

transPLANT milestone report

MS19 (work package 7): 15 reference genomes incorporated in transPLANT hub and submitted to comparative analysis

EMBL-EBI are providing access to many plant genomes described in the registry available for interactive and programmatic analysis through the Ensembl Plants (<http://plants.ensembl.org>) site. These data are shared with other partners through the use of DAS and other web services, and will be available for search through the transPLANT website (<http://www.transplantdb.org>) in the near future. Over the course of the project, we will work to develop further tools to promote interoperability among all the resources in development by project partners.

Ensembl, originally developed in the course of the Human Genome Project but subsequently applied to other domains, is a powerful tool suite for the analysis and display of genome scale data, and Ensembl Plants is the EBI's primary user interface for accessing plant data. We have used transPLANT funding to increase our capacity to include additional reference genomes incorporated in Ensembl Plants. In the second year of the grant, we have made 5 releases of Ensembl Plants, and incorporated the following additional genomes: six additional genomes: the bread wheat D-genome progenitor *Aegilops tauschii*, barley (*Hordeum vulgare*), banana (*Musa acuminata*), barrel clover (*Medicago truncatula*), potato (*Solanum tuberosum*) and the bread wheat A-genome precursor *Triticum urartu*. These new servers take the total number of species for which DAS servers are available through Ensembl Plants to 25, of which 16 have been made available through transPLANT funding. The number of species from the *Poaceae*, which include many important, closely related cereal crops, is now 12.

Genome, and protein-coding, sequences have been analysed comparatively using the Ensembl Compara functional genomics pipeline, which has 3 elements: a protein-based analysis, which infers evolutionary relationships after clustering and alignment (and which are performed over the domain of all plants) and a pairwise DNA-based analysis, performed using the alignment tools blastZ and lastZ. Synteny can be inferred from either primary analysis. The currently available analyses are given in figures 1 and 2.

Figure 1: Pairwise DNA-based comparative analyses undertaken using BLASTZ-net of LASTZ-net available through Ensembl Plants

Figure 2 Synteny analyses available through Ensembl Plants

	Oryza sativa japonica	N/A												
Oryza sativa japonica		N/A												
Arabidopsis thaliana	-		N/A											
Brachypodium distachyon	YES	-		N/A										
Oryza sativa indica	-	-	-		N/A									
Populus trichocarpa	-	YES	-		-		N/A							
Sorghum bicolor	YES	-	YES	-	-	-		N/A						
Vitis vinifera	-	-	-	-	YES	-		N/A						
Arabidopsis lyrata	-	YES	-	-	YES	-		YES		N/A				
Zea mays	YES	-	-	-	-	YES	-	-	-	N/A				
Solanum lycopersicum	-	-	-	-	-	-	-	-	-		N/A			
Solanum tuberosum	-	-	-	-	-	-	-	-	-		YES		N/A	
Hordeum vulgare	YES	-	YES	-	-	-	-	-	-	-	-	-	N/A	
	O.sat	A.tha	B.dis	O.ind	P.tri	S.bic	V.vin	A.lyr	Z.may	S.lyc	S.tub	H.vul		

The bread wheat genome *Triticum aestivum* is not yet sufficiently well assembled to be presented as a species in its own right in Ensembl, but ESTs, DNA contigs and polymorphic loci have been aligned to the syntenyic locations in barley and Brachypodium to enable the wheat sequences to be seen in an appropriate context. Bread wheat will be promoted to a first class status within Ensembl Plants when a more contiguous assembly becomes available.

The protein-centric analysis has (as of July 2013) placed 8,919,135 proteins from 23 plant genomes and selected outlying eukaryotic species (*Homo sapiens*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Ciona intestinalis* and *Saccharomyces cerevisiae*) into 43,771 clusters. The bread wheat precursor genomes, which are the only genomes in Ensembl Plants presently missing from these analyses, will be included in the protein-centric analysis with the release due September 2013. For each cluster, an alignment has been performed and an evolutionary history has been inferred, with putative speciation and gene duplication events inferred. **Figure 3** shows a graphical representation of a typical tree, one of the representations of the data that can be visualised through the Ensembl Plants site.

Figure 3: Visualisation of a gene tree containing the *Arabidopsis thaliana* PAD4 gene. The graphic illustrates a new feature: users can select certain annotations (from the Gene Ontology or InterPro), and highlight the genes within the tree that have been annotated with such terms (in this diagram, the highlighted genes have been annotated with the GO term “regulation of hydrogen peroxide metabolic process”).

