Coordination of information platforms to support use of Plant Genetic Resources

Divseek International AGM
18th January 2022

Sarah Dyer
Non-vertebrate Genomics Team Leader, EMBL-EBI
sdyer@ebi.ac.uk
Many datatypes can be associated with Plant Genetic Resources

- Accession – passport and characterization, SMTA
- Sequence – reads, variants (from sequencing or chips)
- Phenotype – field trials, lab, images, omics data (gene expression, proteomics, metabolomics, metagenomics etc)
- Metadata [data that provides information about other data ]
  - e.g. how trial was conducted, how data was generated or analysed, microbiome context
- Generated by genebanks directly, collaborators or downstream users
Information platforms for Plant data – what can go where?

• Accessions – Genebank databases, Germinate, GRIN-Global, Genesys, EURISCO, GLIS, BioSamples
• Reads – ENA, SRA
• Variants – EVA
• Genome assemblies – ENA, NCBI, CoGe
• Phenotypes - Zenodo, FigShare, Dryad, Plant Genomics and Phenomics (PGP), GnpIS-Ephesis, BreeDB, Expression Atlas etc.

• Many options – consider not only ease of deposition, but ease of re-use
Benefits of data sharing

- Linking all related data to the original accession - Findable
  - Unique, persistent accession identifier e.g. DOIs / BioSamples
  - Capture the relationship between entities – has it been crossed/selfed/subsampled relative to the original genebank accession? Pedigree?
- Showing impact – track citations, DOIs for data sets
- Long-term data security in public repositories
- Added value of being integrated with other public data / resources
- Optimize reuse – FAIR (Findable, Accessible, Interoperable, Reusable)
Data Standards – to share, exchange, combine & understand data

• Before using someone else’s data you will want to understand:
  • the data standards used for recording the data (including file format)
  • the analysis methods used
  • the data quality

• Metadata should be rich, standardised and validated
Data Standards for Plant Genetic Resources

- Multi-Crop Passport Descriptors V.2
- Ontologies (crop, plant trait etc.)
- MIAPPE (plant phenotyping)
- BrAPI (a standard for communicating data)
- GODAN (Global Open Data for Agriculture and Nutrition)

Checklists and validation tools at point of data submission
  e.g. MIAPPE in BioSamples
AgBioData Working groups

• Data Standards for Genetic Variation
• Genotype-Phenotype Data
• Genome Assembly and Annotation Nomenclature
• Pan-genomes
Linking resources

Passport
Genotypes
Trials
Pedigree
Geographic
Climate

Genomic context: Genes, protein models, expression data, predicted effects
Ensembl – integrated genomic data

Wheat assemblies

Ensembl Plants hosts the latest wheat assembly from the IWGSC (RefSeq v1.0), including:
- The IWGSC RefSeq v1.1 gene annotation, with links to wheat-expression.com and KnetMiner.
- 14 wheat cultivars from the 10x genome project.
- Alignment of 98,270 high confidence genes from the TGAC v1 annotation.
- Axiom 35K, 820K SNP arrays from CerealsDB, including QTL links in selected cases and Linkage Disequilibrium display. See QTL example here.
- EMS-induced mutations from sequenced TILLING populations of Cadenza (coding regions and promoters).
- Inter-Homoeologous Variants (IHVs) between the A, B and D genome components.
- Chromosome specific KASP markers were added from the Nottingham BBSRC Wheat Research Centre.
- Whole genome alignments to rice, brachypodium and barley.
- Assembly-to-assembly mapping and gene ID mapping to the previous TGAC v1 assembly, archived at eq37-plants.ensembl.org.
- Polyploid view enabled, allowing users to view alignments among multiple wheat components simultaneously.
- Durum wheat 35K, 90K, 820K and TaBW290K variants
- Chromosome and centromere data can be viewed here.

Archive sites

Archive of release 49 of EnsemblPlants: eq49-plants.ensembl.org (Dec 2020)
Archive of release 45 of EnsemblPlants: eq45-plants.ensembl.org (Sep 2019)
Archive of release 40 of EnsemblPlants: eq40-plants.ensembl.org (July 2018)
Archive of release 37 of EnsemblPlants: eq37-plants.ensembl.org (Oct 2017)
Ensembl Plants browser

Community annotations

Variations from EVA or other sources

Genome

Variants

plants.ensembl.org
Variant:

**Cadenza0598.chr3D.379537481** SNP

**Most severe consequence:** Stop gained

**Alleles:** G/A

**Location:** Chromosome 3D:379537481 (forward strand) | VCF: 3D 379537481 Cadenza0598.chr3D.3795

**HGVS names:** This variant has 5 HGVS names - Show

**Marker/Primer:**
- PRIMER TYPE: chromosome_specific
- SNP TYPE: non-homoeologous
- EMS GENOTYPE: Homozygous
- TOTAL CONTIGS: 3
- MUTANT QUALITY: HIGH
- RESIDUAL HETEROGENEITY: No
- LINK TO PRIMER: Cadenza0598.chr3D.379537481

**External Links**

**Original source:** EMS-induced mutations from sequenced TILLING populations. Seeds can be ordered from UK SeedStore

**About this variant:** This variant overlaps 2 transcripts.

**Explore this variant**

- Genomic context
- Genes and regulation
- Flanking sequence
- Genotype frequency
- Phenotype data
- Sample genotypes
- Linkage disequilibrium
- Phylogenetic context
- Citations
Explore this variant
Population Genetics - Human gene HBB (haemoglobin subunit beta)

### Population genetics

#### 1000 Genomes Project Phase 3 allele frequencies

<table>
<thead>
<tr>
<th>Sub-populations</th>
<th>ALL</th>
<th>T: 97%</th>
<th>A: 3%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>APR</td>
<td>T: 90%</td>
<td>A: 10%</td>
</tr>
<tr>
<td></td>
<td>AMR</td>
<td>T: 90%</td>
<td>A: 1%</td>
</tr>
<tr>
<td></td>
<td>EAS</td>
<td>T: 100%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>EUR</td>
<td>T: 100%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SAS</td>
<td>T: 100%</td>
<td></td>
</tr>
</tbody>
</table>

### 1000 Genomes Project Phase 3 (32)

<table>
<thead>
<tr>
<th>Population</th>
<th>Allele: frequency (count)</th>
<th>Genotype: frequency (count)</th>
<th>Genotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>T: 0.973 (4871) A: 0.027 (137)</td>
<td>T/T: 0.945 (2367) A/T: 0.055 (137)</td>
<td>Show</td>
</tr>
<tr>
<td>APR</td>
<td>T: 0.906 (1190) A: 0.094 (132)</td>
<td>T/T: 0.800 (529) A/T: 0.200 (132)</td>
<td>Show</td>
</tr>
<tr>
<td>ACB</td>
<td>T: 0.950 (183) A: 0.047 (6)</td>
<td>T/T: 0.906 (87) A/T: 0.094 (9)</td>
<td>Show</td>
</tr>
<tr>
<td>ASW</td>
<td>T: 0.984 (120) A: 0.016 (2)</td>
<td>T/T: 0.967 (59) A/T: 0.033 (2)</td>
<td>Show</td>
</tr>
<tr>
<td>ESN</td>
<td>T: 0.879 (174) A: 0.121 (24)</td>
<td>T/T: 0.758 (75) A/T: 0.242 (24)</td>
<td>Show</td>
</tr>
<tr>
<td>GWD</td>
<td>T: 0.885 (200) A: 0.115 (26)</td>
<td>T/T: 0.770 (67) A/T: 0.230 (26)</td>
<td>Show</td>
</tr>
<tr>
<td>LWK</td>
<td>T: 0.899 (178) A: 0.101 (20)</td>
<td>T/T: 0.798 (79) A/T: 0.202 (20)</td>
<td>Show</td>
</tr>
<tr>
<td>MSL</td>
<td>T: 0.876 (149) A: 0.124 (21)</td>
<td>T/T: 0.753 (64) A/T: 0.247 (21)</td>
<td>Show</td>
</tr>
<tr>
<td>YRI</td>
<td>T: 0.861 (186) A: 0.139 (30)</td>
<td>T/T: 0.722 (76) A/T: 0.278 (30)</td>
<td>Show</td>
</tr>
</tbody>
</table>
AlphaFold
- v52 Arabidopsis

Transcript: AT4G16660.1
Description: heat shock protein 70 (Hsp 70) family protein [Source:TAIR;Acc:AT4G16660]
Location: Chromosome 4: 9,376,699-9,381,607 forward strand.
About this transcript: This transcript has 15 exons, is annotated with 27 domains and features, is associated with 884 variant alleles and maps to 47 cDNA probes.
Gene: This transcript is a product of gene AT4G16660

AlphaFold predicted model
Ensembl protein: AT4G16660.1

Deleterious
Tolerated

Variants:
SIFT

Model Confidence:
- Very high (pLDDT > 90)
- Confident (90 > pLDDT > 70)
- Low (70 > pLDDT > 50)
- Very low (pLDDT < 50)
Variant: Cadenza0598.chr3D.379537481 SNP

Most severe consequence: stop gained

Alleles: G/A

Location: Chromosome 3D:379537481 (forward strand) | VCF: 3D 379537481 Cadenza0598.chr3D.3795

HGVS names: This variant has 5 HGVS names - Show

Marker/Primer: PRIMER TYPE: chromosome_specific
SNP TYPE: non-homoeologous
EMS GENOTYPE: Homozygous
TOTAL CONTIGS: 3
MUTANT QUALITY: HIGH
RESIDUAL HETEROGENEITY: No
LINK TO PRIMER: Cadenza0598.chr3D.379537481

External Links:
Original source: EMS-induced mutations from sequenced TILLING populations up to dot (e.g Kronos3128).

About this variant: This variant overlaps 2 transcripts.

Explore this variant:

SeedStor2

Shopping Cart
Cadenza and Kronos TILLING Resources

The TILLING resource hosted on SeedStor is a specific set of lines that have known mutations and as such are best browsed via the genome browser tool at www.wheat-tilling.com. You can paste a list of lines to request from that genome browser into the box below or if you know the lines you wish to order then you can select them via the “Wheat TILLING” icon in the ‘Browse Collectors’ search module and import them below.

With this resource please note that the genome browser may show matches to lines that we have not been able to regenerate in sufficient amounts for distribution. Sadly these are often the most interesting lines precisely because they show mutations in some of the most interesting targets. We will contact you if we are unable to provide specific requested lines in the hope that their may be alternative lines for your target.

Please complete the fields below so that we can fulfill your request, please click here to see our privacy policy.
Conclusion

• Sharing data has benefits and can increase impact
• Deposit data into appropriate repositories
• Comply with (or develop!) data standards
• Add as much metadata as you can to ensure usability
• Enjoy your added impact!
Thanks!

If you’d like to learn more about Ensembl visit:
training.ensembl.org