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

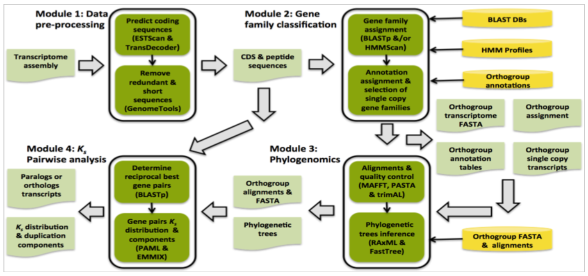
Eric Kenneth Wafula[1], Greg Von Kuster[1], Joshua P. Der[2], Loren Honaas[1], Saravanaraj Ayyampalayam[3], Norman Wickett[4], Jim Leebens-Mack[5] and Claude dePamphilis[1]

[1]Penn State University, University Park, PA, [2]California State University, Fullerton, Fullerton, CA, [3]The University of Georgia, Athens, GA,
[4]Chicago Botanic Garden, Glencoe, IL, 5University of Georgia, Athens, GA

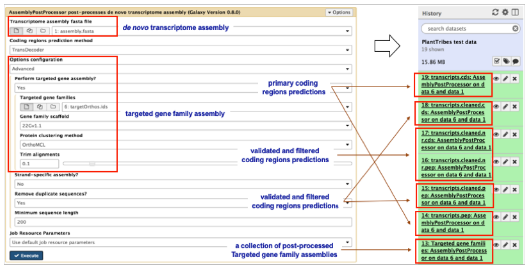
Abstract

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/non-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/non-synonymous substitution rates for either orthologous or paralogous transcript pairs, (3) results of significant duplication components in the distribution of Ks (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 predicted coding sequences derived from a transcriptome assembly or gene predictions from a sequenced genome. Galaxy PlantTribes is freely available on the Galaxy main portal (https://usegalaxy.org). In addition, the standalone version of the pipeline is available for download on GitHub (https://github.com/dePamphilis/PlantTribes) as a command-line interface for batch processing of many datasets.

Workflow


Module 1: Data pre-processing
Transcriptome assembly
Predict coding sequences (ESTScan & TransDecoder)
Remove redundant & short sequences (GenomeTools)
Module 2: Gene family classification
CDS & peptide sequences
Gene family assignment (BLASTp &/or HMMScan)
Annotation assignment & selection of single copy gene families
BLAST DBs
HMM Profiles
Orthogroup annotations
Orthogroup transcriptome FASTA
Orthogroup assignment
Orthogroup annotation tables
Orthogroup single copy transcripts
Module 3: Phylogenomics
Alignments & quality control (MAFFT, PASTA & trimAL)
Phylogenetic trees inference (RAxML & FastTree)
Orthogroup FASTA & alignments
Orthogroup alignments & FASTA
Phylogenetic trees
Module 4: K_s Pairwise analysis
Determine reciprocal best gene pairs (BLASTp)
Gene pairs K_s distribution & components (PAML & EMMIX)
Paralogs or orthologs transcripts
K_s distribution & duplication components

Post Assembly QC


AssemblyPostProcessor post-processes de novo transcriptome assembly (Galaxy Version 0.8.0)

Transcriptome assembly fasta file: 1: assembly.fasta

Coding regions prediction method: TransDecoder

Options configuration: Advanced

Perform targeted gene assembly?: Yes

Targeted gene families: 6: targetOrthos.ids

Gene family scaffold: 22Gv1.1

Protein clustering method: OrthoMCL

Trim alignments: 0.1

Strand-specific assembly?: No

Remove duplicate sequences?: Yes

Minimum sequence length: 200

Job Resource Parameters: Use default job resource parameters

✓ Execute

History: PlantTribes test data, 19 shown, 15.86 MB

19: transcripts.cds: AssemblyPostProcessor on data 6 and data 1

18: transcripts.cleaned.cds: AssemblyPostProcessor on data 6 and data 1

17: transcripts.cleaned.nr.cds: AssemblyPostProcessor on data 6 and data 1

16: transcripts.cleaned.nr.pep: AssemblyPostProcessor on data 6 and data 1

15: transcripts.cleaned.pep: AssemblyPostProcessor on data 6 and data 1

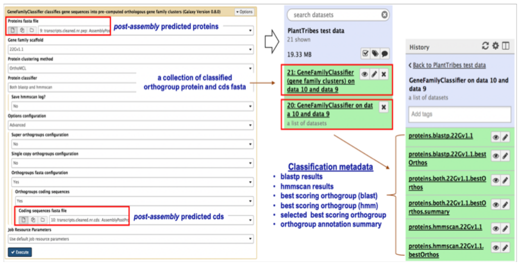
14: transcripts.pep: AssemblyPostProcessor on data 6 and data 1

13: Targeted gene families: AssemblyPostProcessor on data 6 and data 1

de novo transcriptome assembly
primary coding regions predictions
targeted gene family assembly
validated and filtered coding regions predictions
validated and filtered coding regions predictions
a collection of post-processed Targeted gene family assemblies

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.

Assembly Classification


GeneFamilyClassifier classifies gene sequences into pre-computed orthologous gene family clusters (Galaxy Version 0.8.0)

Proteins fasta file: 9: transcripts.cleaned.nr.pep: AssemblyPos

Gene family scaffold: 22Gv1.1

Protein clustering method: OrthoMCL

Protein classifier: Both blastp and hmmscan

Save hmmscan log?: No

Options configuration: Advanced

Super orthogroups configuration: No

Single copy orthogroups configuration: No

Orthogroups fasta configuration: Yes

Orthogroups coding sequences: Yes

Coding sequences fasta file: 10: transcripts.cleaned.nr.cds: AssemblyPostPro

Job Resource Parameters: Use default job resource parameters

✓ Execute

History: PlantTribes test data, 21 shown, 19.33 MB

21: GeneFamilyClassifier (gene family clusters) on data 10 and data 9

20: GeneFamilyClassifier on data 10 and data 9 — a list of datasets

‹ Back to PlantTribes test data — GeneFamilyClassifier on data 10 and data 9 — a list of datasets

proteins.blastp.22Gv1.1

proteins.blastp.22Gv1.1.bestOrthos

proteins.both.22Gv1.1.bestOrthos

proteins.both.22Gv1.1.bestOrthos.summary

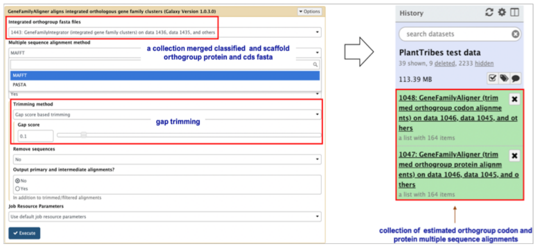
proteins.hmmscan.22Gv1.1

proteins.hmmscan.22Gv1.1.bestOrthos

post-assembly predicted proteins
a collection of classified orthogroup protein and cds fasta
post-assembly predicted cds
Classification metadata
blastp results
hmmscan results
best scoring orthogroup (blast)
best scoring orthogroup (hmm)
selected best scoring orthogroup
orthogroup annotation summary

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 orthogoups") at multiple stringencies, and functional annotations.

Alignment Estimation and QC


GeneFamilyAligner aligns integrated orthologous gene family clusters (Galaxy Version 1.0.3.0)

Integrated orthogroup fasta files: 1443: GeneFamilyIntegrator (integrated gene family clusters) on data 1436, data 1435, and others

Multiple sequence alignment method: MAFFT (MAFFT, PASTA)

Trimming method: Gap score based trimming

Gap score: 0.1

Remove sequences: No

Output primary and intermediate alignments?: ● No ○ Yes

In addition to trimmed/filtered alignments

Job Resource Parameters: Use default job resource parameters

✓ Execute

History: PlantTribes test data, 39 shown, 9 deleted, 2233 hidden, 113.39 MB

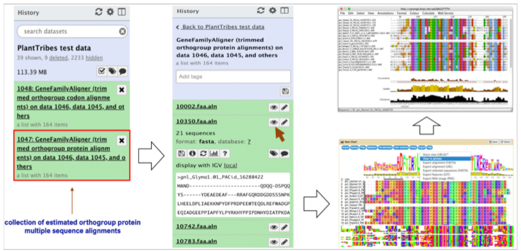
1048: GeneFamilyAligner (trimmed orthogroup codon alignments) on data 1046, data 1045, and others — a list with 164 items

1047: GeneFamilyAligner (trimmed orthogroup protein alignments) on data 1046, data 1045, and others — a list with 164 items

a collection merged classified and scaffold orthogroup protein and cds fasta
gap trimming
collection of estimated orthogroup codon and protein multiple sequence alignments

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 scaffold orthogroup backbone gene models with gene coding sequences classified into the scaffold by the **GeneFamilyClassifier**.

Alignment Visualization and Editing


History: PlantTribes test data, 39 shown, 9 deleted, 2233 hidden, 113.39 MB

1048: GeneFamilyAligner (trimmed orthogroup codon alignments) on data 1046, data 1045, and others — a list with 164 items

1047: GeneFamilyAligner (trimmed orthogroup protein alignments) on data 1046, data 1045, and others — a list with 164 items

‹ Back to PlantTribes test data — GeneFamilyAligner (trimmed orthogroup protein alignments) on data 1046, data 1045, and others — a list with 164 items

10002.faa.aln

10350.faa.aln — 21 sequences, format: fasta, database: ? — display with IGV local

```
>gnl_Glyma1.01_PACid_16288422
MAND----------------------QDQQ-DSPQQ
YS------YDEAEDEAF---RRAFGQRDDGDDSSSNPK
LHEELDPLIAEKKNPYDFPRDPEEWTEQDLREFWADGP
EQIADGEEPPIAPFYLPYRKHYPPIPDNHYDIATPKDA
```

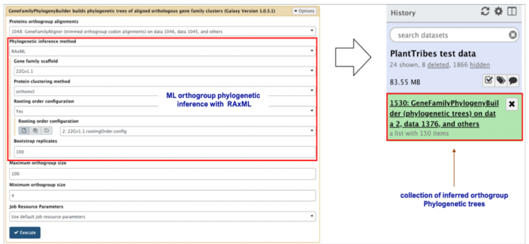
10742.faa.aln

10783.faa.aln

collection of estimated orthogroup protein multiple sequence alignments

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

Phylogenetic Inference


GeneFamilyPhylogenyBuilder builds phylogenetic trees of aligned orthologous gene family clusters (Galaxy Version 1.0.3.1)

Proteins orthogroup alignments: 1048: GeneFamilyAligner (trimmed orthogroup codon alignments) on data 1046, data 1045, and others

Phylogenetic inference method: RAxML

Gene family scaffold: 22Gv1.1

Protein clustering method: orthomcl

Rooting order configuration: Yes

Rooting order configuration: 2: 22Gv1.1.rootingOrder.config

Bootstrap replicates: 100

Maximum orthogroup size: 100

Minimum orthogroup size: 4

Job Resource Parameters: Use default job resource parameters

✓ Execute

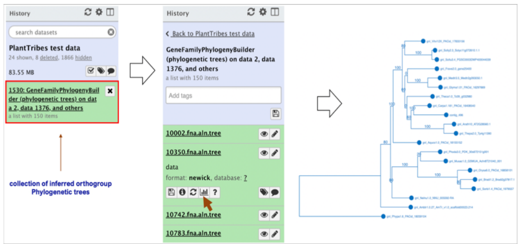
History: PlantTribes test data, 24 shown, 8 deleted, 1866 hidden, 83.55 MB

1530: GeneFamilyPhylogenyBuilder (phylogenetic trees) on data 2, data 1376, and others — a list with 150 items

ML orthogroup phylogenetic inference with RAxML
collection of inferred orthogroup Phylogenetic trees

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 Visualization and Editing


History: PlantTribes test data, 24 shown, 8 deleted, 1866 hidden, 83.55 MB

1530: GeneFamilyPhylogenyBuilder (phylogenetic trees) on data 2, data 1376, and others — a list with 150 items

‹ Back to PlantTribes test data — GeneFamilyPhylogenyBuilder (phylogenetic trees) on data 2, data 1376, and others — a list with 150 items

10002.fna.aln.tree

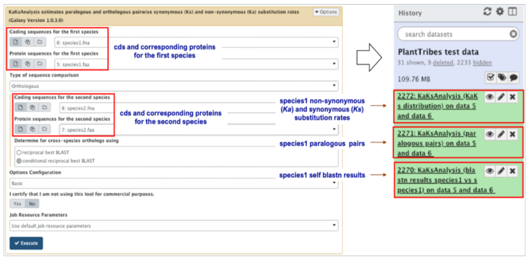
10350.fna.aln.tree — data, format: newick, database: ?

10742.fna.aln.tree

10783.fna.aln.tree

Orthogroup newick phylogentic trees produced by the **GeneFamilyPhylogenyBuilder** can be visualized in Galaxy using either the Phylogenetic Tree Visualization plugin or the PHYLOViZ plugin. The Phylogenetic Tree Visualization plugin provides several options of tree rendering.

Selection Analysis


KaKsAnalysis estimates paralogous and orthologous pairwise synonymous (Ks) and non-synonymous (Ka) substitution rates (Galaxy Version 1.0.3.0)

Coding sequences for the first species: 6: species1.fna

Protein sequences for the first species: 5: species1.faa

Type of sequence comparison: Orthologous

Coding sequences for the second species: 8: species2.fna

Protein sequences for the second species: 7: species2.faa

Determine for cross-species orthologs using: ○ reciprocal best BLAST ● conditional reciprocal best BLAST

Options Configuration: Basic

I certify that I am not using this tool for commercial purposes. Yes No

Job Resource Parameters: Use default job resource parameters

✓ Execute

History: PlantTribes test data, 31 shown, 9 deleted, 2233 hidden, 109.76 MB

2272: KaKsAnalysis (KaKs distribution) on data 5 and data 6

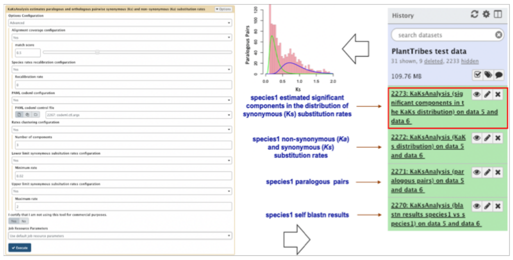
2271: KaKsAnalysis (paralogous pairs) on data 5 and data 6

2270: KaKsAnalysis (blastn results species1 vs species1) on data 5 and data 6

cds and corresponding proteins for the first species
cds and corresponding proteins for the second species
species1 non-synonymous (*Ka*) and synonymous (*Ks*) substitution rates
species1 paralogous pairs
species1 self blastn results

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** or from an external source.

Whole Genome Duplication Estimation


KaKsAnalysis estimates paralogous and orthologous pairwise synonymous (Ks) and non-synonymous (Ka) substitution rates

Options Configuration: Advanced

Alignment coverage configuration: Yes — match score: 0.5

Species rates recalibration configuration: Yes — Recalibration rate: 0

PAML codeml configuration: Yes — PAML codeml control file: 2267: codeml.ctl.args

Rates clustering configuration: Yes — Number of components: 3

Lower limit synonymous substitution rates configuration: Yes — Minimum rate: 0.02

Upper limit synonymous substitution rates configuration: Yes — Maximum rate: 2

I certify that I am not using this tool for commercial purposes. Yes No

Job Resource Parameters: Use default job resource parameters

✓ Execute

History: PlantTribes test data, 31 shown, 9 deleted, 2233 hidden, 109.76 MB

2273: KaKsAnalysis (significant components in the KaKs distribution) on data 5 and data 6

2272: KaKsAnalysis (KaKs distribution) on data 5 and data 6

2271: KaKsAnalysis (paralogous pairs) on data 5 and data 6

2270: KaKsAnalysis (blastn results species1 vs species1) on data 5 and data 6

Paralogous Pairs
0
20
40
60
80
100
120
0.0
0.5
1.0
1.5
2.0
Ks
species1 estimated significant components in the distribution of synonymous (Ks) substitution rates
species1 non-synonymous (*Ka*) and synonymous (*Ks*) substitution rates
species1 paralogous pairs
species1 self blastn results

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 event(s) in a species or a pair of species. The **KsDistribution** (not shown) can be to plot the distribution of Ks rates and fit the estimated significant normal mixtures component(s) onto 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

Acknowledgement

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 Discovery and Functional Characterization of Genes Regulating Plant Immunity in Perennial Crops (NSF Grant 1546863) with additional support from the One Thousand Plants (1KP) transcriptome project.