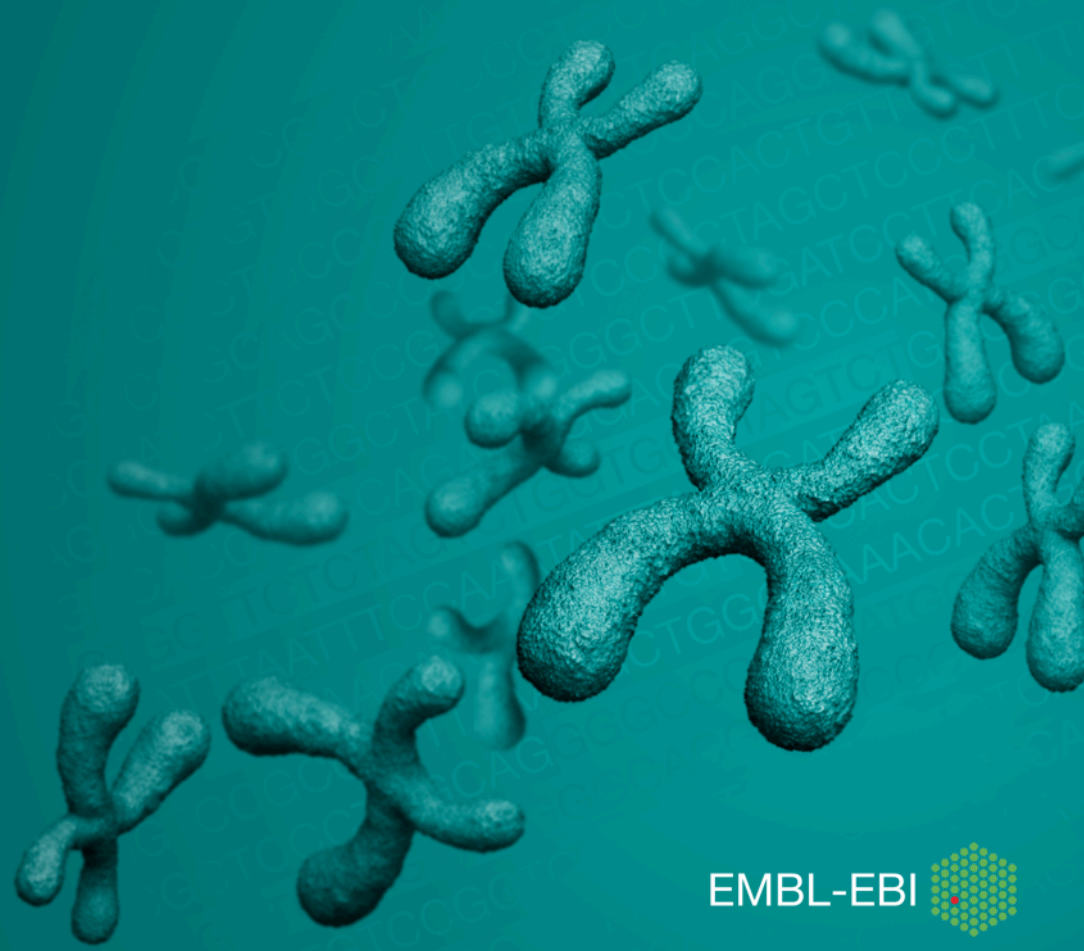
transPLANT-Elixir Plant Informatics Meeting

Paul Kersey



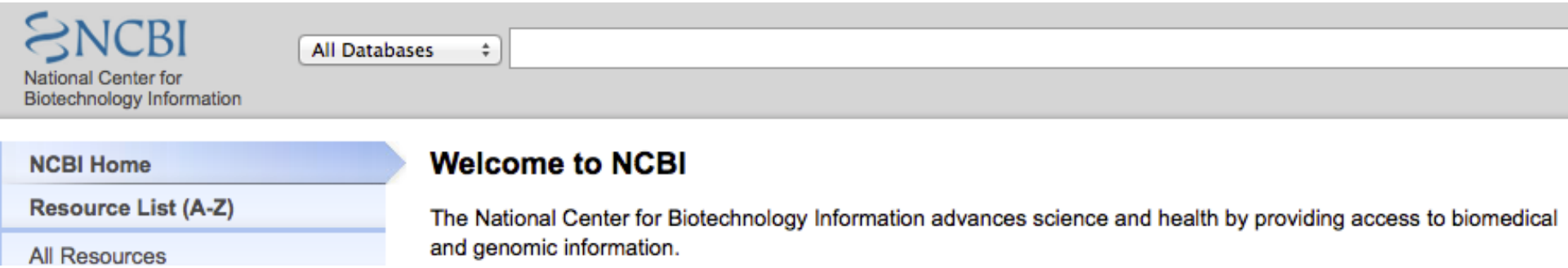
EMBL-EBI

Plant genomics/informatics is a rapidly advancing field

- Increasing numbers of species have sequenced reference genomes
  - Even large genomes from species such as wheat – a hexaploid with roughly ~5x the DNA content of human – are currently being deciphered
- Large scale resequencing, genotyping and phenotyping underway in most major crops
  - Possibility for direct application of knowledge in crop improvement
- Pest, pathogen, pollinator and symbiont genomes also being sequenced from across the taxonomic space

But plant genomics/informatics is traditionally underfunded

Genomics England, with the consent of participants and the support of the public, is creating a lasting legacy for patients, the NHS and the UK economy through the sequencing of 100,000 genomes: the 100,000 Genomes Project.



http://www.transplantdb.eu

transPLANT

- A 4 year EU FP7-funded project (DG CONNECT) coordinated by EMBL-EBI
- An I3 (e-infrastructure) project with elements of coordination, service and RTD
- Duration 4 years
  - (September 2011-August 2015)
  - What do we do next?

What is a scientific data infrastructure?

- Hardware, software, “peopleware”
- Data repositories
- Algorithms
- Standards for interoperability
  - Syntactic – data formats, APIs, controlled vocabularies
  - Semantic – minimum information, quality metrics, annotation practice etc.

Is there still a place for hardware in “scientific infrastructure”?

- Compute is a commercially available commodity
  - No-one expects us to build our laboratories ourselves
- Economics of commercial compute are not yet completely competitive if use is sufficiently intensive
- Distributed models maybe less suited to operations constrained bandwidth not CPU cycles
- A “science grid” may still make sense as part of an infrastructure
  - Is biology a sufficiently broad domain to need its own solution?

Is a place for plant biology as a specific domain within an infrastructure program for the life sciences?

- Many data structures, algorithms, and viewers common to all domains of life
- Arguably two models:
  - Generic, pre-competitive: fits all life
  - Specialised, adapted to commercial use cases: if viable, can support itself from market funds
- So what's left?

But complex data doesn't self organise

- Slow progress of semantic web cf. rapid progress of non-semantic web
- Tools/databases have some generic potential, but do need to be adapted to specific use cases
- With large numbers of data generators, making high-quality data available to users requires quality control
- Increasingly, reference data "is" the infrastructure
  - Illustration – think "Wikipedia" vs. "MediaWiki" vs. wherever the Wikipedia data center is….

elixir Data for Life

Search

about | services | organisation | events | news & media | members area

Welcome to ELIXIR

Building a sustainable European infrastructure for biological information, supporting life science research and its translation to medicine, agriculture, bioindustries and society.

"ELIXIR unites Europe's leading life science organisations in managing and safeguarding the massive amounts of data being generated every day by publicly funded research. It is a pan-European research infrastructure for biological information.

"ELIXIR will provide the facilities necessary for life science researchers - from bench biologists to cheminformaticians - to make the most of our rapidly growing store of information about living systems, which is the foundation on which our understanding of life is built."

- Dr Niklas Blomberg, ELIXIR Director

Celebrating Elixir

This meeting

- **Informational** – what are we all doing, what do we expect to be doing, what would we like to be doing?
  - How can we collaborate more closely
- **Strategic**: what do we mean by infrastructure, and how we create the infrastructure we need?
- **Financial**: what funding streams will be available to allow us to build the infrastructure we need?

This meeting

- What do the nodes plan to do in the field of plant genomics? (and what resources do they have to do this with)
- What do the transPLANT partners (many of whom are also involved in infrastructure provision, and some of whom will be directly involved in Elixir nodes), plan to do?
- Where will our activities be synergistic, and what can we do to make the benefits of these synergies available to the broadest possible user community?

This meeting

- Where do we expect the needs of plant researchers to be met by generic infrastructure, and where do we need plant-specific implementations
- What are other related infrastructures doing?
- Are there unplugged gaps in infrastructure provision?
- Are there activities that would benefit from being pursued collaboratively, and if so, how?
  - Are the funding opportunities (I3 or VRE) that we should pursue? transPLANT2 + certain Elixir nodes?

http://www.transplantdb.eu

What are the goals of transPLANT?

- A common set of reference data to be shared between different researchers and service provides
- Construction of missing data archives
- Provision of tools to manipulate and mine plant genomic data
- Provision of an integrating point of interactive access to diverse data sets
- Provision of a compute environment for programmatic access to plant genomic data
- Developing common standards for use within transPLANT and a wider community
- Training potential users
- Engaging with other related communities to share experiences, tools and roadmaps

Overall structure

- An I3 project funded by DG Connect under the framework 7 program
- 12 work packages, 4 types of activities
  - WP1 Management
  - WP2-4 Coordination activities
  - WP5-6 Service activities
  - WP7-12 RTD activities

Coordination and Support Activities

- WP2 Interaction with relevant communities
- WP3 Standards development
- WP4 User training

Service Activities

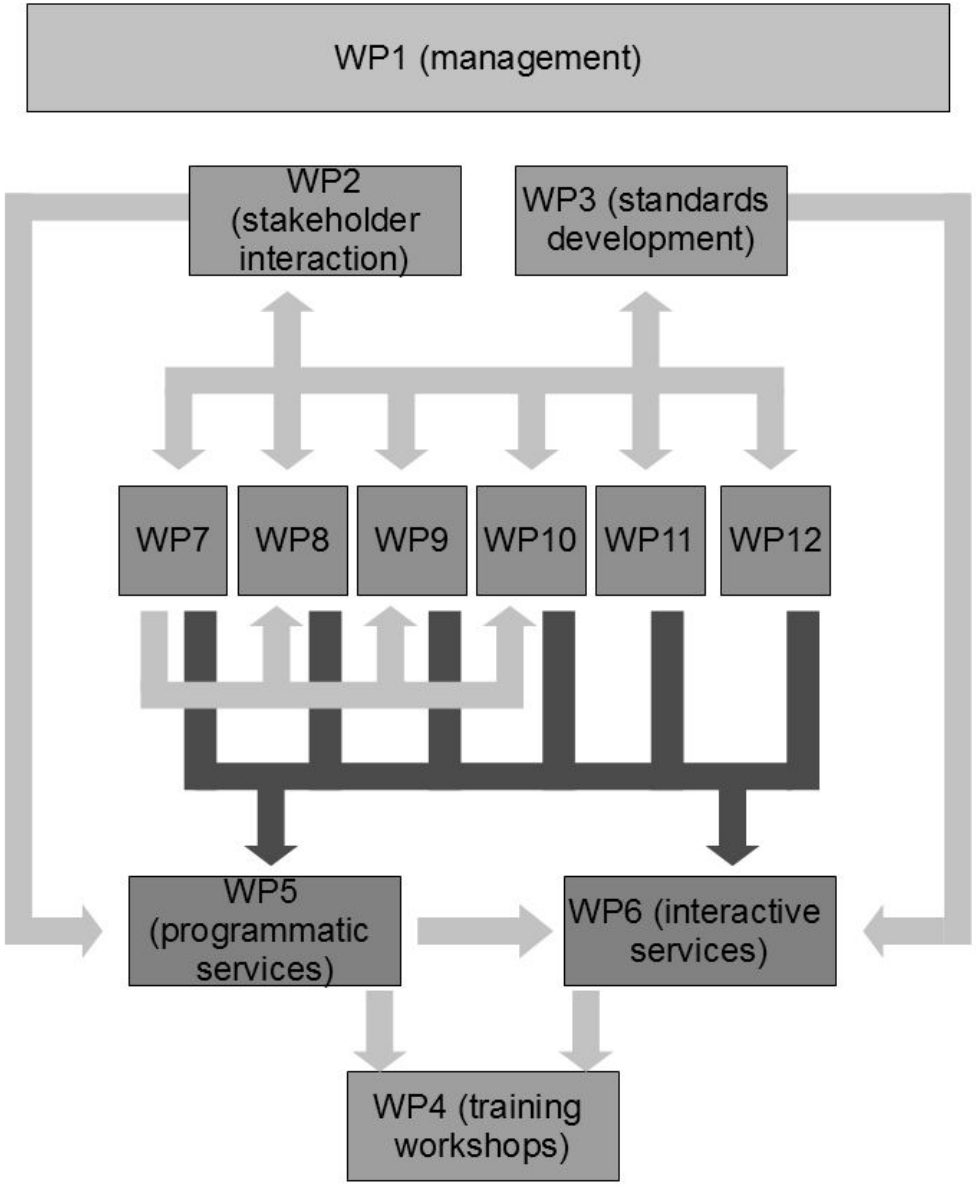
- WP5 Services for computational access
- WP6 transPLANT portal: a single point of access to distributed data

RTD Activities

- Develop the core data infrastructure components to which access will be provided in WPs 5 and 6
  - Reference genomic sequence (WP7)
  - A repository for variation data (WP9)
  - An associated toolset (WP8, 10, 11, 12)

RTD Activities

- WP7 - A reference repository for genomes
- WP8 - An architecture for plant genomic complexity
- WP9 - An archive for plant variation data
- WP10 - Linking genomes to phenotype
- WP11 - Information retrieval systems
- WP12 - Algorithm/tool evaluation, extenson and development


WP1 (management)
WP2 (stakeholder interaction)
WP3 (standards development)
WP7
WP8
WP9
WP10
WP11
WP12
WP5 (programmatic services)
WP6 (interactive services)
WP4 (training workshops)

From Genome to Variome

- A set of reference genomic data lies at the heart of the transPLANT (WP7)
- But developing a variation archive for plant data (WP9) is the critical new component of the infrastructure
  - dbSNP, run by NCBI, is not well-attuned to the needs of the plant community
  - Possible routes forward:
    - Brokered submission to dbSNP/intermediate data management
    - Collaboration with NCBI
    - Independent plant-focused resource

transPLANT activities

- Training, standards development and broader strategic planning
- Integrated search (model development and integration)
- Data coordination and exchange
- Variation archiving and tool development
- Collaborative work around representation of and computation with large genomic data sets

transPLANT

trans-National Infrastructure for Plant Genomic Science

INSIDE THIS ISSUE

TRAINING AND EVENTS

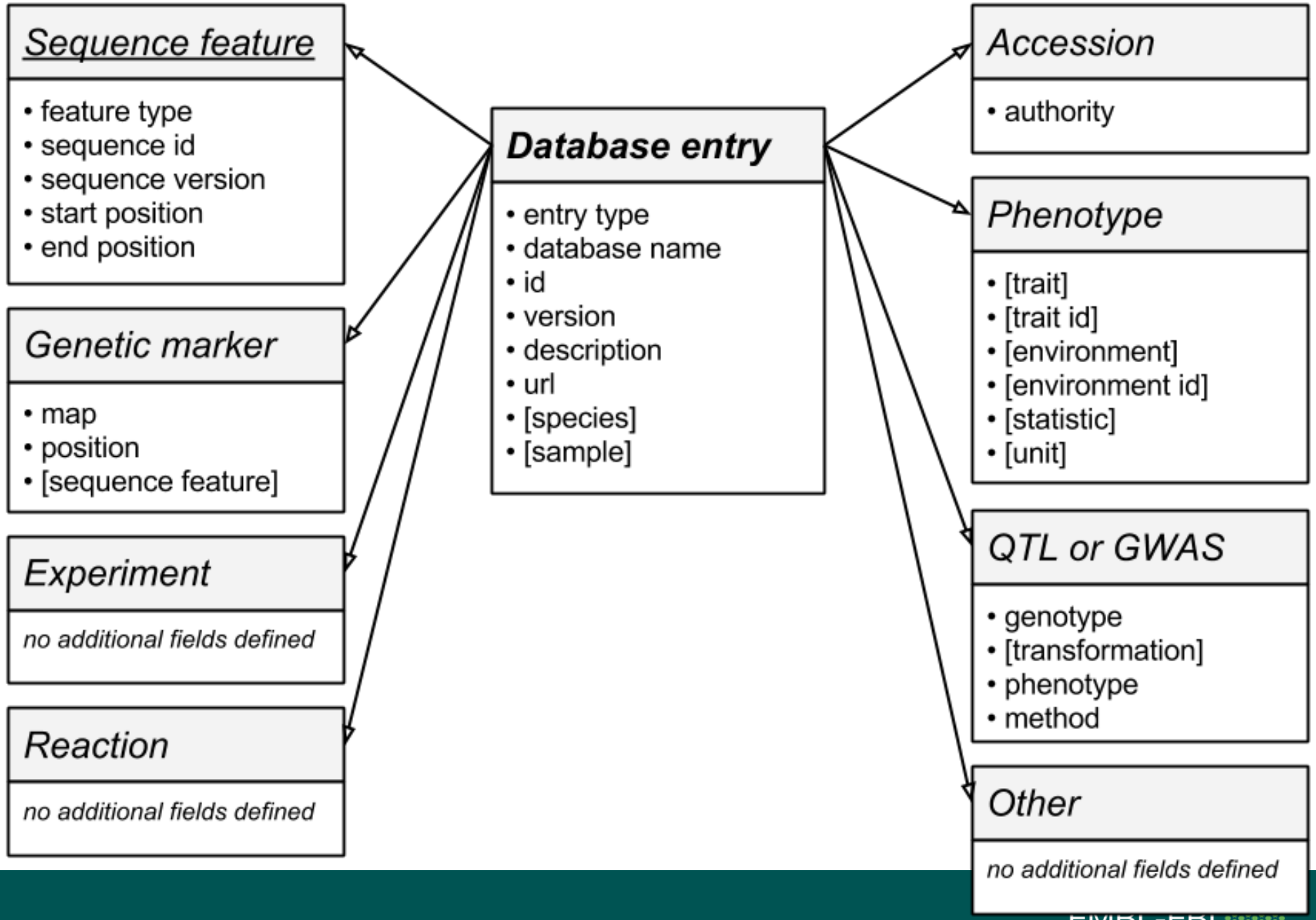
- 13-14 October 2014: 3rd transPLANT User Training Workshop on "Exploiting and Understanding Solonaceous Genomes". To be held at the DLO, Wageningen, Holland. Registration details will follow.

COMMENTS?

- We want to hear from you! Any comments or suggestions please contact us at transplant_help@ebi.ac.uk

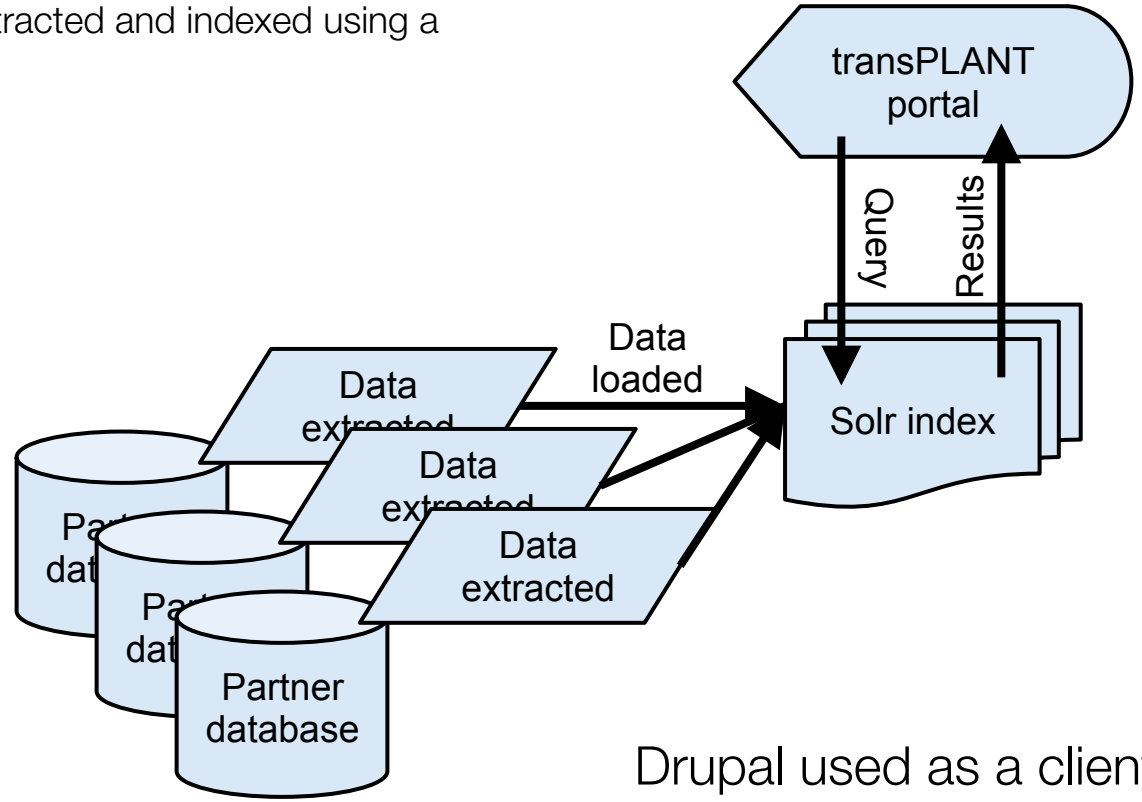
About the transPLANT project

transPLANT is a consortium of 11 European partners gathered to develop a trans-national infrastructure for plant genomic science. Bringing together groups with strengths in data analysis, plant science, computer science and from


Sequence feature
• feature type
• sequence id
• sequence version
• start position
• end position
Genetic marker
• map
• position
• [sequence feature]
Experiment
no additional fields defined
Reaction
no additional fields defined
Database entry
• entry type
• database name
• id
• version
• description
• url
• [species]
• [sample]
Accession
• authority
Phenotype
• [trait]
• [trait id]
• [environment]
• [environment id]
• [statistic]
• [unit]
QTL or GWAS
• genotype
• [transformation]
• phenotype
• method
Other
no additional fields defined

Current search implementation

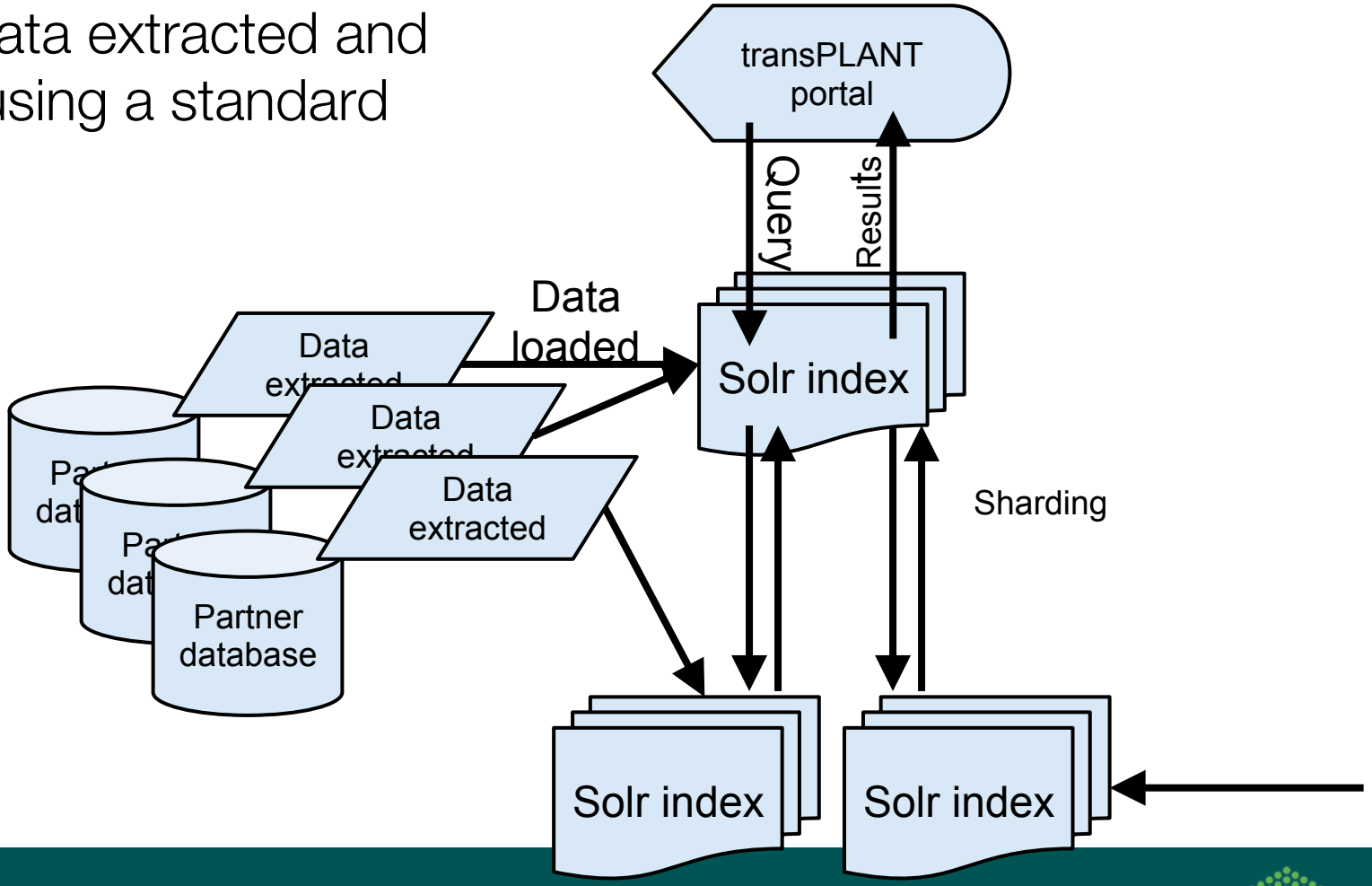
Partner data extracted and indexed using a simple schema



Drupal used as a client, querying a single Solr index

Truly distributed search

Partner data extracted and indexed using a standard schema



Current search

Search found 971 items

- rubisco

Filter by database:

- CR-EST (864)
- Ensembl Plants (62)
- PlantsDB (45)

Filter by data type:

- expressed sequence tags (864)
- protein_coding (62)
- transcript (45)

Filter by species:

- Hordeum vulgare (835)
- Arabidopsis thaliana (25)
- Oryza sativa (21)
- Solanum tuberosum (17)
- Brassica rapa (13)
- Pisum sativum (9)
- Musa acuminata (8)
- Nicotiana tabacum (8)
- Triticum urartu (7)
- Chlamydomonas reinhardtii (6)
- Aegilops tauschii (5)
- Medicago truncatula (5)

Search across plant genomics resources

Enter terms rubisco Search

Search results

Results from PlantsDB:

1. Sb01g000380.1

   similar to RuBisCO large subunit-binding protein subunit alpha, chloroplast precursor - ID=Sb01g000380;Description="similar to RuBisCO large subunit-binding protein subunit alpha, chloroplast precursor" ...

2. Sb09g014430.1

   similar to RuBisCO large subunit-binding protein subunit alpha, chloroplast precursor - ID=Sb09g014430;Description="similar to RuBisCO large subunit-binding protein subunit alpha, chloroplast precursor" ...

More results in PlantsDB...

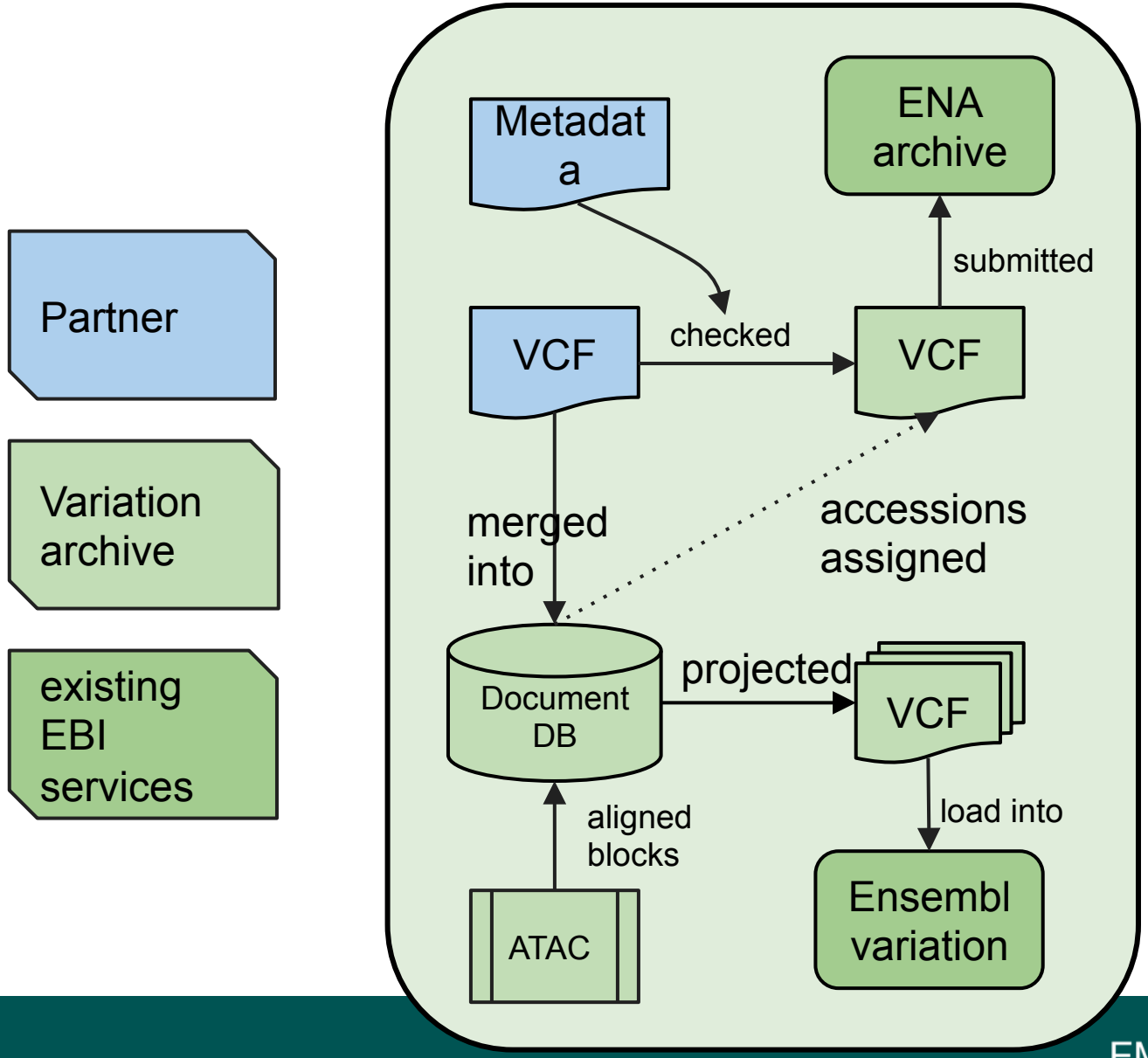
Results from Ensembl Plants:

1. Bra025431

   Bra025431 AT5G38410 (E=6e-092) | ribulose bisphosphate carboxylase small chain 3B / RuBisCO small subunit 3B (RBCS-3B) (ATS3B) ...

2. Bra028174

   Bra028174 AT5G38430 (E=7e-093) | ribulose bisphosphate carboxylase small chain 1B / RuBisCO small subunit 1B (RBCS-1B) (ATS1B) ...


Partner
Variation archive
existing EBI services
Metadat a
ENA archive
submitted
VCF
checked
VCF
merged into
accessions assigned
Document DB
projected
VCF
aligned blocks
load into
ATAC
Ensembl variation

transPLANT trans-National Infrastructure for Plant Genomic Science

Home | Resources | Events | Variation Archive | About

Variation Archive

- Introduction
- Accessioning
- Submit
- Download
- Search
- Query (beta)

Submit data to the variation archive

The transPLANT variation archive accepts variant calls in VCF format (version 4 and above) on known reference sequences, i.e. sequences present in the databases of the INSDC: ENA, GenBank, or DDBJ. Submitters must supply valid VCF and appropriate meta-data: Institute, Title, Study reference, Assembly reference, Sample references, and Sequence references.

For a more detailed description of the submission process, see the VCF submission in detail page.

Overview

1. Sign up
2. Upload VCF
3. Enter meta-data
4. Submit

- After signing up, you will receive email instructions for uploading your VCFs.
- After uploading, VCFs will be automatically processed.
- When processing is complete, you will receive an email with instructions for adding the required meta-data: Title, Study reference, Assembly reference, Sample references, Sequence references.
- Your VCF will be submitted to ENA and the assigned ENA Submission and Analysis accessions will be returned.

Sign up for a VCF submission account

Email address

Institute

Sign Up

Basic information

Please provide a description of your analysis.

Alias

Center name

Title

Description

Experiment type

References to ENA

Please provide existing ENA accessions for the study and assembly your analysis is referencing.

Study accession

Assembly accession

Sample mapping

Found these sample labels in your VCF, please provide existing ENA accessions for them.

IRGC103469/IRGC103469_aln_sorted.bam

TOG5457/TOG5457_aln_sorted.bam

TOG5467/TOG5467_aln_sorted.bam

TOG5923/TOG5923_aln_sorted.bam

TOG5949/TOG5949_aln_sorted.bam

TOG7025/TOG7025_aln_sorted.bam

TOG7102/TOG7102_aln_sorted.bam

Sequence mapping

Found these CHROMO labels in your VCF, please provide existing ENA accessions for them.

1

Submit

If you don't like forms...

Download the form as JSON and upload it after you have filled in all missing values.

Choose File No file chosen

Submit

vcf submission system

-user registers on website

-send email with FTP instructions

-user uploads vcf to ENA FTP

-validate and parse for required metadata

-send email to web form

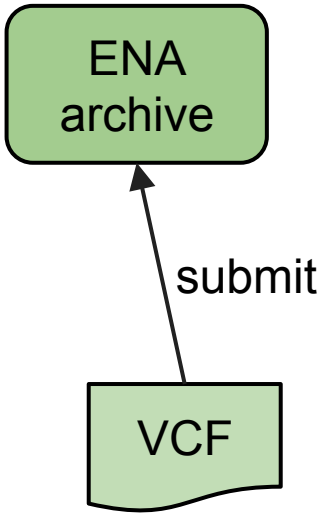
-user provides missing metadata

(sequence, samples, study)

-generate XML

-submit to ENA

-return accessions or errors



ENA European Nucleotide Archive

ENA Home | Search & Browse | Submit & Update | About ENA | Contact

• Please subscribe to ena-announce@ebi.ac.uk to receive alerts about ENA services.

Text search | Advanced search | Sequence search

Enter or paste text or ENA accession number: Search

Upload file of accessions: Choose File No file chosen Search

Study: ERP001451 : Whole genome shotgun survey sequencing of Barley cv. Haruna Nijo for de novo assembly of gene-space

View: XML

Download: XML

Send Feedback

Submitting Centre	Study Type	Read Count	Base Count
Leibniz Institute of Plant Genetics and Crop Plant Research	Whole Genome Sequencing		

Secondary accession(s)

PRJEB3042

Abstract

"Barley (Hordeum vulgare L.) is amongst the oldest domesticated crop plants and remains one of the world's most important crop species. It is diploid with a haploid genome of 5.1 gigabases (Gb), twice the size of those of human and maize, and closely related to the most widely grown crop, hexaploid wheat. To meet global demand for food, feed and fibre, it is commonly agreed that reference genome sequences of our crop plants are urgently required to enable genome-assisted crop improvement. In the framework of the The International Barley Genome Sequencing Consortium (IBSC) we present here the whole genome shotgun rawdata of different barley cultivars and genotypes to build a physical ordered genetic and functional sequence resource."

Navigation | Read Files | **Analysis Files** | Attributes

Download files

View: TEXT

Download: TEXT

Select columns

Showing results 1 - 1 of 1 results

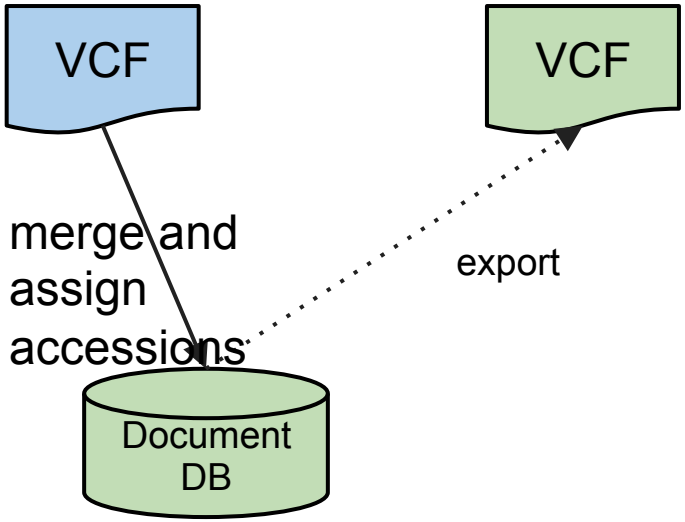
Analysis accession	Study accession	Secondary study accession	Sample accession	Secondary sample accession	Scientific name	Submitted files (ftp)	Submitted files (galaxy)	CoL tax ID	CoL scientific name
ERZ016065	PRJEB3042	ERP001451	SAMEA1557037	ERS140602	Hordeum vulgare subsp. vulgare	VCF File 1	VCF File 1		

merge and accession

merge variants based on position

stable variant accessions

project variants when reference assembly updates



merge in detail

RNA example

- load existing 15.3 million variants from db
  into bloom filter 5 min
- merge 67k variants 3 min
- assign accessions to 15k novel variant positions same 3 min

merge result

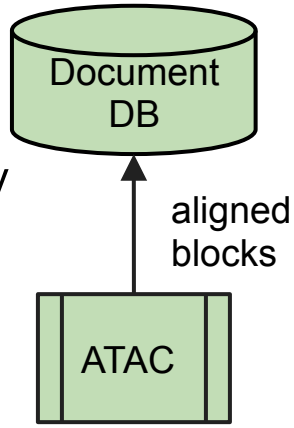
#CHROM	POS	ID	REF	ALT	...	FORMAT	barke	bowman	...
contig_2	152	vcZ00001	A	G	...	GT:GQ:PL	.:.:.	.:.:.	
contig_4	251	vcZ00002	A	C	...	GT:GQ:PL	1/1:99:255,60,0	1/1:99:255,141,0	
contig_4	268	vcZ00003	C	T	...	GT:GQ:PL	.:.:.	.:.:.	
contig_4	297	vcZ00004	G	A	...	GT:GQ:PL	.:.:.	.:.:.	
contig_4	478	vcZ00005	G	A	...	GT:GQ:PL	1/1:99:255,51,0	1/1:99:255,129,0	
contig_4	581	vcZ00006	C	A	...	GT:GQ:PL	1/1:99:255,63,0	1/1:99:255,138,0	
contig_4	808	vcZ00007	G	A	...	GT:GQ:PL	.:.:.	1/1:69:212,36,0	

projecting variants

ATAC - Assembly To Assembly Comparison

mapping between two genome assemblies

generates genome-wide list of assembly-to-assembly blocks that are at least 95% identical



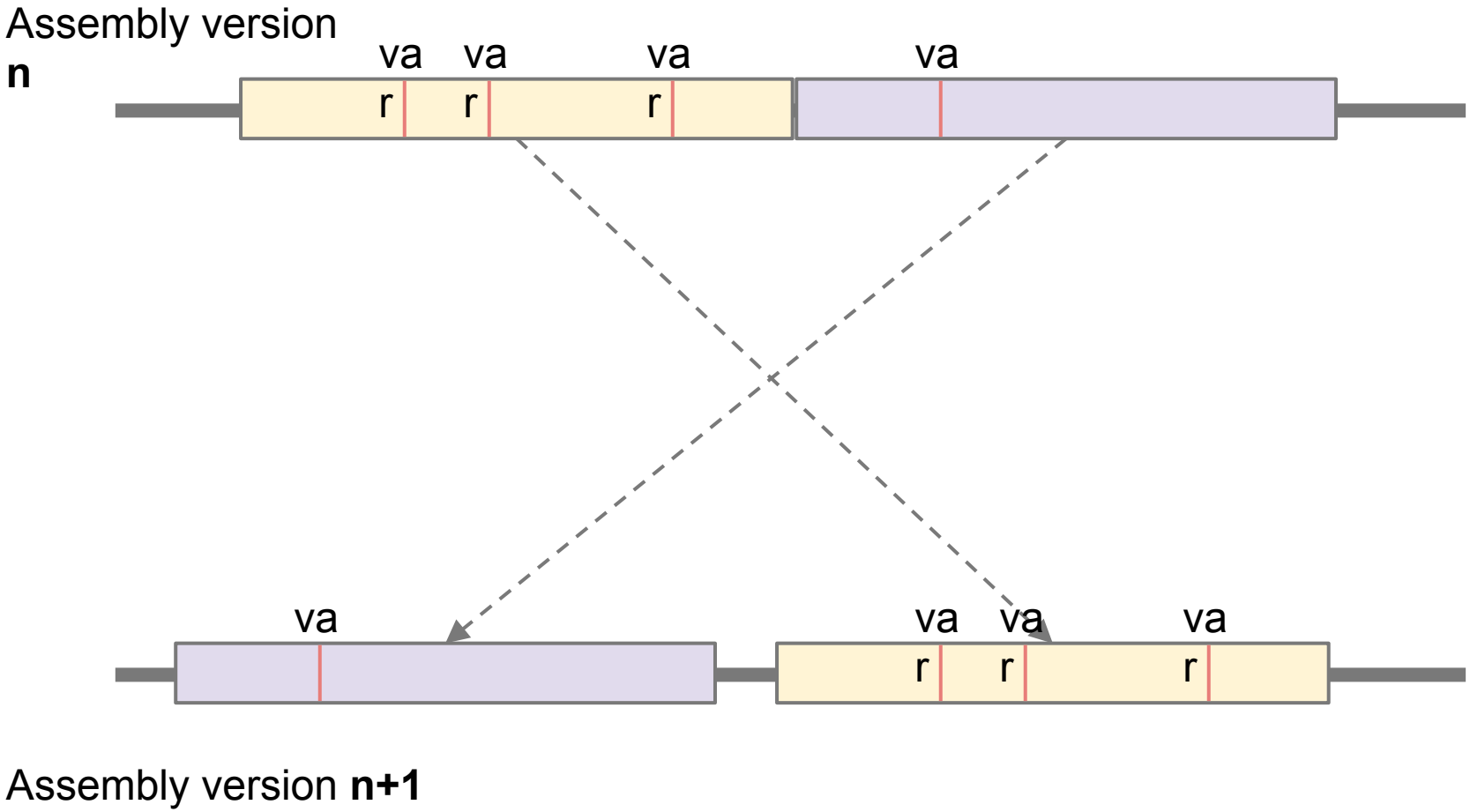
atac assembly mapper

variation archive can lift-over/project variants from one assembly version to the next

make mappings available via the Ensembl website

- find all Ensembl Plants databases with distinct assembly versions
- run ATAC on their DNA sequences
- load assembly-to-assembly matches into the Ensembl schema

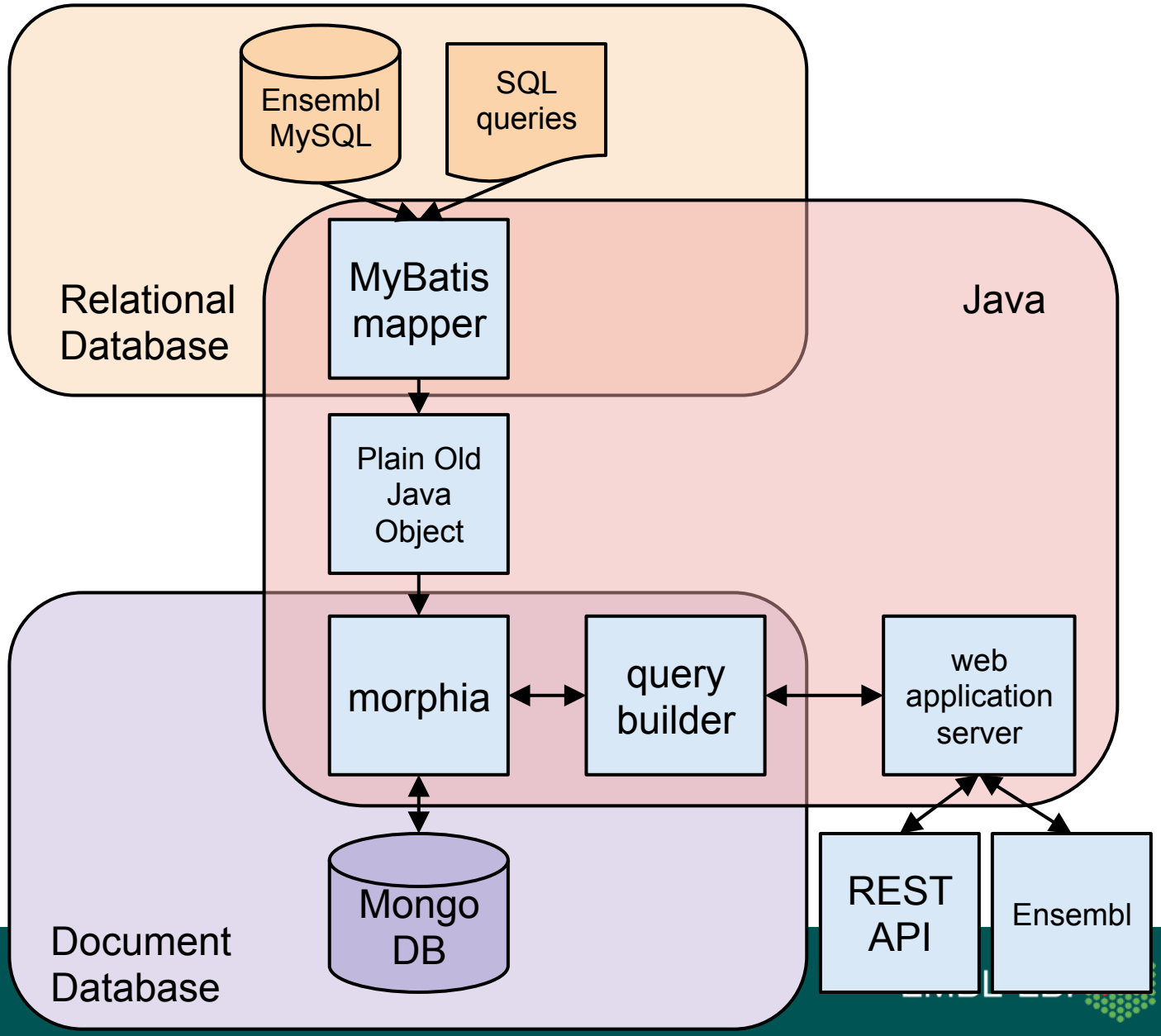
assembly projection via whole genome alignment



query tool

16 **columns**
from
10 **tables**
per variant

1 **document**
per variant



A Virtual European Plant Database

- No single resource is adequately funded to run a “European Plant Database”
- Expertise in different crops, experimental approaches and analysis techniques are distributed throughout many countries
- Different interfaces serve difference purposes

A Virtual European Plant Database

- This can be an advantage to users if:
  - Valuable data is persistently stored and remains accessible
  - Users can find the data they want
  - Users can combine data that resides in different places
    - Requires the use of common identifiers and descriptors for sequence, phenotype, cultivars, etc.
  - Users can compute against the data
    - Software and hardware components to this

5 challenges of the post-genomic era

- Data storage
- Data compute
- Data interpretation
- Data integration
- Genotype to phenotype

EnsemblPlants ▾ BLAST | Sequence Search | BioMart | Tools | Downloads | Help | Documentation

Login/Register

Search Ensembl Plants…

Search: All species for Go

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)
  - Five new rice genomes:
    - *Oryza barthii*
    - *Oryza glumaepatula*
    - *Oryza meridionalis*
    - *Oryza nivara*
    - *Oryza punctata*
- Updated genomes
  - *Triticum aestivum* (bread wheat) gene models updated to v2.1.
- New data
  - Whole genome alignments within *T. aestivum*, between its component A, B, and D genomes.
  - Calculation of homoeologous genes between the *T. aestivum* component A, B and D genomes.
  - Whole genome alignments between *Amborella trichopoda* and the two following genomes, *Oryza sativa* Japonica and *Arabidopsis thaliana*.

Did you know...?

The Assembly Converter Tool allows the conversion of a variety of annotation formats between different versions of a genome 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 URGI

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.

Databases are constructed in a direct collaboration with the Gramene resource, funded by the United States National Science Foundation award #1127112. More information about our collaboration with Gramene in available here.

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

Ensembl Plants Species

Key

Species V P G A
Provider | *Scientific name* | Taxonomy ID

V - has a variation database, P - is in pan-taxonomic compara, G - has whole genome DNA alignments, A - has other alignments

Liliopsida


Aegilops tauschii G A
BGI | *Aegilops tauschii* | 37682


Brachypodium distachyon V G A
Brachypodium.org | *Brachypodium distachyon (L.) Beauv* | 15368


Hordeum vulgare V G A
IBSC | *Hordeum vulgare* | 112509


Musa acuminata G
CIRAD | *Musa acuminata Doubled-haploid Pahang (DH-Pahang)* | 214687


Oryza barthii G A
OGE | *Oryza barthii* | 65489


Oryza brachyantha G A
OGE | *Oryza brachyantha* | 4533


Oryza glaberrima V G A
AGI | *Oryza glaberrima* | 4538


Oryza glumaepatula G A
OGE | *Oryza glumaepatula* | 40148


Oryza meridionalis G A
OGE | *Oryza meridionalis* | 40149


Oryza nivara G A
OGE | *Oryza nivara* | 4536


Oryza punctata G A
OGE | *Oryza punctata* | 4537


Oryza sativa Indica Group V G A
RIS | *Oryza indica 93-11 (Indica rice)* | 39946


Oryza sativa Japonica (Rice) V P G
RAP-DB | *Oryza sativa* | 39947


Setaria italica G A
JGI | *Setaria italica* | 4555


Sorghum bicolor V G A
JGI | *Sorghum bicolor BTX623* | 4558


Triticum aestivum G A
IWGSP1 | *Triticum aestivum Chinese Spring* | 4565


Triticum urartu G A
BGI | *Triticum urartu* | 4572


Zea mays V G A
MaizeSequence.org | *Zea mays* | 4577

eudicotyledons


Arabidopsis lyrata G A
JGI | *Arabidopsis lyrata* | 81972


Arabidopsis thaliana V P G A
TAIR | *Arabidopsis thaliana* | 3702


Brassica rapa G A
IVFCAAS | *Brassica rapa* | 51351


Glycine max G A
JGI | *Glycine max* | 3847


Medicago truncatula G A
IMGAG | *Medicago truncatula A17* | 3880


Populus trichocarpa G A
JGI | *Populus trichocarpa* | 3694


Prunus persica G A
International Peach Genome Initiative | *Prunus persica* | 3760


Solanum lycopersicum P G A
ITGSP | *Solanum lycopersicum* | 4081


Solanum tuberosum G A
PGSC | *Solanum tuberosum* | 4113


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

Lycopodiophyta

Selaginella moellendorffii G A
European Nucleotide Archive JGI | *Selaginella moellendorffii* | 88036

Bryophyta

Physcomitrella patens P G A
JGI | *Physcomitrella patens* | 145481

Chlorophyta

Chlamydomonas reinhardtii P G A
JGI | *Chlamydomonas reinhardtii* | 3055

Rhodophyta

Cyanidioschyzon merolae P G
European Nucleotide Archive Cyanidioschyzon merolae Genome Project | *Cyanidioschyzon merolae* | 280699

Amborellales

Amborella trichopoda P G
Amborella Genome Database | *Amborella trichopoda* | 13333

EnsemblPlants

Home > Arabidopsis thaliana [TAIR9]

Login / Register | BLAST | BioMart | FTP | Docs & FAQs

Location: 3:19,431,371-19,434,403 | Gene: UEV1D-4 | Transcript: AT3G52560.1

Gene-based displays
- Gene summary
- Splice variants (3)
- Supporting evidence
- Sequence
- External references (3)
- Regulation
- Plant Compara
  - Genomic alignments (7)
  - Gene Tree (image)
    - Gene Tree (text)
    - Gene Tree (alignment)
  - Orthologues (16)
  - Paralogues (3)
- Pan-taxonomic Compara
  - Gene Tree (image)
    - Gene Tree (text)
    - Gene Tree (alignment)
  - Orthologues (43)
  - 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: UEV1D-4 (AT3G52560-TAIR-G)

UEV1D-4 (UBIQUITIN E2 VARIANT 1D-4); protein binding / ubiquitin-protein ligase; MMZ4/UEV1D encodes a protein that may play a role in DNA damage responses and error-free post-replicative DNA repair by participating in lysine-63-based polyubiquitination reactions. UEV1D-4, a predicted splice variant, can interact relatively weakly with UBC35/UBC13A and UBC36/UBC13B in a yeast-2-hybrid UEV1D-4 can also significantly, but not totally, functionally complement an mms2 mutation in budding yeast by increasing mms2 mutant viability in the presence of the DNA damaging agent MMS. uev1d-1 mutants are more sensitive than wild type plants to the DNA damaging agent MMS in seed germination and pollen germination assays.

Location Chromosome 3: 19,494,381-19,496,084 reverse strand.

Transcripts There are 3 transcripts in this gene: hide transcripts

Name	Transcript ID	Protein ID	Description
AT3G52560.1	AT3G52560.1-TAIR	AT3G52560.1-P	protein_coding
AT3G52560.2	AT3G52560.2-TAIR	AT3G52560.2-P	protein_coding
AT3G52560.3	AT3G52560.3-TAIR	AT3G52560.3-P	protein_coding

Gene summary help

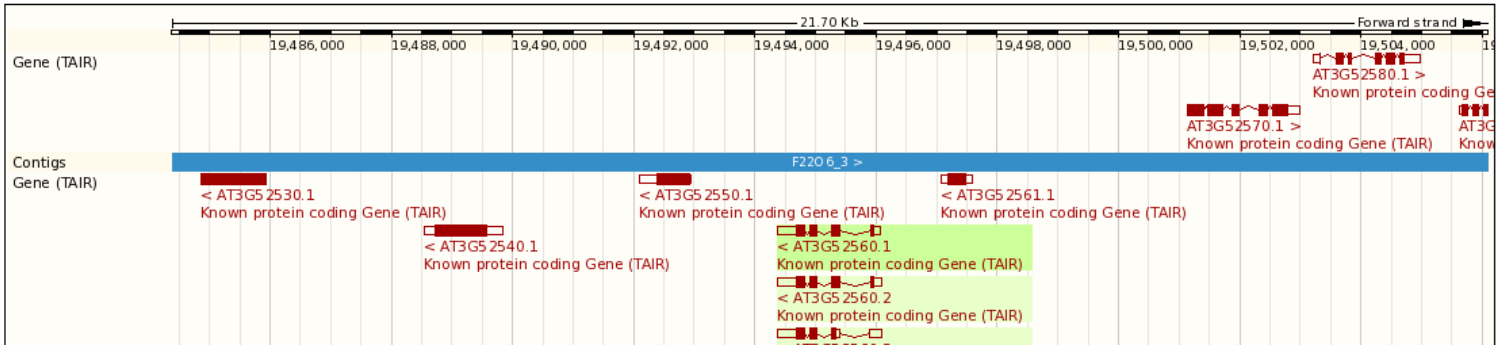
Splice variants »

Name UEV1D-4 (TAIR Gene Name)

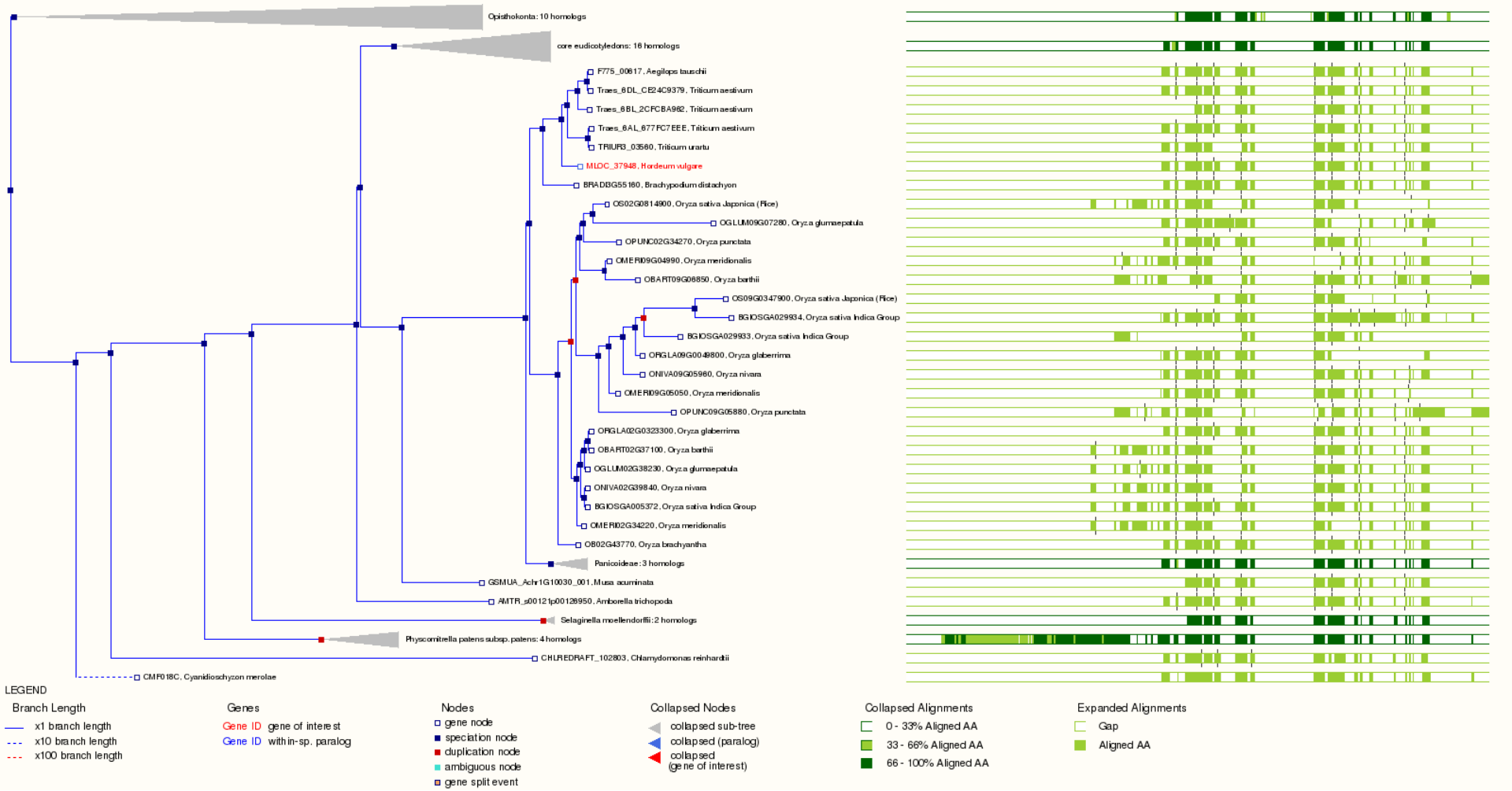
Gene type Known protein coding

Prediction Method Gene annotation by TAIR through a process of automatic and manual curation

Transcripts



1:1 orthology calls over 19 cereals including the three sub-genomes of bread wheat



EMBL-EBI

EnsemblPlants

Home > Arabidopsis thaliana

Login / Register | BLAST | BioMart | FTP | Docs & FAQs

Location: 1:8,001-18,000 | Gene: KEA1 | Transcript: AT1G01790.1

Gene-based displays

- Gene summary
- Splice variants (1)
- Supporting evidence
- Sequence
- External references (3)
- Regulation
- Plants Compara
  - Genomic alignments (7)
  - Gene Tree (image)
    - Gene Tree (text)
    - Gene Tree (alignment)
  - Orthologues (12)
  - Paralogues (5)
- Pan-taxonomic Compara
  - Gene Tree (image)
    - Gene Tree (text)
    - Gene Tree (alignment)
  - Orthologues (6)
  - Paralogues (5)
  - 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

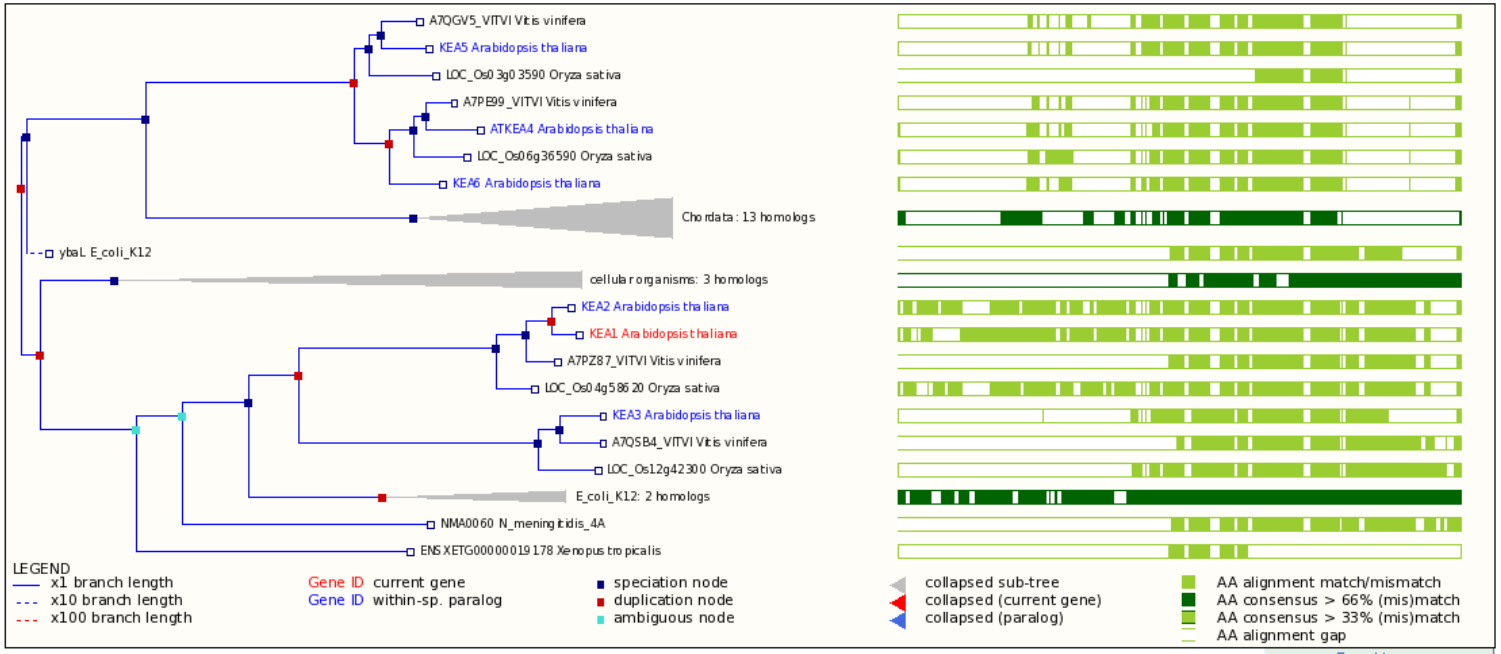
Gene: KEA1 (AT1G01790-TAIR-G)

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.

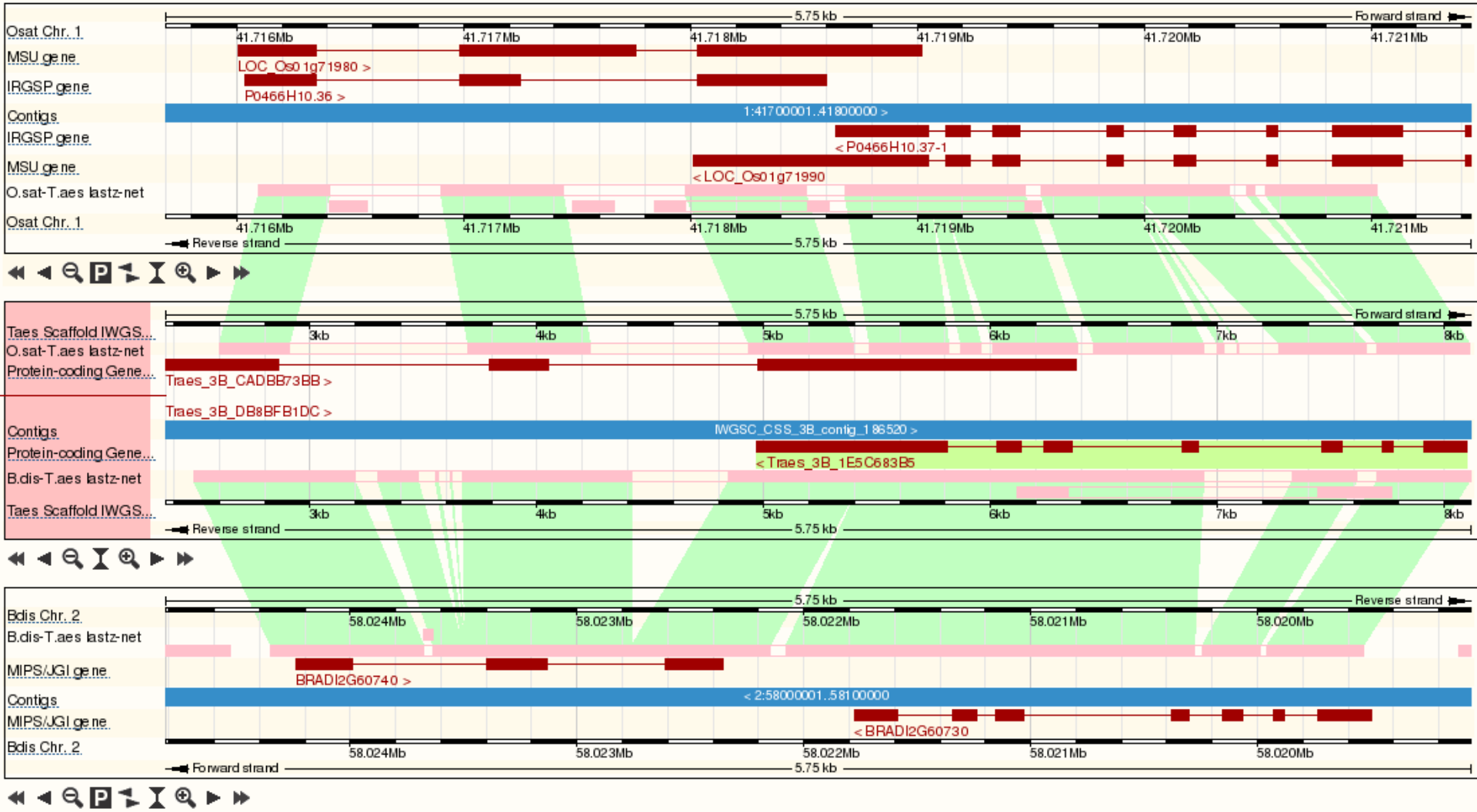
Transcripts There is 1 transcript in this gene: show transcripts

« Paralogues | **Gene Tree (image)** | Orthologues »



View options:

- View current gene only
- View paralogs of current gene
- View all duplication nodes


Osat Chr. 1
MSU gene
IRGSP gene
Contigs
IRGSP gene
MSU gene
O.sat-T.aes lastz-net
Osat Chr. 1
5.75 kb
Forward strand
41.716Mb
41.717Mb
41.718Mb
41.719Mb
41.720Mb
41.721Mb
LOC_Os01g71980 >
P0466H10.36 >
1:41700001..41800000 >
< P0466H10.37-1
< LOC_Os01g71990
Reverse strand
Taes Scaffold IWGS...
O.sat-T.aes lastz-net
Protein-coding Gene...
Contigs
Protein-coding Gene...
B.dis-T.aes lastz-net
Taes Scaffold IWGS...
3kb
4kb
5kb
6kb
7kb
8kb
Traes_3B_CADBB73BB >
Traes_3B_DB8BFB1DC >
IWGSC_CSS_3B_contig_186520 >
< Traes_3B_1E5C683B5
Bdis Chr. 2
B.dis-T.aes lastz-net
MIPS/JGI gene
Contigs
MIPS/JGI gene
Bdis Chr. 2
58.024Mb
58.023Mb
58.022Mb
58.021Mb
58.020Mb
BRADI2G60740 >
< 2:58000001..58100000
< BRADI2G60730

EMBL-EBI

EnsemblPlants

Home > Arabidopsis thaliana

Login / Register | BLAST | BioMart | FTP | Docs & FAQs

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

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

Ensembl Plants is produced in collaboration with Gramene

DB built by NASC

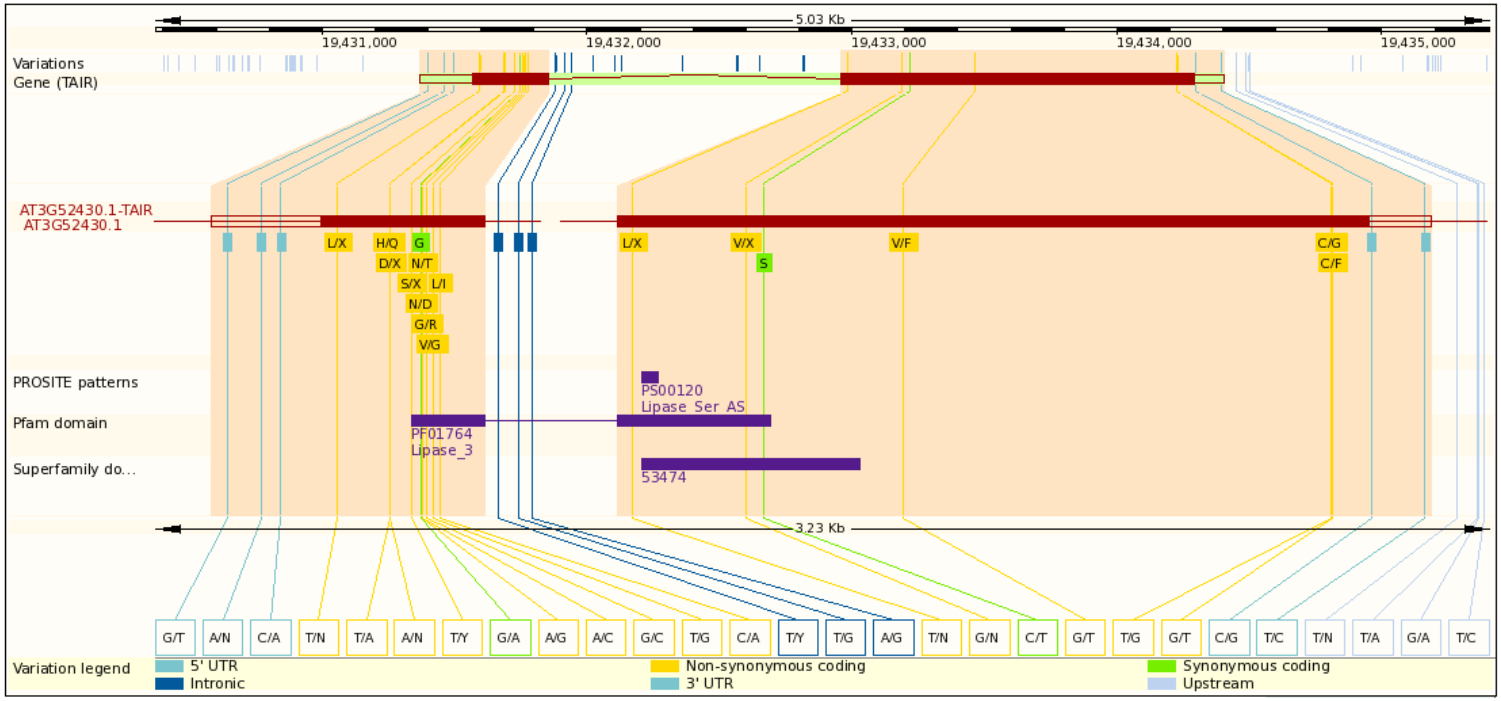
Gene: PAD4 (AT3G52430-TAIR-G)

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 »



EMBL-EBI

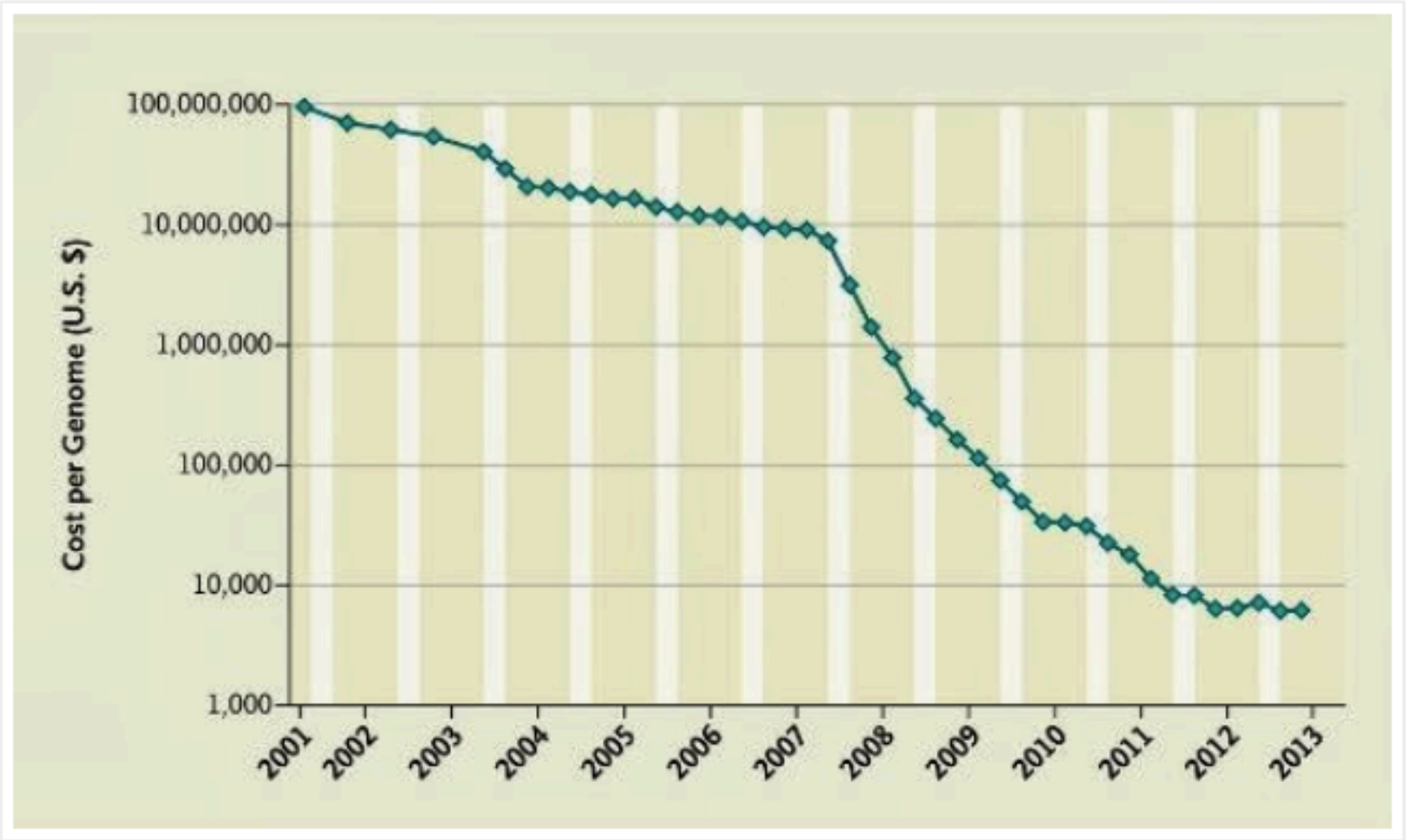
Example – accessing data in Ensembl

- Perl API

```
my $genes = $first_clone->get_all_Genes();
while ( my $gene = shift @{$genes} ) {
        print $gene->stable_id(), "\n";
}
```

- REST-ful API
- http://beta.rest.ensemblgenomes.org/lookup/id/AT3G52430?content-type=application/json;expand=1
- {"source":"ensembl","object_type":"Gene","logic_name":"tair","species":"arabidopsis_thaliana","description":"alpha/beta-Hydrolases superfamily protein [Source:TAIR_LOCUS;Acc:AT3G52430]","display_name":"PAD4","biotype":"protein_coding","end”: ….

Cost of Sequencing a Human Genome 2001-2013



What do the next five years hold for plant genomics?

- Every important model and crop genome sequenced
  - Improving reference assemblies for difficult crops, but unlikely to have complete molecular assemblies
    - Longer read technology likely to be helping
  - Structural diversity likely to continue to be poorly organised

    (but of course it's always possible to believe)
- Extensive genotyping of gene bank accessions
- Extensive sequencing of crop wild relatives

What can we expect in plant phenomics?

- Increasingly automated phenotyping
  - Both in phenotyping centres and in the field
  - Both imaging and molecular characterisation
- Large scale “conventional” characterisation of economically relevant traits in multiple lines in genome-wide associated studies

What can we expect in plant phenomics?

- Increasingly automated phenotyping
  - Both in phenotyping centres and in the field
  - Both imaging and molecular characterisation
- Large scale “conventional” characterisation of economically relevant traits in multiple lines in genome-wide associated studies

GxPxEek!

- Genome: one genome per species, per population, per individual
- Phenotype: one phenotype per individual per experimental condition
- Field trials: one trait per crop per temporal/spatial location
  - May be measured in increasing resolution
  - Deployment of monitoring technology directly in agricultural context

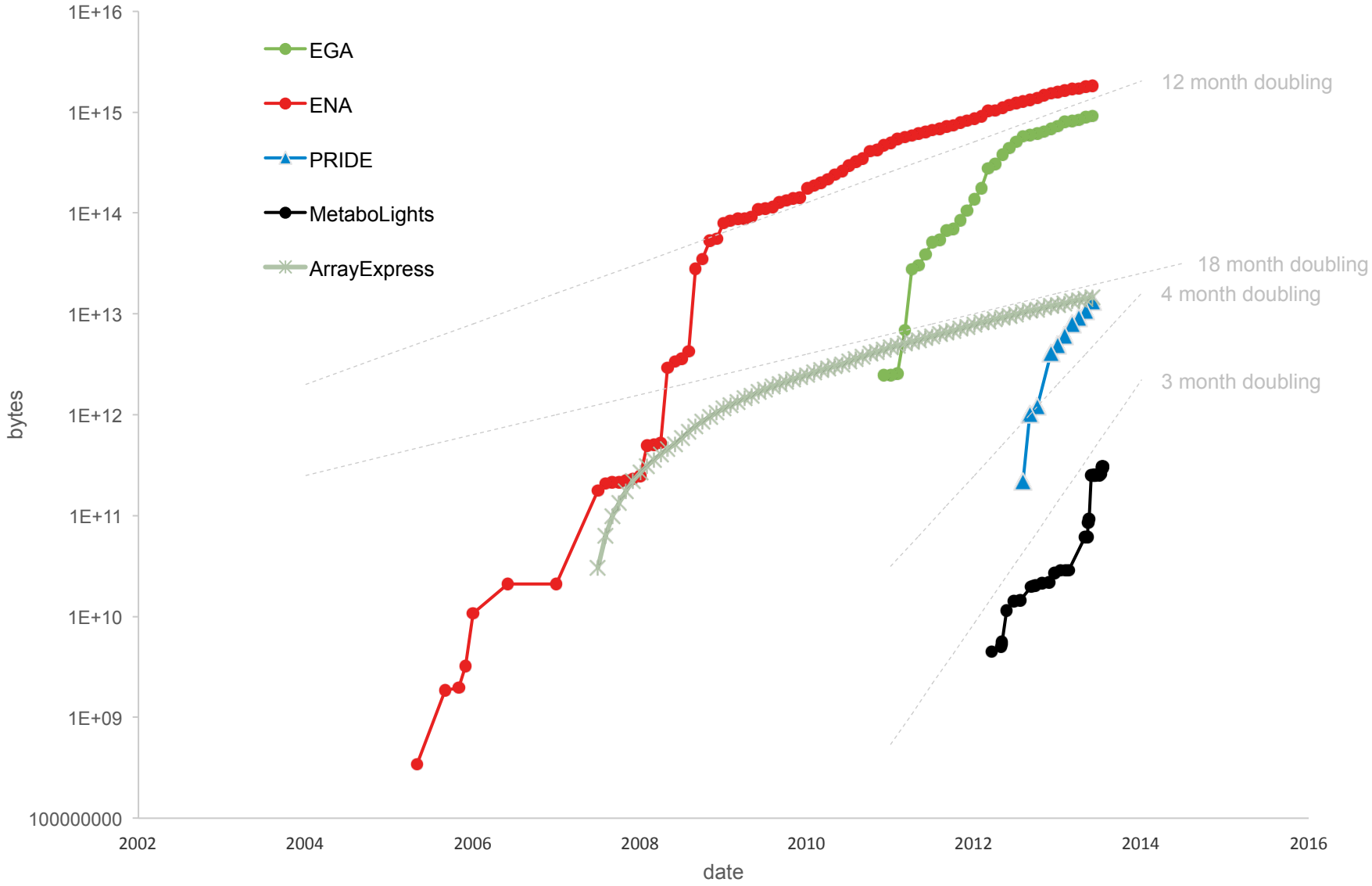
The EBI mission

- 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.

EBI mission

- Primarily focused on molecular data
  - Archive and interpretative services
- We aim to capture
  - All “reference data”
  - All molecular data from scientific experiments
    - Supporting/comprising “the literature”
- Medical informatics is out of scope, but medical research is very much in scope
  - Use of common technologies/open APIs to enable use of reference data in the widest range of contexts

Growing data



The next 5-10 years

- Expected move from petabytes (10^{15}) of storage to exabytes (10^{18}) of storage
- As biology becomes more data intensive, we can anticipate some increase in storage budgets
  - Sequencing technology, CPU is progressing more rapidly than storage technology
  - Small chance of exponentially increasing budgets

If sequencing is so cheap, do we need to archive?

- If we don't keep (and distribute) (some record of) the data, then the data wasn't worth producing, either
  - EBI mission
  - Scientific accountability
  - Open raw data to multiple interpretations
  - Additive value of multiple experiments (population, comparative studies)
- Data may be cheaper to reproduce than to store
  - Not yet
  - Storing data is inherently cheaper than storing samples
    - Some samples (e.g. cancer patient data) may not be recoverable

Some comments

- Sequence read archive already compresses raw image data 200-500 fold
  - Most data never even leaves the machine
  - Same will apply to images for phenotypes
- Storing data is easier than storing samples (e.g. cancer tissues, etc.)
- Electronic records are easier to distribute than samples
- Archiving old data is effectively “free” (in terms of disk capacity)

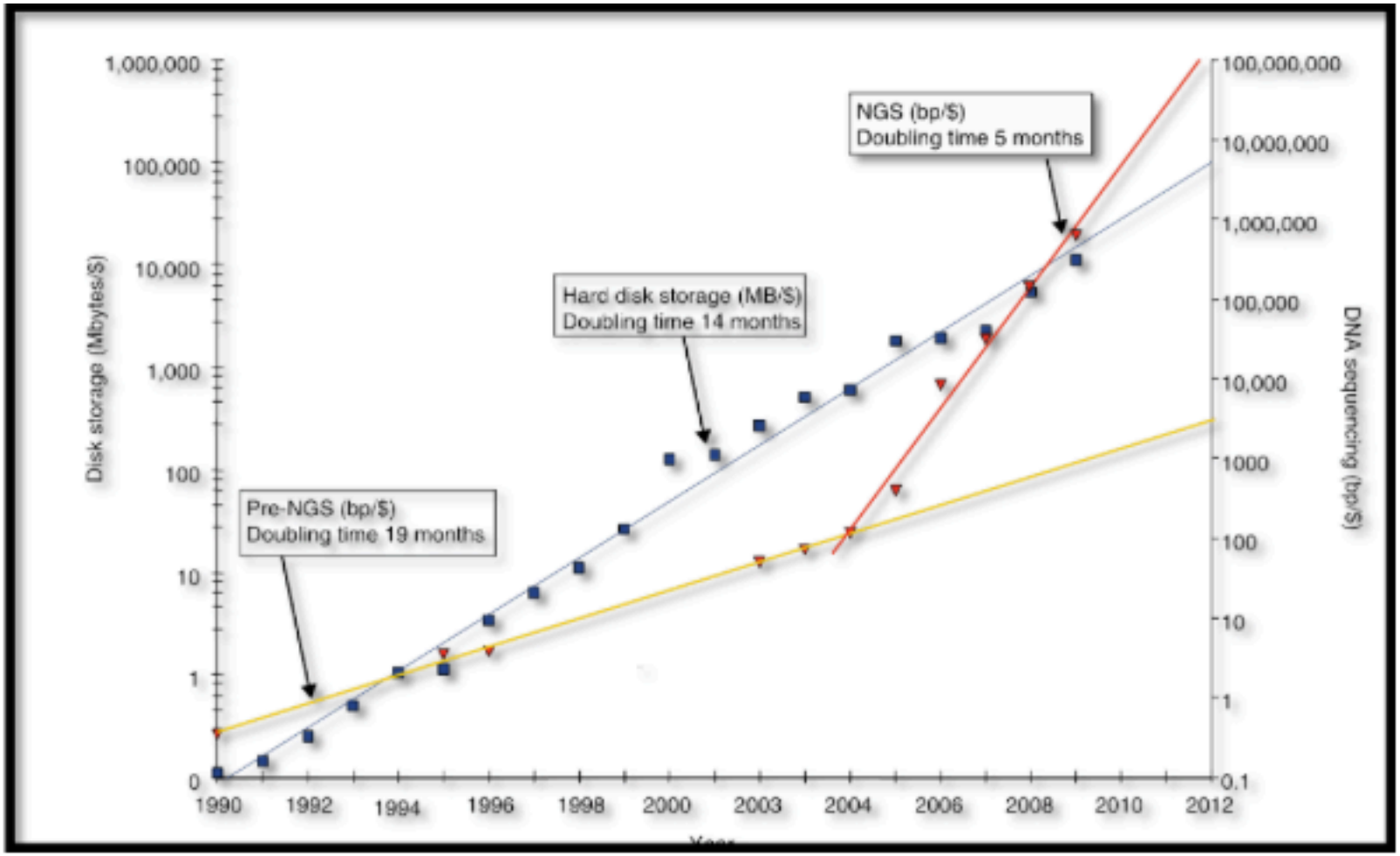
Reference-based compression

- 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

Distributed or centralised storage

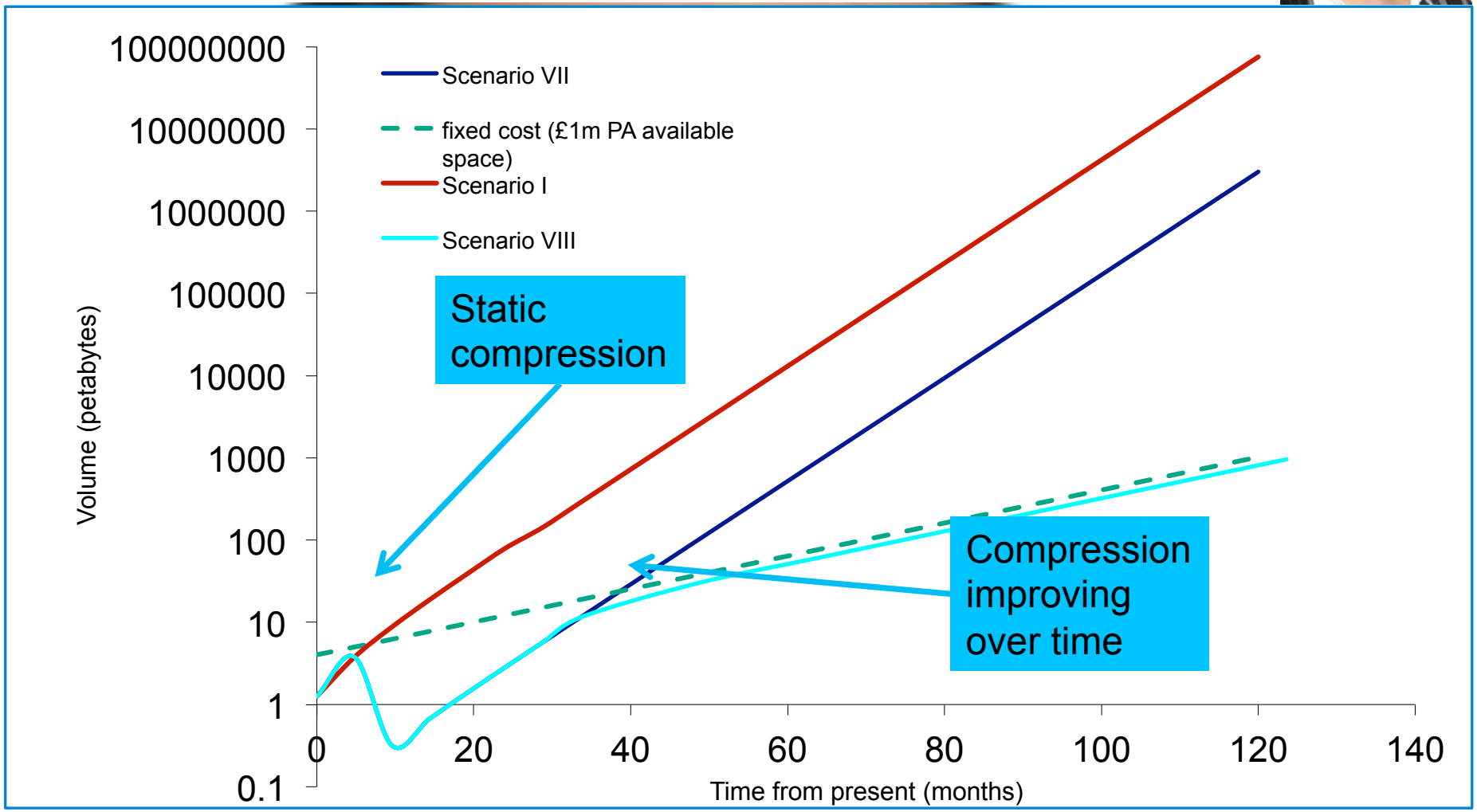
- Distributed data still needs storing
  - Communication costs, potentially insufficient concentration of expertise to get economies of scale
  - Where data cannot be centralised, common technology frameworks keep this transparent from users

A bleak prediction



Historical trends in storage prices versus DNA sequencing costs (reprinted from Stein, L.D., *Genome Biology* 2010, 11:207)

Compression options



CSH PRESS GENOME RESEARCH

Efficient storage of high throughput sequencing data using reference-based compression

Markus Hsi-Yang Fritz, Rasko Leinonen, Guy Cochrane and Ewan Birney[1]

[+] Author Affiliations

* Corresponding author; email: birney@ebi.ac.uk

Abstract

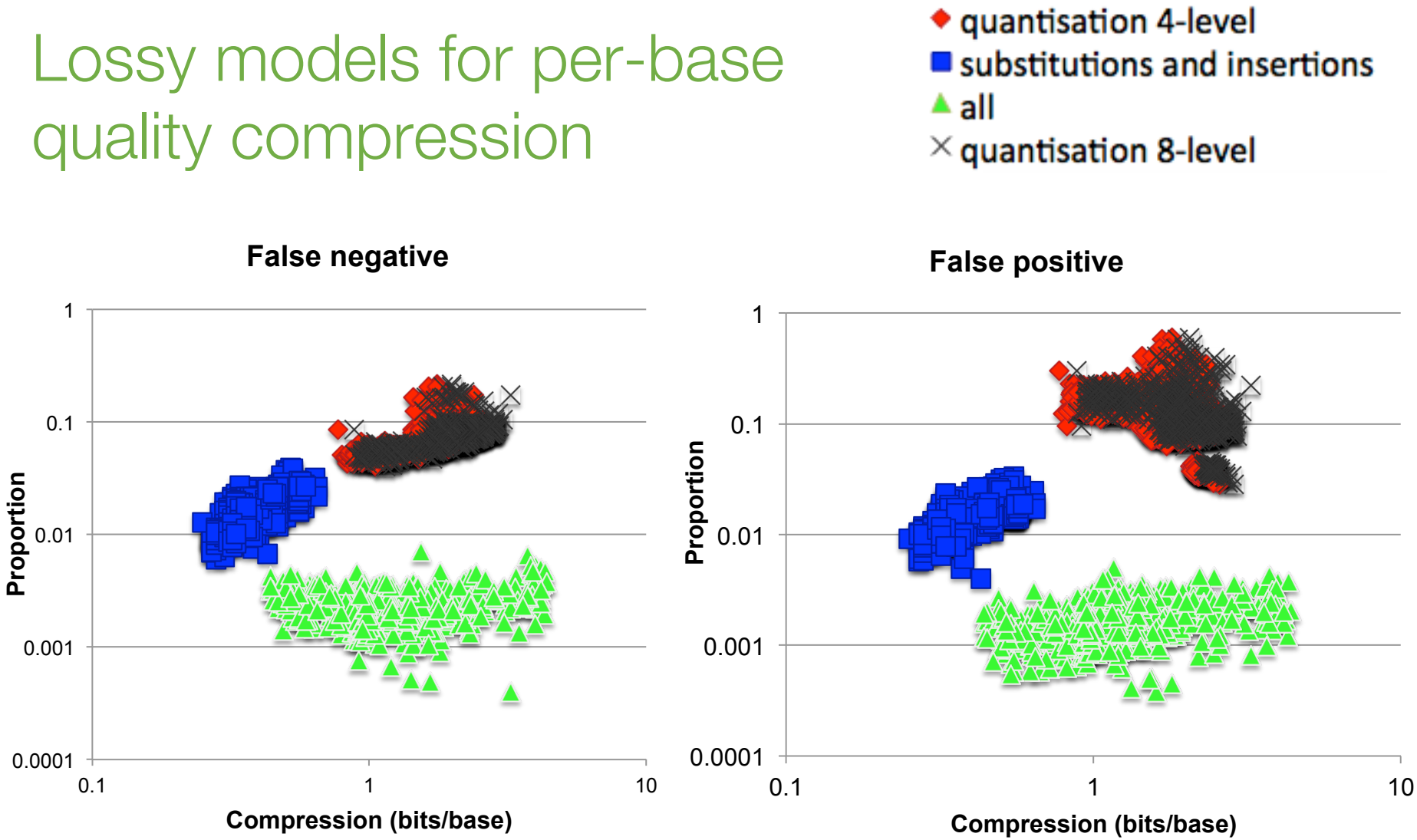
Data storage costs have become an appreciable proportion of total cost in the creation and analysis of DNA sequence data. Of particular concern is that the rate of increase in DNA sequencing is significantly outstripping the rate of increase in disk storage capacity. In this paper we present a new reference-based compression method that efficiently compresses DNA sequences for storage. Our approach works for re-sequencing experiments that target well-studied genomes. We align new sequences to a reference genome and then encode the differences between the new sequence and the reference genome for storage. Our compression method is most efficient when we allow controlled loss of data in the saving of quality information and unaligned sequences. With this new compression method we observe exponential efficiency gains as read lengths increase, and the magnitude of this efficiency gain can be controlled by changing the amount of quality information stored. Our compression method is tunable: the storage quality scores and unaligned sequences may be adjusted for different experiments to conserve information or to minimize storage costs, and provides one opportunity to address the threat that increasing DNA sequence volumes will overcome our ability to store the sequences.

Received September 2, 2010.
Accepted January 13, 2011.

Published in Advance January 18, 2011, doi: 10.1101/gr.114819.110

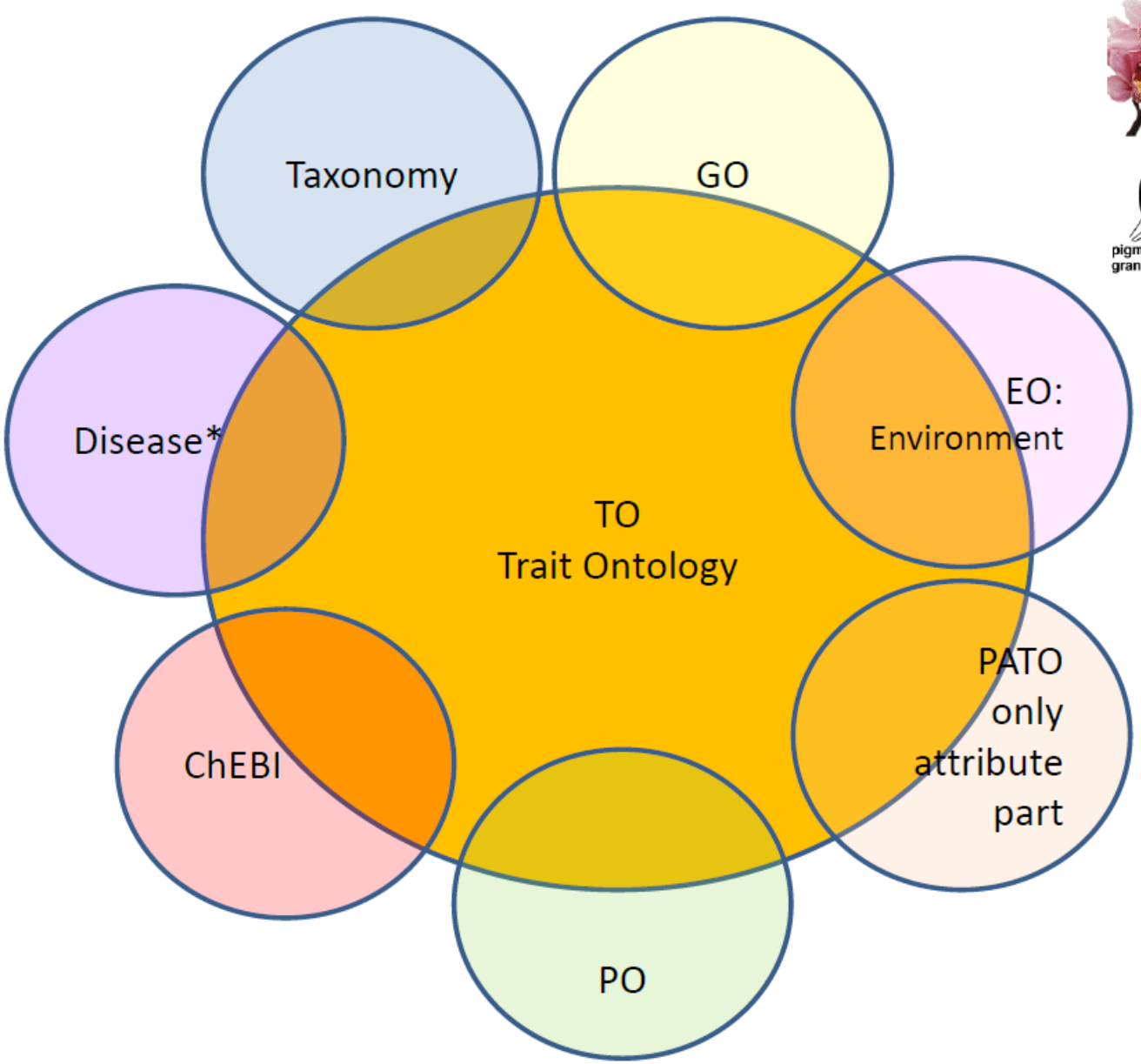
Genome Res. 2011.

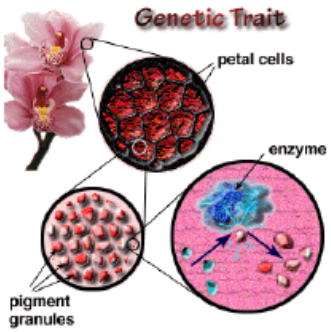
Lossy models for per-base quality compression



EMBL-EBI

Traits are the visible markers from multiple dimensions





Trait vs Phenotype

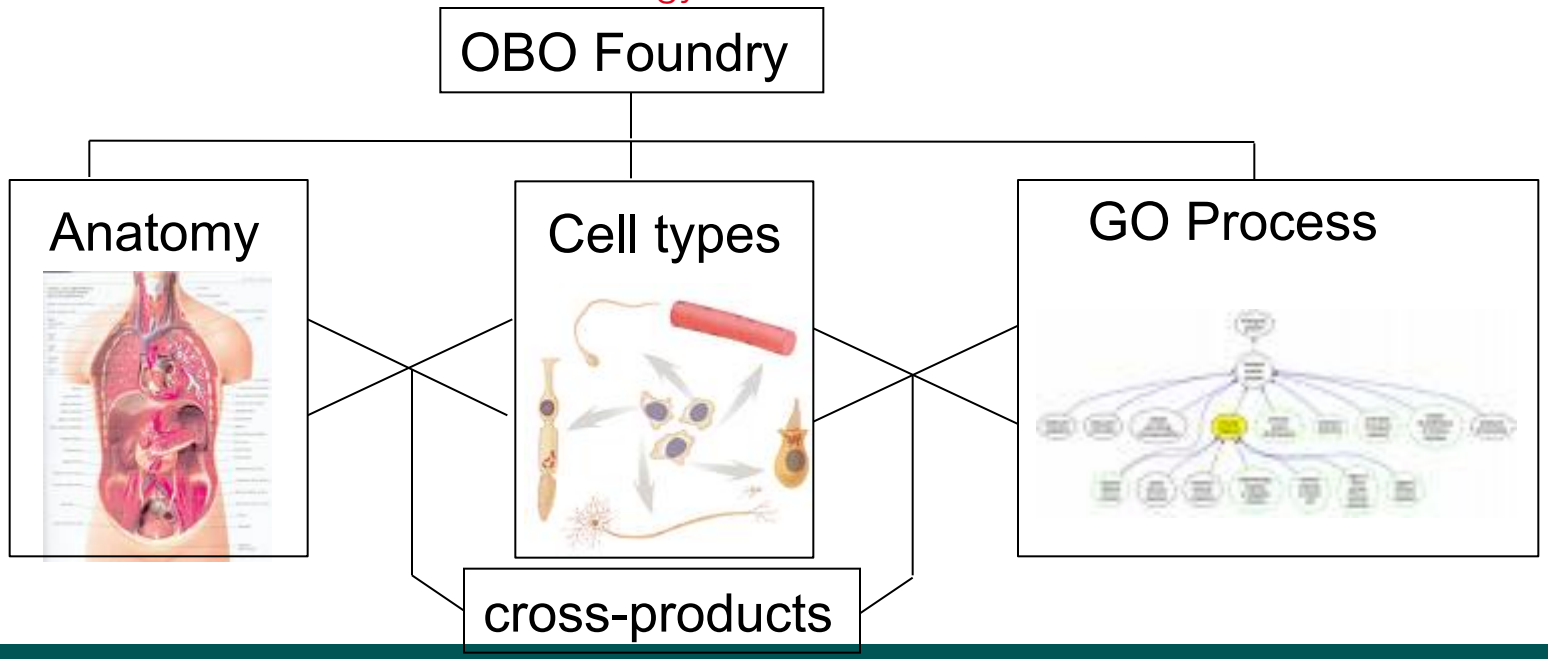
- Entity+Attribute = Trait [**observable**]
  E.g. *Leaf(PO) + color (PATO-A)* = *Leaf color(TO)*

- [Entity+Attribute+Value] = Phenotype [**observed**]
  e.g. *Leaf(PO) + color (PATO-A) + yellow (PATO-V)* = *Leaf color yellow* [EAV model:old]

- [Entity+(Attribute+Value)] = Phenotype [**observed**]
  e.g. *Leaf(PO) + color yellow (PATO-AV)* = *Leaf color yellow*
  [EA model:NEW]

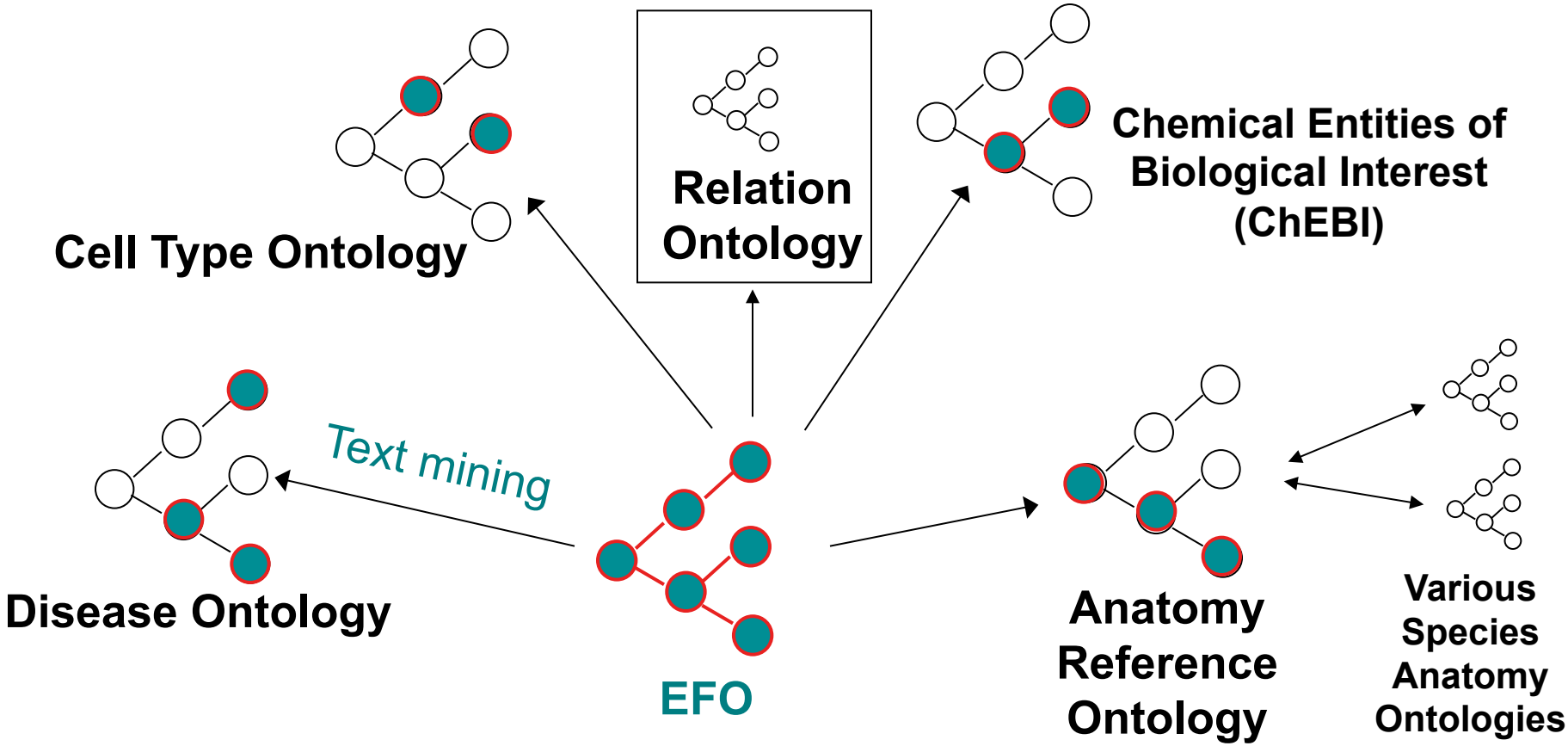
Different kinds of ontologies - Canonical

- Ontologies that represent *knowledge space*
  - Clear scope e.g. 'Normal processes'
  - And purpose – annotation of gene products
  - Applied for more e.g. Enrichment analysis and text mining
  - (Mostly) orthogonal – there is only one Cell Type Ontology
  - Foundational or Canonical Ontology

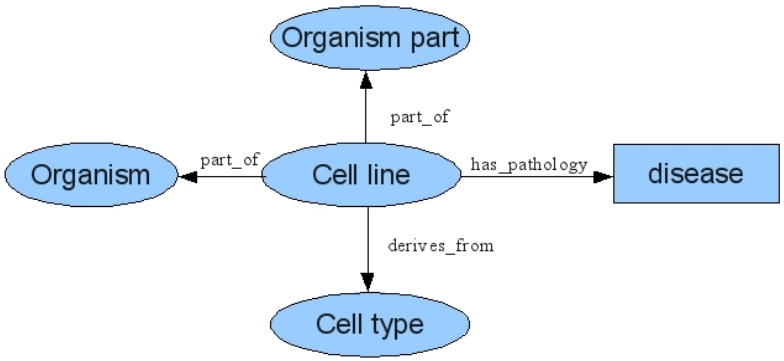


Building the Experimental Factor Ontology

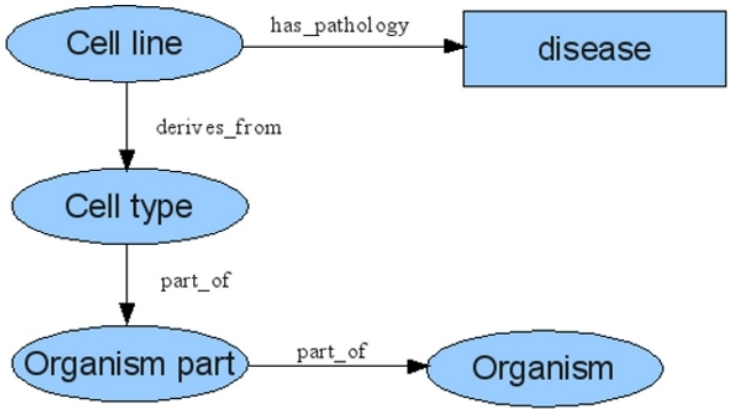
- Position of EFO in the 'bigger picture'
- Key is orthogonal coverage, reuse of existing resources and shared frameworks



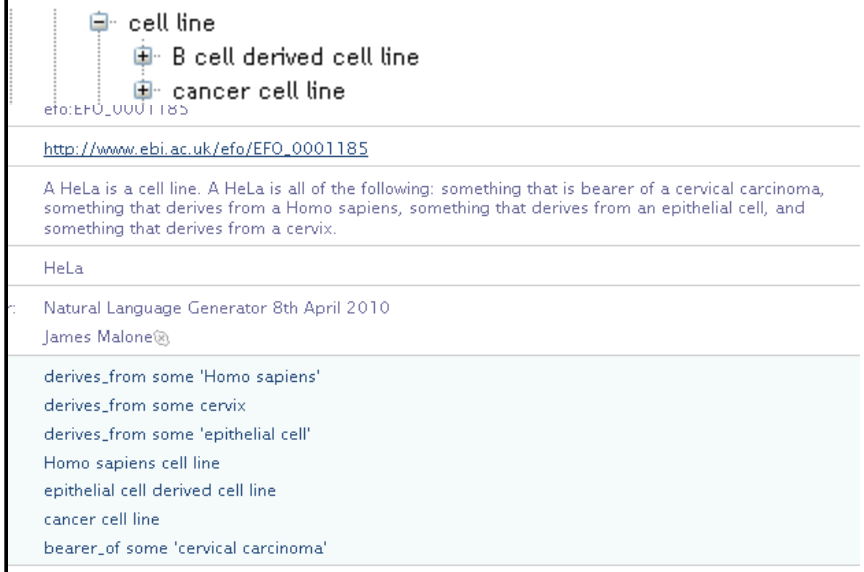
Capturing complex relations – Cell Lines



A. Experimental Factor Ontology cell line pattern



B. Coriell cell line ontology pattern



Annotation of traits in “man-machine readable” form is expensive

- Old style “manual” curation
- The vocabularies themselves still need development as well as use
- Important activities but not scalable
- Much valuable “legacy” data in non-standard representations

The future

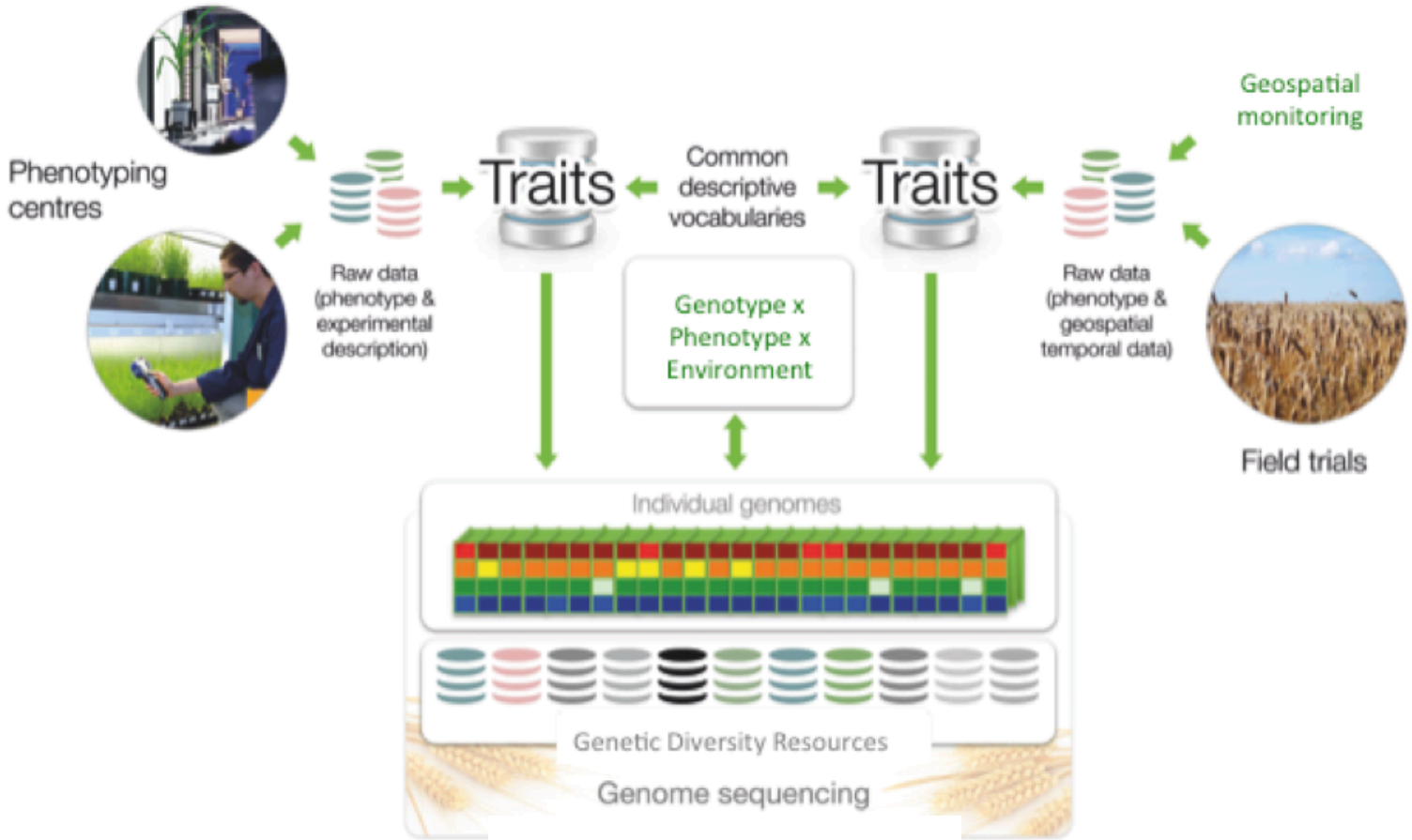
- More data
- More dimensions (more data types)
- More intersection (GxPxE)
- More distribution
- Smart queries

The future

- Universal identifiers for all populations/individuals/ samples of biological material
- High-quality, well-annotated reference genomes
- Reference catalogues of genomic variation
  - Solved problem for representation of structural variation
- Distributed archives of phenotypic data using standard vocabularies for high-level summary data

The money

- Plant data will (probably) be sufficiently small to be captured within whatever universal archives exist at EBI without requiring dedicated budgets
- Interpretative services will require additional budget to support them, which will only be secured through demonstration of demand and expected impact
- Elixir should bring additional national funding in a coordinated way across Europe


Phenotyping centres
Raw data (phenotype & experimental description)
Traits
Common descriptive vocabularies
Traits
Geospatial monitoring
Raw data (phenotype & geospatial temporal data)
Field trials
Genotype x Phenotype x Environment
Individual genomes
Genetic Diversity Resources
Genome sequencing

The money

- Plant data will (probably) be sufficiently small to be captured within whatever universal archives exist at EBI without requiring dedicated budgets
- Interpretative services will require additional budget to support them, which will only be secured through demonstration of demand and expected impact
- Elixir should bring additional national funding in a coordinated way across Europe

Breakout session 1: possible questions

- What are the big questions that plant scientists are going to be asking in the next 10 years?
- What your the key national priorities?
- How is data going to be used in answering these questions?
  - What are the different use cases of researchers, “-omics” centres, plant breeders, etc.?
- What data needs to be interoperable, and in what ways, and what data needs to be private?
- GxPxE – what does this mean to you?
  - What questions would you like to be able to ask of phenotypic data?
  - What is the interface between biological and other data (e.g. geospatial data) and how much should we be worrying about this?

Breakout session 2: possible questions

- What technologies will be needed to address the scientific drivers (databases, compute infrastructure, standards, etc.)?
- What problems are solved/funded?
- Where can plant science ride on solutions being developed elsewhere?
- Are there areas where where you see a major unfunded gap in infrastructure?
- If you could fund 3 components of an infrastructure, what would they be?
- Who else in this meeting would you like to work with to solve problems/ unlock potentialities? What would you do together?

The money

- Plant data will (probably) be sufficiently small to be captured within whatever universal archives exist at EBI without requiring dedicated budgets
- Interpretative services will require additional budget to support them, which will only be secured through demonstration of demand and expected impact
- Elixir should bring additional national funding in a coordinated way across Europe

European Commission

RESEARCH & INNOVATION

Infrastructures

Research Infrastructures

ESFRI

Council Conclusions on the implementation of the ESFRI roadmap

In its conclusions of 26th May 2014, the Council acknowledges the work done by ESFRI to identify priority projects which are mature enough to be under implementation in 2015-2016 and whose timely implementation is considered essential to extend the frontiers of knowledge in the fields concerned.

The Council also confirms the Member States' commitment to focus their available national resources on the respective prioritised projects they are financially participating in and invites the Commission, under Horizon 2020, to complement the Member States' own financial commitments through a one-time financial contribution for the priority projects, and to financially support the other projects (preparation and implementation) identified by ESFRI and listed in the Annex.

The Council also welcomes the plans of ESFRI to update its roadmap in 2015/2016.

- Council Conclusions of 26 May 2014 (see list of priority projects in Annex)
- Prioritisation of Support to ESFRI Projects for Implementation, ESFRI report, 7 April 2014 670 KB
- Letter from John Womersley, ESFRI Chair, to the Greek presidency 66 KB

ESFRI, the European Strategy Forum on Research Infrastructures, is a strategic instrument to develop the scientific integration of Europe and to strengthen its international outreach. The competitive and open access to high quality Research Infrastructures supports and benchmarks the quality of the activities of European scientists, and attracts the best researchers from around the world.

The **mission of ESFRI** is to support a coherent and strategy-led approach to policy-making on research infrastructures in Europe, and to facilitate multilateral initiatives leading to the better use and development of research infrastructures, at EU and international level.

ESFRI's delegates are nominated by the Research Ministers of the Member and Associate Countries, and include a representative of the Commission, working together to develop a joint vision and a common strategy. This strategy aims at overcoming the limits due to fragmentation of individual policies and provides Europe with the most up-to-date Research Infrastructures, responding to the rapidly evolving Science frontiers, advancing also the knowledge-based technologies and their extended use.

Since it was formed in 2002 at the behest of the European Council, ESFRI has witnessed significant advances towards unity and international impact in the field of research infrastructures. The publication of the first Roadmap for pan-European research infrastructures in 2006, and its update in 2008 was a key contributing factor, and several projects are now entering the realization phase. The Forum is determined to sustain the momentum in the implementation of the projects on the Roadmap, to expand the outreach to those scientific fields which are still evolving their conceptual

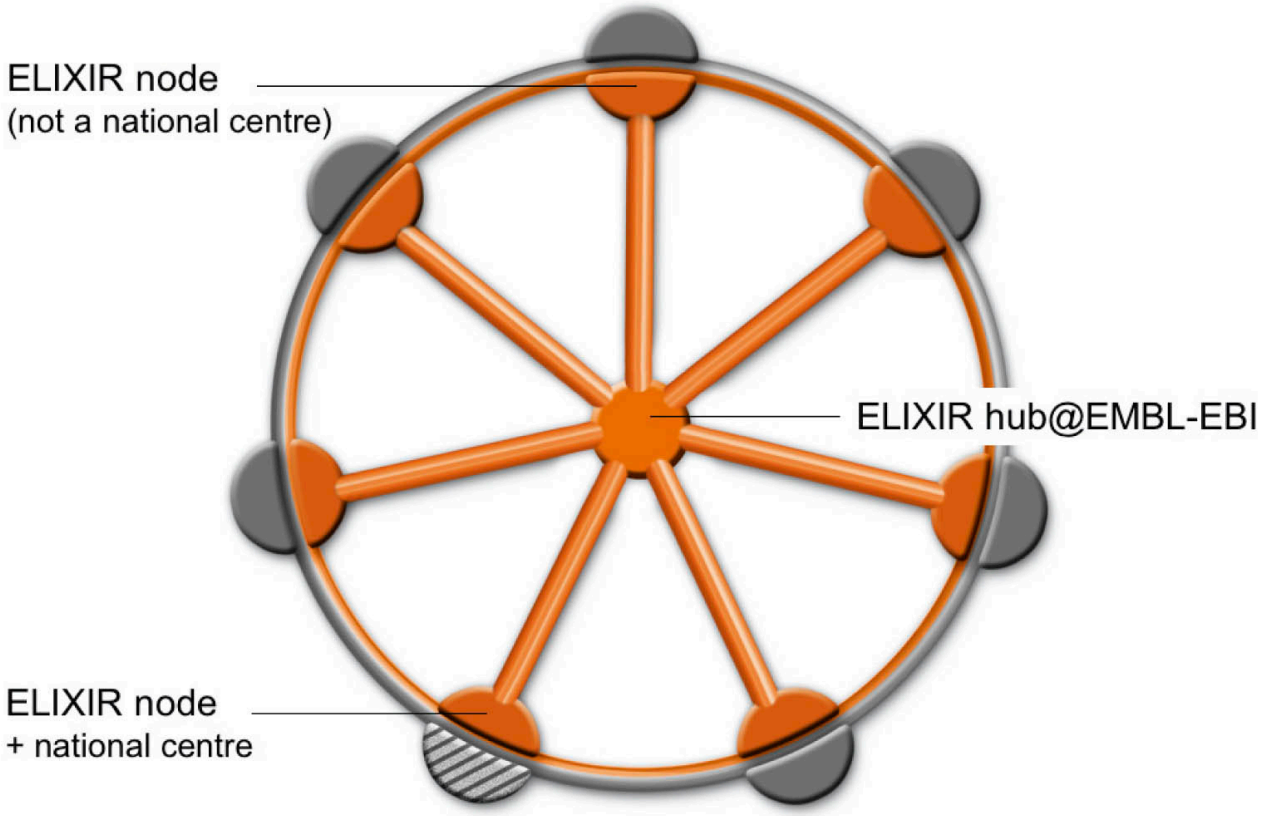
ESFRI

Highlights

- Indicators of pan-European relevance of research infrastructures 174 KB

Assessing the projects on the ESFRI roadmap: A high level expert group report 83.8 MB

A distributed pan-European infrastructure



EBI is a node, not the hub

- We are hosting the Elixir hub
  - But the hub has its own independent governance mechanisms, finance, etc.
- We expect the hub to utilise our expertise and services for bringing the nodes together to provide an integrated infrastructure for Europe
  - But EBI can't tell the other nodes what to do, and has no say in terms of which activities get funded

How will Elixir be funded?

- Individual nodes to be funded by national governments
- Hub to be funded from from a special EU infrastructure call
  - Elixir identified as one of 3 priority phase 1 ESFRI projects targeted for funding by the Commission

How will Elixir be funded?

- European Commission to identify I3s infrastructure projects as suitable for application by consortia with links to Elixir
  - Calls will be open to all, but ability to demonstrate links to Elixir (and other ESFRI infrastructures) where appropriate will clearly benefit chances of funding
  - I3 calls are a good chance for nodes to seek funds for collaboration, and even for infrastructures to seek funds for collaboration
  - Non-nodes need to demonstrate appropriate links to nodes

Models for nodes

- National centres of excellence/points of contact between national and European infrastructure
- Domain-specific experts for all-Europe
- Actual institutes versus distributed networks