**transPLANT milestone report**

**MS21 (work package 8): Initial release of data for syntenic and duplicated regions**

To analyse syntenic relationships, intra-genomic duplications and rearrangements in and between a number of finished but also unfinished plant genomes, a pipeline for computation and analysis of orthologous/paralogous gene pairs and conserved gene order was developed.

The pipeline was designed to perform a variety of different sub-tasks: i) determining orthologs and paralogs, ii) calculating conserved synteny, and iii) calculating features such as syntenic quality index and dS/dN ratios.

*Identification of reference centric homology*

Homologous gene pairs were derived by searching for similarity of all protein coding sequences of a reference genome against all protein coding sequences of a target genome allowing for a n:1 relationship. To be classified as a homologous gene pair all high-scoring segment pairs (HSPs) were ordered corresponding to the query sequence.

*Identification of orthologous genes*

We used the 'best bidirectional hit' (BBH) method and criterion to calculate orthologous gene pairs. A gene pair is assigned to be orthologous if each protein coding sequence of the pair is found as best hit (reciprocal best hit criterion). The similarity search was carried out with TBlastX using default parameters and filtering. For each gene pair all high-scoring segment pairs (HSPs) in the protein coding sequence have to be arranged in the same order corresponding to the query sequence.

*Identification of homologous synteny blocks (HSBs)*

To construct pairwise HSBs between two genomes we followed the rules proposed by Murphy et al. (Science; July 2005, Vol. 309 no. 5734 pp. 613-617; DOI: 10.1126/science.1111387) using previously calculated orthologous gene pair sets. However, instead of whole genome synteny blocks (GRIMM-Synteny-based) we used a DAG based method to avoid overlapping synteny blocks and the need for orthologous genes shared over ALL plant genomes compared. This is especially critical since the number of genomes is increasing dramatically and plants contain a large portion of duplications and rearrangements.

*Calculation of the relative syntenic conservation index (RSCI)*

All orthologous genes which were part of a HSB were classified as syn+. Non- collinear genes in these segments were classified as syn-. The ratio of syn+ and syn- of a HSB reflects its syntenic quality. Using a sliding window across all HSBs a relative syntenic conservation index dRSCI can be calculated by normalizing the syn+ / syn- ratio according to the sum of the orthologous gene pairs.

*Identification of paralogous genes*

Protein coding sequences were pairwise aligned using gapped BLAST with filtering for low sequence complexity and default parameters. To be classified as a paralogous gene all high-scoring segment pairs (HSPs) in the target sequence have to be arranged in the same order corresponding to the query sequence, the total length of the aligned target sequence needs to cover more than 60% of the query sequence and the overall similarity is higher than 70%. All the criteria need to be met in a symmetrical manner (A vs. B and B vs. A).

*Calculation of global and local Ka/Ks ratios*

Divergence (D) was calculated according to the proportion distance (p-distance) for each orthologous pair and all pairs with a divergence >5% were excluded from further analysis to prevent inaccurate alignments. The nucleotide sequences were translated with Transeq (EMBOSS: <http://emboss.sourceforge.net/apps>) and were aligned with CLUSTALW. Ka/Ks ratios were calculated using YN00. In addition a local Ka/Ks ratio was calculated using a sliding window of 30 amino acids across the entire alignment length.

*Calculating the Tandem Gene Array Index*

The tandem gene array index was calculated using a genome-wide similarity search with TBlastX against the genome itself. The index at position x is defined as the number of similar genes upstream and downstream within a certain window divided by the total sum of the genes in the whole window. Typically the window is defined by the number of consecutive genes. Here we used a window size of 50 genes upstream and the same number of genes downstream.

As of February 2014, the pipeline described above has been applied to the genomes and gene predictions of the following species: *B.* *distachyon*, *S. bicolor (sorghum)*, *O. sativa* (rice), *H. vulgare* (barley), *Ae. tauschii* (progenitor of the wheat D sub-genome) as well as *P. dactylifera* (datepalm) versus *E. oleifera* (oilpalm). The genomes of *S. lycopersicum* (tomato), *S. tuberosum* (potato), *C. annuum* (chili pepper), *Ae. urartu* and *Z.mays* (maize) are currently being prepared for computation and will be available shortly.

The results from the computation of orthologous gene pairs and conserved gene order for selected species combinations of interest were released as download files from:

**http://seacow.helmholtz-muenchen.de/mips\_cviewer/data\_dumps.html**

Files are in tab-delimited format and define syntenic relationships between two plant species and provide a starting point for the downstream identification and analysis of duplicated regions, chromosomal rearrangements etc.

To assist the identification of duplicated regions, inversions, deletions and chromosomal rearrangements a **visualization tool,** **CrowsNest,** was developed and populated with the results from the computation pipeline.

CrowsNest is a whole genome interactive comparative mapping and visualization tool comparing genetic, physical and hierarchical (fingerprinted contigs) maps in the plant kingdom. Syntenic relationships can be investigated in a hierarchical manner from macro level views down to more detailed micro level views.

By incorporating and comparing data from both model (such as *B. distachyon*) and crop (such as barley) plants, knowledge transfer regarding genome structure, gene function and conservation between related plant species is supported and facilitated.

CrowsNest can be accessed at:

**http://mips.helmholtz-muenchen.de/plant/crowsNest/index.jsp**

and currently incorporates data from *B. distachyon*, *S. bicolor*, *O. sativa* (rice), *H. vulgare* (barley), *Ae. tauschii* (progenitor of the wheat D sub-genome) as well as *P. dactylifera* (datepalm) versus *E. oleifera* (oilpalm).

****

**Figure 1**: The syntenic regions between the organisms can be browsed in the CrowsNest tool in a hierarchical way from macro-synteny (3A-3C) down to micro-synteny views (3F). Example is given for the plant genomes of *B. distachyon* and *S. bicolor*.