function; a bioinformatics approach

This thesis focuses on deriving function from

## The transPLANT variation archive

Store, accession and update plant variation data. Now accepting submissions of the materials.



## transPLANT

is a European-Union funded e-infrastructure to support computational analysis of genomic data from crop and model plants. The project funds coordination and research activities; and provides free access to tools, training and data standards.

You can use this site to:

- Find content from ten different plant genomics databases in a single click e.g. [search for "rubisco"](#).
- Find out what genomic databases are available for your species of interest in our genome resource registry e.g. [look for "Arabidopsis" here](#).
- Archive, accession and update plant variation data: [find out more](#)

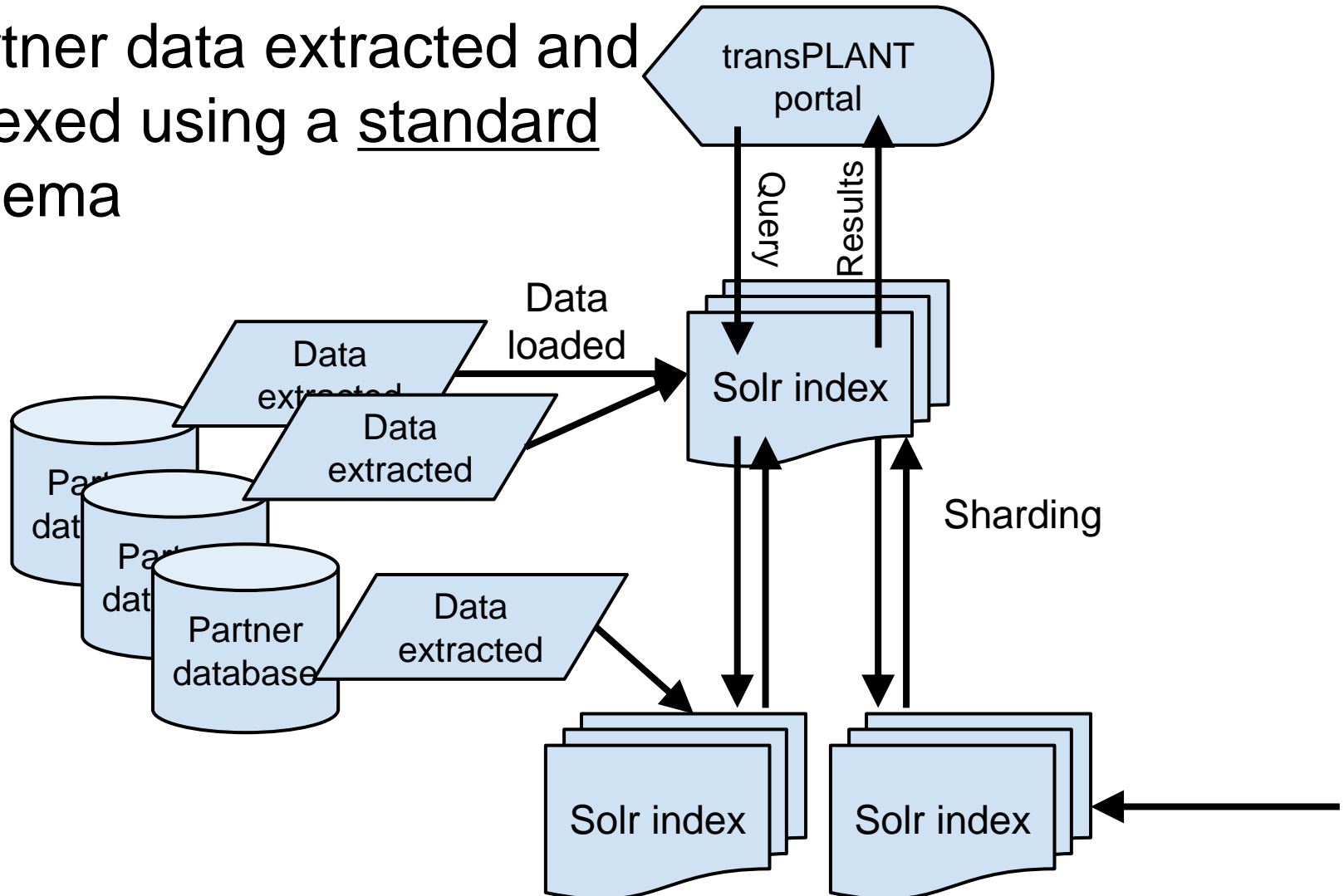
# The transPLANT distributed search



- Search backend implemented using
  - A fast, free, powerful and widely adopted search server based on Lucene search library.
  - Scalable - Allows distributed search
  - Standards Based Open Interfaces
    - XML and HTTP, hence, many clients
- Gives us
  - Distributed search over partner resources via sharding
  - Allows faceted search for rapid resource discovery
  - Integrates with the Drupal website

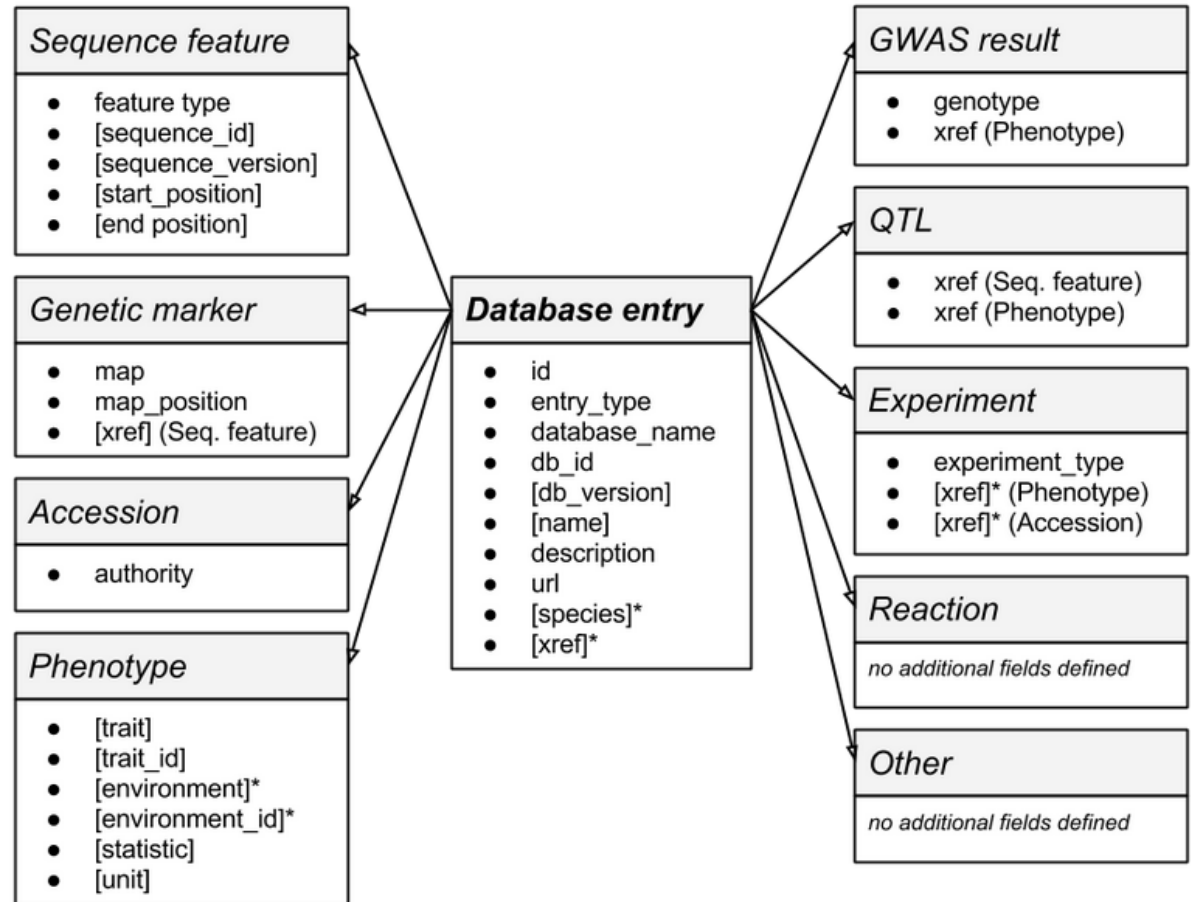
# Implementation

- Partner data extracted and indexed using a standard schema



# Data model

- A conceptual schema for indexing data across partner databases for the purpose of search.
- Designed to be simple enough to accommodate results from different and varied sources, yet detailed enough to provide meaningful free text search and subsequent exploration of results via faceting, filtering and rich-snippets.



# Implementation

Partner	Database name	Database description	Number of entries	Number of species	Entry types	Local or remote?
EBI	Ensembl Plants	A genome-centric portal for plant species of scientific interest	1,523,622	38	Sequence feature (protein_coding, ncRNA)	Local
PAS	PolapgenDB		31	Hordeum vulgare	Phenotype, Accession, Experiment, Genetic marker, Sequence feature	Remote
MIPS	PlantsDB	Tools for Comparative analysis of plant genomes and transcriptome data.	263,401	6	Sequence feature (transcript)	Local
MIPS	Crows Nest	A comparative map viewer for genome-wide chromosome organization and synteny between two or more plant genomes.	140,138	4	Sequence feature	Local
IPK	OPTIMAS-DW	A comprehensive transcriptomics, metabolomics, ionomics, proteomics and phenomics data resource for maize.	33,919	Zea mays	Sequence feature	Local
IPK	MetaCrop	Diverse information about metabolic pathways in crop plants and the creation of detailed metabolic models.	586	>50	Reaction	Local
IPK	GBIS	GBIS/I allows to retrieve information from the German federal ex situ collection.	148,696	>50	Accession	Local
IPK	CR-EST	Access to sequence, classification, clustering, and annotation data of crop EST projects at the IPK-Gatersleben.	218,927	6	Sequence feature (EST)	Local
INRA	GnpIS	GnpIS is a multispecies integrative information system dedicated to plant and fungi pests.	27,366	>50	Accession, Experiment	Remote
GMI	GWAPortal	GWAPortal is a resource for phenotypes and GWAS studies in plants.	828	Arabidopsis thaliana	GWAS, Experiment, Phenotype	Remote
6	10		2,357,514			

Providing tools, training and standards for PLANT genomic science

transPLANT search

Resource registry

Site search

## CURRENT SEARCH

Found 37 results

## FILTER BY DATABASE:

- Ensembl Plants (16)
- PlantsDB (13)
- OPTIMAS-DW (8)

## FILTER BY DATA TYPE:

- Sequence feature (16)
- transcript (13)
- Sequence (8)

## FILTER BY SPECIES:

- Arabidopsis thaliana (11)
- Zea mays (8)
- Medicago truncatula (6)
- Brassica rapa subsp. pekinensis (4)
- Theobroma cacao (2)

Show more

Search » Search results » pad4

## Search



## Search results

### MTR\_8g094370

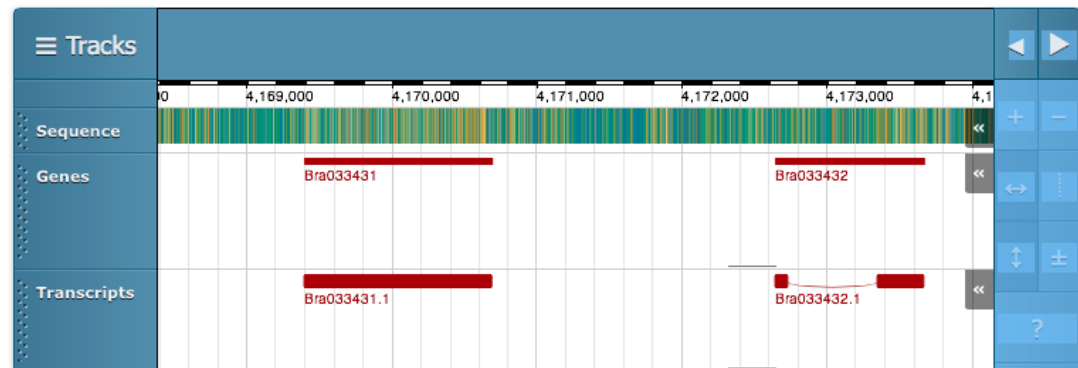
**Description** PAD4 [Source:UniProtKB/TrEMBL;Acc:G7LEZ0]  
**Species** Medicago truncatula  
**Location** 8:27218509-27222910

[View in Geniverse](#)

### Bra033432

**Description** AT3G52430 (E=1e-022) PAD4, ATPAD4 | PAD4 (PHYTOALEXIN DEFICIENT 4); lipase/ protein binding / triacylglycerol lipase  
**Species** Brassica rapa subsp. pekinensis  
**Location** A04:4172644-4173668

[View in Geniverse](#)



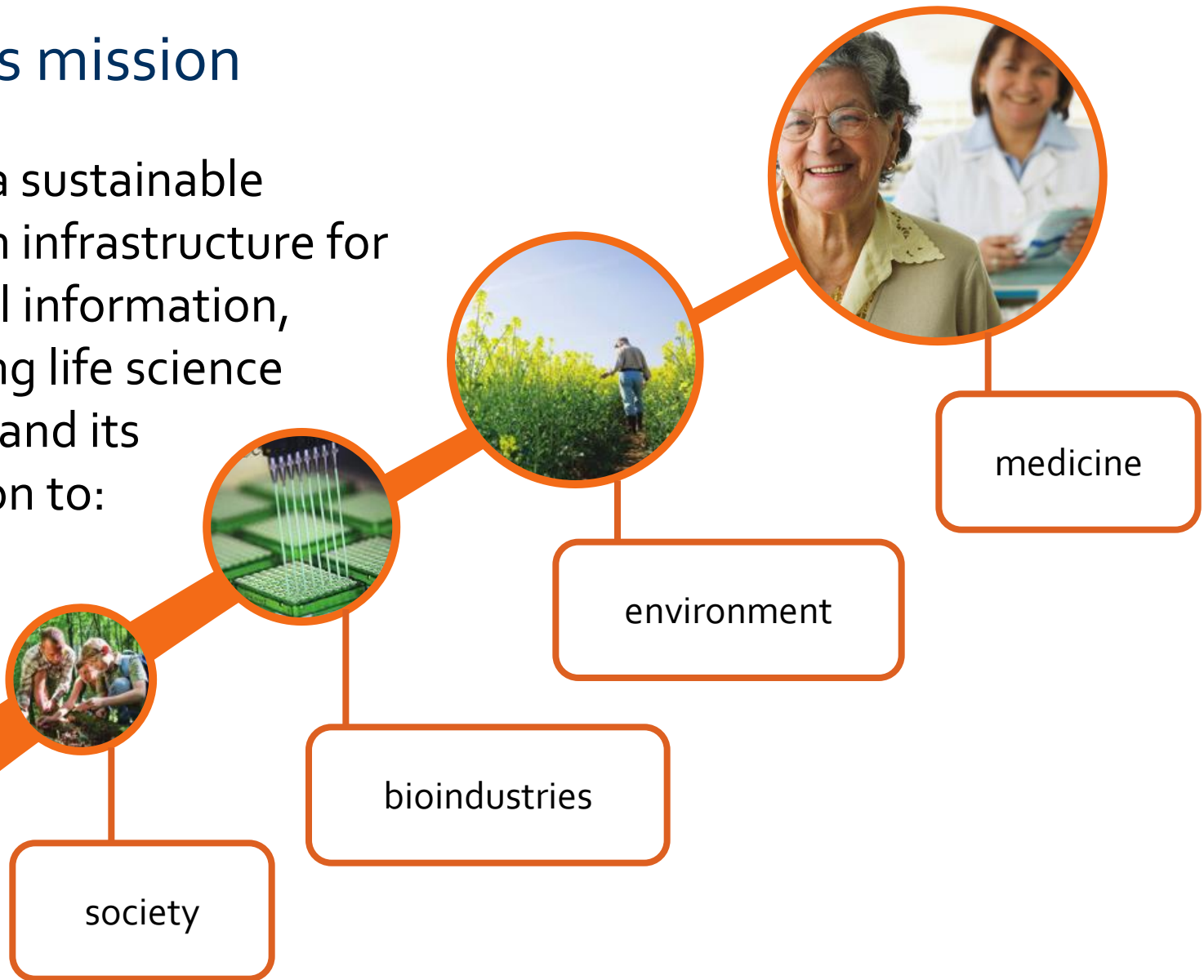
Powered by Geniverse

### Bra006922

**Description** AT3G52430 (E=8e-195) PAD4, ATPAD4 | PAD4 (PHYTOALEXIN DEFICIENT 4);

# ELIXIR's mission

To build a sustainable European infrastructure for biological information, supporting life science research and its translation to:





# ELIXIR Infrastructure = Nodes + ELIXIR Platforms

## Data

*Sustain core data resources*

## Tools

*Services & connectors to drive access and exploitation*

## Interoperability

*Integration and interoperability of data and services*

## Compute

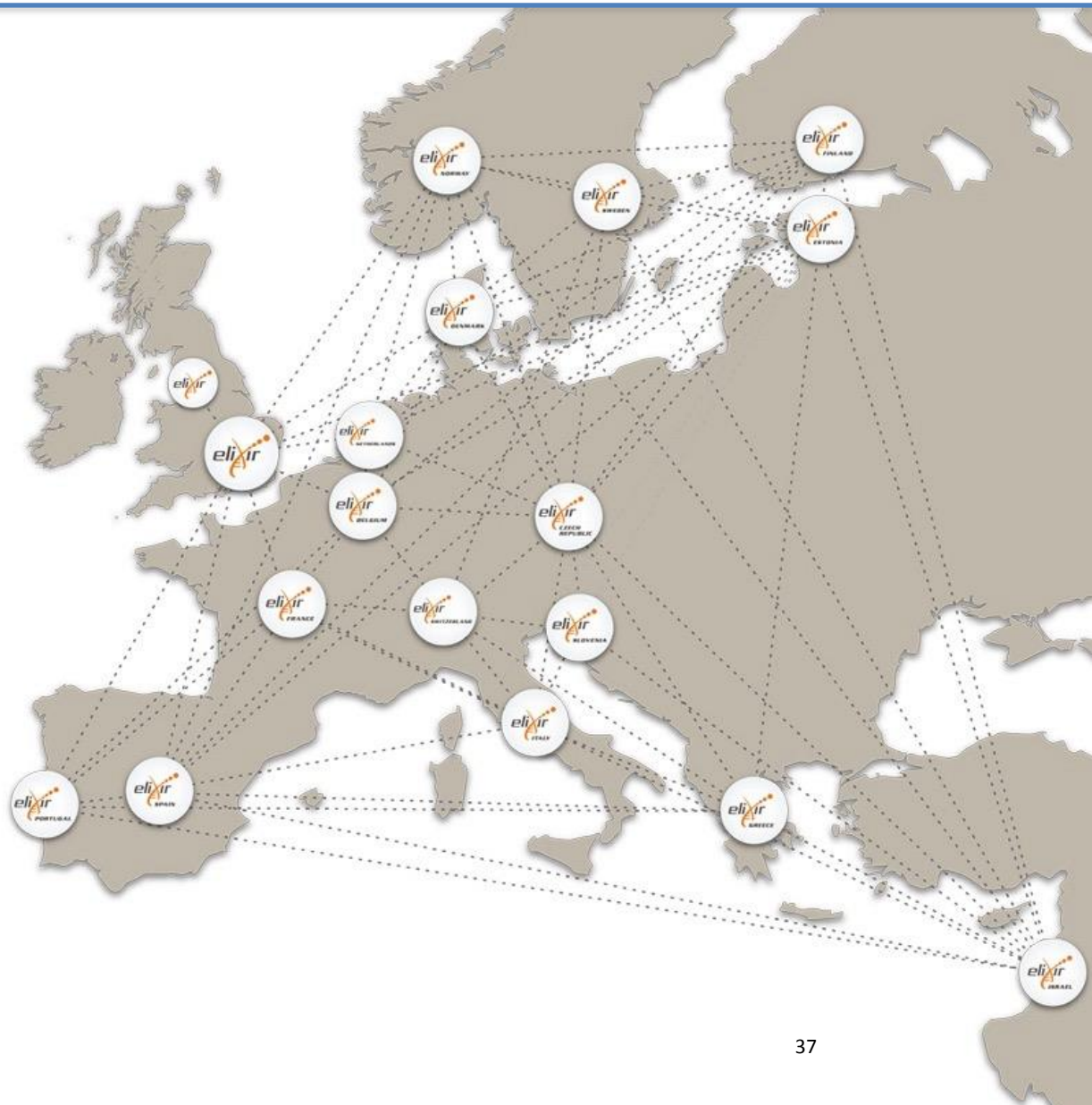
*Access, Exchange & Compute (incl. sensitive data)*

## Training

*Professional skills for managing and exploiting data*

## User Communities

*Infrastructure usage and impact*




## EXCELERATE



EXCELERATE funding will help ELIXIR coordinate and extend national and international data resources to ensure the delivery of world-leading life-science data services. It will support a pan-European training programme, anchored in national infrastructures, to increase bioinformatics capacity and competency. It will also provide efficiencies in management and operation throughout the infrastructure, which is distributed amongst 17 countries.

Dedicated use cases, defined in close partnership with diverse research communities, will help ELIXIR grow in keeping with the needs of scientists working on rare diseases, biomedical and human genomics, marine and plant sciences and other specialized areas. This will ensure that ELIXIR's services for data, tools, interoperability, compute, training and industry support are rooted in user needs, and can deliver benefits to existing and future research projects.

ELIXIR was invited to apply to a dedicated call within Horizon 2020 following the ESFRI and European Council decision in 2014 to categorize ELIXIR as one of [Europe's three priority new Research Infrastructures](#). ELIXIR-EXCELERATE represents ELIXIR's submission to this Call.

[Download EXCELERATE presentation.](#) 

# ELIXIR green use case: needs of the plant community

---

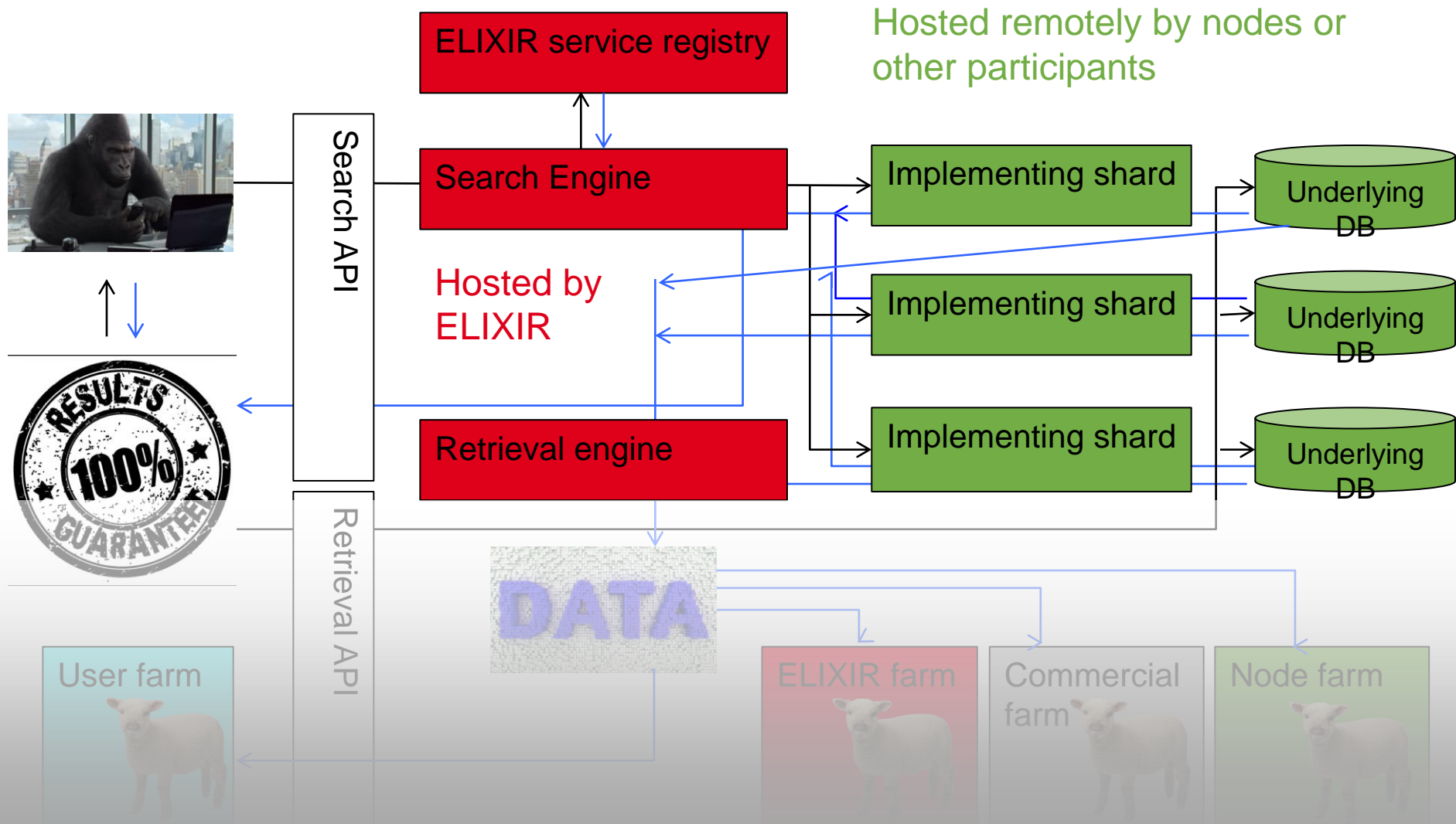
- For nucleotide sequence data, there are existing archives
  - Submission of nucleotide data to European Nucleotide Archive/GenBank/DDBJ, variation data to European Variation Archive/dbSNP
- Concentrate on filling resource gaps/improving “social” practices: phenotype data is a particular challenge
  - Heterogeneous
  - Poorly annotated
  - Dispersed
  - In production in increasing scale

# Plant genome-phenotype work in ELIXIR-EXCELERATE

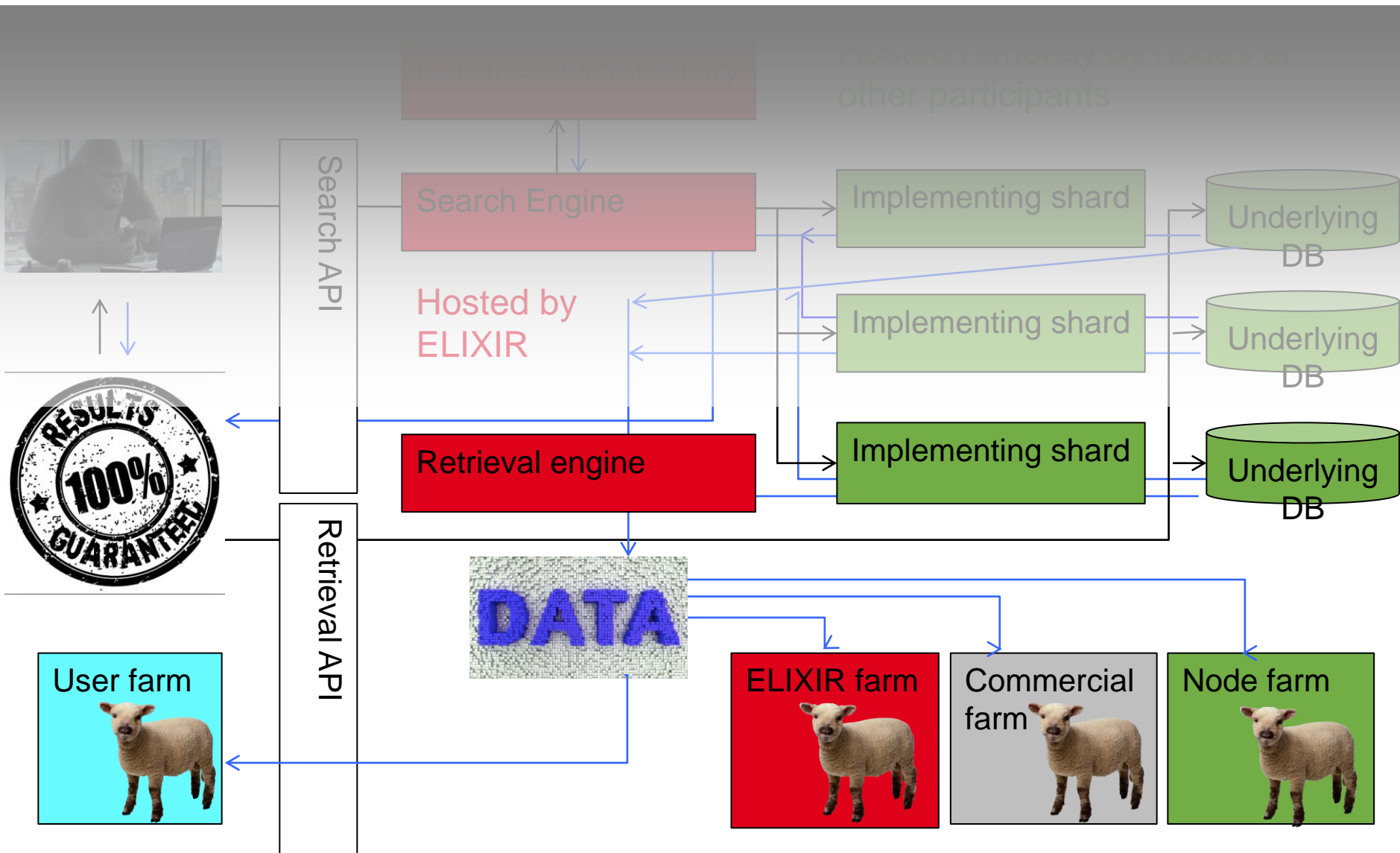
---

- Aims of work
  - Each partner will publish some phenotypic data sets emerging from national activities
  - Agreement on standards (minimal information, choice of ontologies, data formats)
  - Design and implementation of a common API
    - Initial focus will be on data discovery and retrieval
      - What data is available for this species?
      - What types of experimental data are available?
    - Comprehensive data modelling is a later step

# Workflow: Data Discovery



# Workflow (user data access)

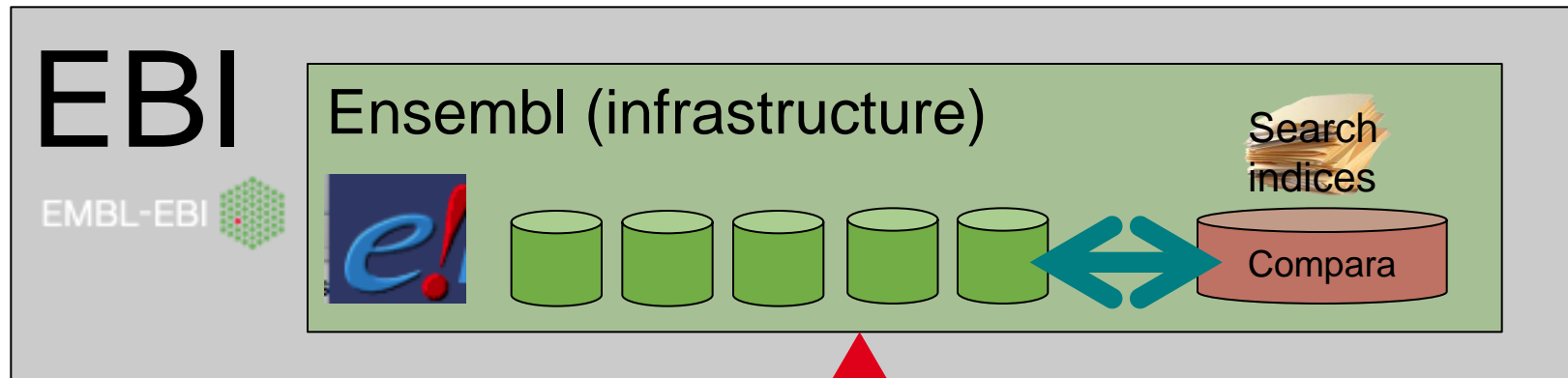


# Compara is changing

- New, HMM-based methodology will scale to building larger families
- One Compara from across the taxonomic space
- Compara + distributed search + ELIXIR = a new operational model for Ensembl databases

# A distributed future for Ensembl databases

---

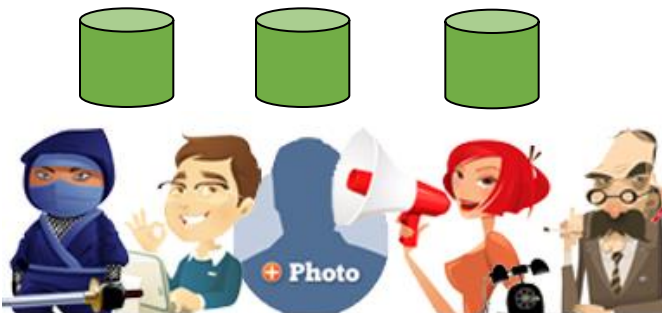
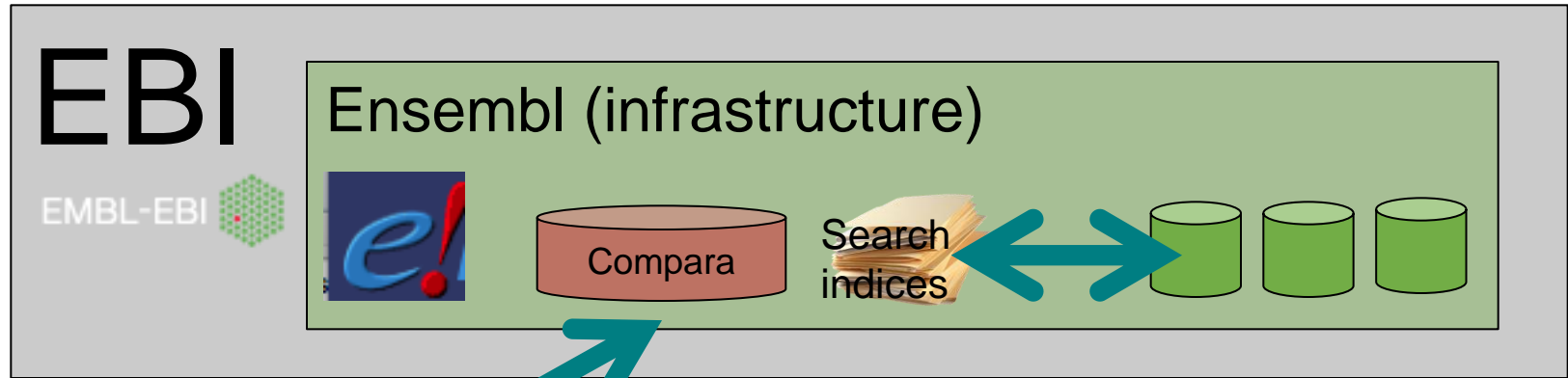


Collaborators



# A distributed future for Ensembl databases

---



Advanced  
collaborators  
(ELIXIR  
nodes?)



Collaborators

# People

- James Allen, Sanjay Boddu, **Dan Bolser**, Bruce Bolt, Mikkel Christensen, Paul Davis, Thomas Down, Christoph Grabmueller, Kevin Howe, **Arnaud Kerhornou**, Julia Khobdova, Eugene Kulesha, Naveen Kumar, Nick Langridge, Dan Lawson, Mark McDowall, Uma Maheswari, Gareth Maslen, Michael Nuhn, Chuang Kee Ong, Michael Paulini, Helder Pedro, Anton Petrov, Dan Staines, Electra Tappanari, Brandon Walts, Gary Williams
- My colleagues at in the various Ensembl teams at EBI
- Guy Cochrane (data growth slide)

# 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)