PlantTribes: Galaxy tools for comparative gene family analysis in plant genomes

Eric Kenneth Wafula1, Jing Van Kuster2, Joshua P. Der1, Loren Honaza1, Saravanaraj Ayyampalayam3, Norman Wickett4, Jim Leebens-Mack5 and Claude dePamphilis6

Penn State University, University Park, PA, – California State University, Fullerton, Fullerton, CA, - The University of Georgia, Athens, GA, - Chicago Botanic Garden, Glencoe, IL, - University of Georgia, Athens, GA

Abstract

Galaxy PlantTribes is a collection of automated modular analysis pipelines that utilize objective classifications of complete protein sequences from sequenced genomes for comparative and evolutionary analyses of genome-scale gene families and transcriptomes. It post-processes de novo assembly transcripts into putative coding sequences and their corresponding amino acid translations, estimates paralogous/orthologous pairwise synonymous/synonymous substitution rates for a set of gene sequences, classifies gene sequences into pre-computed orthologous plant gene family clusters, and builds gene family multiple sequence alignments and their corresponding phylogenies. A user provides de novo assembly transcripts and Galaxy PlantTribes produces: (1) predicted coding sequences and their corresponding translations, (2) a table of pairwise synonymous/synonymous substitution rates for either orthologous or paralogous transcript pairs, (3) results of significant duplication components in the distribution of its (synonymous substitutions) values, (4) a summary table for transcripts classified into orthologous plant gene family clusters with their corresponding functional annotations, (4) gene family amino acid and nucleotide fasta sequences, (6) multiple sequence alignments, and (5) inferred maximum likelihood phylogenies. Optionally, a user can provide an external gene family scaffold and/or externally classified sequences and their corresponding amino acid translations and optimally aligns 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.

Workflow

Assembly

GeneFamilyClassifier classifies gene coding sequences either produced by the AssemblyPostProcessor 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 orthogroup") at multiple stringencies, and functional annotations.

Alignment Estimation and QC

GeneFamilyAligner estimates protein and codon multiple sequence alignments of integrated orthologous gene family fasta files produced by the GeneFamilyIntegrator (not shown). The GeneFamilyIntegrator integrates PlantTribes scaffolded orthogroup backbone gene models with gene coding sequences classified into the scaffold by the GeneFamilyClassifier.

Phylogenetic Inference

GeneFamilyPhylogenyBuilder performs gene family phylogenetic inference of multiple sequence alignments produced by the GeneFamilyAligner. Orthogroup maximum likelihood (ML) phylogenetic trees are inferred using either RAxML or FastTree algorithms. Trees are rooted using the most distant taxon present in the orthogroup backbone taxa if root rooting order list is not provided.

Phylogenetic Inference

Orthogroup fasta multiple sequence alignments produced by the GeneFamilyAligner can be visualized in Galaxy using the MSAViewer plugin and manually edited with Jalview.

Selection Analysis

KatoKa Анаlysis estimates paralogous and orthologous pairwise Ks values can be clustered into components using a mixture of multivariate normal distributions to identify significant duplication events in a species or a group of species. The Kato Ka Distribution (not shown) can be to plot the distribution of its rates and fit the estimated significant normal mixtures components into the distribution.

Whole Genome Duplication Estimation

The resulting set of estimated paralogous and orthologous pairwise Ks values can be clustered into components using a mixture of multivariate normal distributions to identify significant duplication events in a species or a group of species. The Kato Ka Distribution (not shown) can be to plot the distribution of its rates and fit the estimated significant normal mixtures components into the distribution.

Application

1. Targeted gene family assembly
2. Improving genome annotation quality
3. Transcriptome coverage and functional annotation
4. Species tree inference using single copy genes
5. Gene tree – species tree reconciliation
6. Timing of gene duplications and polyploidy
7. Ancestral gene content reconstruction
8. Gene family expansions/contractions
9. Timing of new gene function evolution
10. Studies of horizontal transfers among species

Funding and support: Funding for this work was provided by the NSF Plant Genome Research Program through the Amborella Genome Sequencing Project (NSF Grant 0922742), Parasitic Plant Genome Project (NSF Grant 0701748), and Immunity in Perennial Crops (NSF Grant 1546863) with additional support from the One Thousand Plants (1000) transcriptome project.