Galaxy / PlantTribes

Greg consulted with Dr. Claude DePamphillis in the Department of Biology to build the Galaxy / PlantTribes analysis environment. PlantTribes is a collection of automated modular analysis pipelines that utilize objective classifications of complete protein sequences from sequenced plant genomes to perform comparative evolutionary studies.

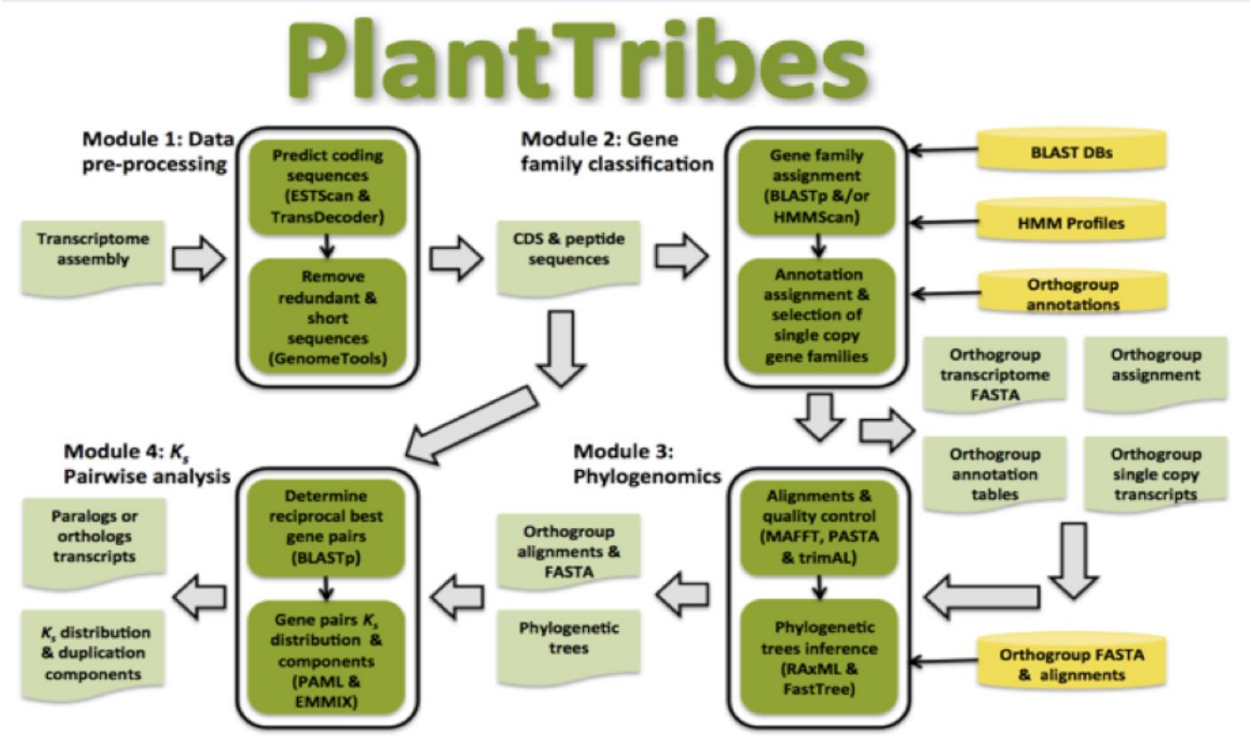

Figure 12: PlantTribes analysis process flow

The Galaxy PlantTribes tools currently include the following, but more are coming since work continues on this environment.

- **AssemblyPostProcessor** - post-processes de novo assembled transcripts into putative coding sequences and their corresponding amino acid translations and optionally assigns transcripts to circumscribed gene families (orthogroups). After transcripts have been assigned to gene families, overlapping contigs can be identified and merged to reduce fragmentation in the de novo assembly.
- **GeneFamilyClassifier** - classifies gene coding sequences either produced by the AssemblyPostProcessor tool or from an external source into pre-computed orthologous gene family clusters (orthogroups) of a PlantTribes

scaffold. Classified sequences are then assigned with the corresponding orthogroups' metadata that includes gene counts of backbone taxa, super clusters (super orthogoups) at multiple stringencies, and functional annotations from sources such as Gene Ontology (GO), InterPro protein domains, TAIR, UniProtKB/TrEMBL, and UniProtKB/Swiss-Prot. Additionally, sequences belonging to single/low-copy gene families that are mainly utilized in species tree inference can be determined.

- **GeneFamilyIntegrator** - integrates PlantTribes scaffold orthogroup backbone gene models with gene coding sequences classified into the scaffold by the GeneFamilyClassifier.
- **GeneFamilyAligner** - estimates protein and codon multiple sequence alignments of integrated orthologous gene family fasta files produced by the GeneFamilyIntegrator.
- **GeneFamilyPhylogenyBuilder** - performs gene family phylogenetic inference of multiple sequence alignments produced by the GeneFamilyAligner.
- **KaKsAnalysis** - estimates paralogous and orthologous pairwise synonymous (Ks) and non-synonymous (Ka) substitution rates for a set of gene coding sequences either produced by the AssemblyPostProcessor tool or from an external source. Optionally, the resulting set of estimated Ks values can be clustered into components using a mixture of multivariate normal distributions to identify significant duplication event(s) in a species or a pair of species.
- **KsDistribution** - uses the analysis results produced by the KaKsAnalysis tool to plot the distribution of synonymous substitution (Ks) rates and fit the estimated significant normal mixtures component(s) onto the distribution.

The Galaxy PlantTribes analysis workbench performs the following functions.

- Post-processes de novo assembly transcripts into putative coding sequences and their corresponding amino acid translations
- Locally assembles targeted gene families
- Estimates paralogous/orthologous pairwise synonymous/non-synonymous substitution rates for a set of gene sequences
- Classifies gene sequences into pre-computed orthologous plant gene family clusters
- Builds gene family multiple sequence alignments and their corresponding phylogenies.

A user provides de novo assembly transcripts, and PlantTribes produces:

- Predicted coding sequences and their corresponding translations
- A table of pairwise synonymous/non-synonymous substitution rates for reciprocal best blast transcript pairs
- Results of significant duplication components in the distribution of Ks (synonymous substitutions) values
- A summary table for transcripts classified into orthologous plant gene family clusters with their corresponding functional annotations
- Gene family amino acid and nucleotide fasta sequences
- Multiple sequence alignments
- Inferred maximum likelihood phylogenies

Optionally, a user can provide externally predicted coding sequences and their corresponding amino acid translations derived from a transcriptome assembly or gene predictions from a sequenced genome.