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

Eric Wafula, Greg Von Kuster, Joshua Der, Loren Honaas, Saravanaraj Ayyampalayam, Norman Wickett, Jim Leebens-Mack, and Claude dePamphilis.

January 14, 2019

Galaxy Meeting, PAG

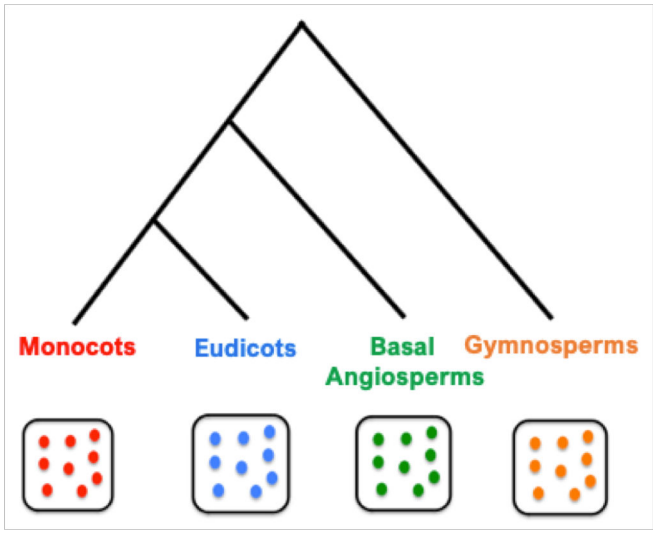
Motivation

- Massive genome and transcriptome data for non-model organisms
- Existing gene family resources are static
- A scalable global gene family framework is needed

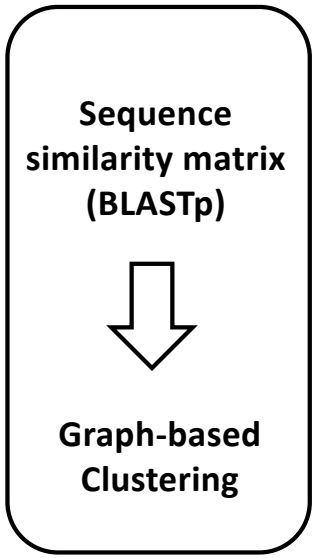
Objectives

- Generate objective gene family (“**orthogroup**”) classifications (“**scaffolds**”) that leverage diverse genomes
- Optimize circumscription of gene families from both evolutionary and functional perspectives
- Develop modular set of analysis tools for comparative phylogenomics of new datasets
- Make these tools broadly available as a stand-alone package (GitHub) and through the Galaxy framework

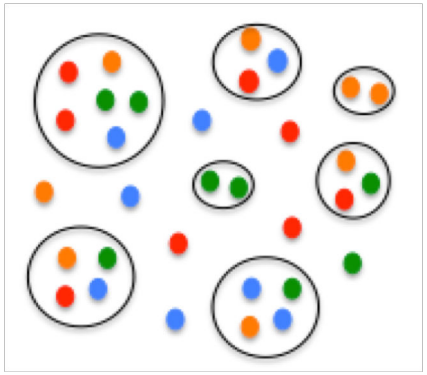
Scaffold construction



selected representative proteomes from clades of interest



primary gene family circumscription



orthogroups, paralogous clusters, and singletons

Scaffold construction

Data

- Orthogroups protein and cds fasta
- Orthogroups protein multiple sequence alignment fasta
- Orthogroup hidden Markov model based profiles (HMM-profiles)
- Scaffold database of profile HMM-profiles
- BLAST database of all scaffold proteins

Metadata

- UniProtKB/Swiss-Prot protein functional annotations
- TAIR *Arabidopsis thaliana* gene functional annotations
- InterPro and Pfam functional domain annotations
- Gene Ontology functional terms
- Gene family sizes (orthogroup gene counts) for scaffold taxa
- Secondary gene family clusters (super-orthogroups)

Scaffold data usage

Genomes

Genome of the long-living sacred lotus (*Nelumbo nucifera* Gaertn.)

Fern genomes elucidate land plant evolution and cyanobacterial symbioses

The *Amborella* Genome and the Evolution of Flowering Plants

Gene family studies

Ecological genomics of tropical trees: how local population size and allelic diversity of resistance genes relate to immune responses, cosusceptibility to pathogens, and negative density dependence

Horizontal gene transfer is more frequent with increased heterotrophy and contributes to parasite adaptation

PlantTribes: a gene and gene family resource for comparative genomics in plants

P. Kerr Wall[1], Jim Leebens-Mack[1,2], Kai F. Müller[1,3], Dawn Field[4], Naomi S. Altman[5] and Claude W. dePamphilis[1,*]

2008 –NAR– 99 citations

http://bigdata.bx.psu.edu/PlantTribes_scaffolds/

Gene and genome duplications

A genome triplication associated with early diversification of the core eudicots

Ancestral polyploidy in seed plants and angiosperms

The butterfly plant arms-race escalated by gene and genome duplications

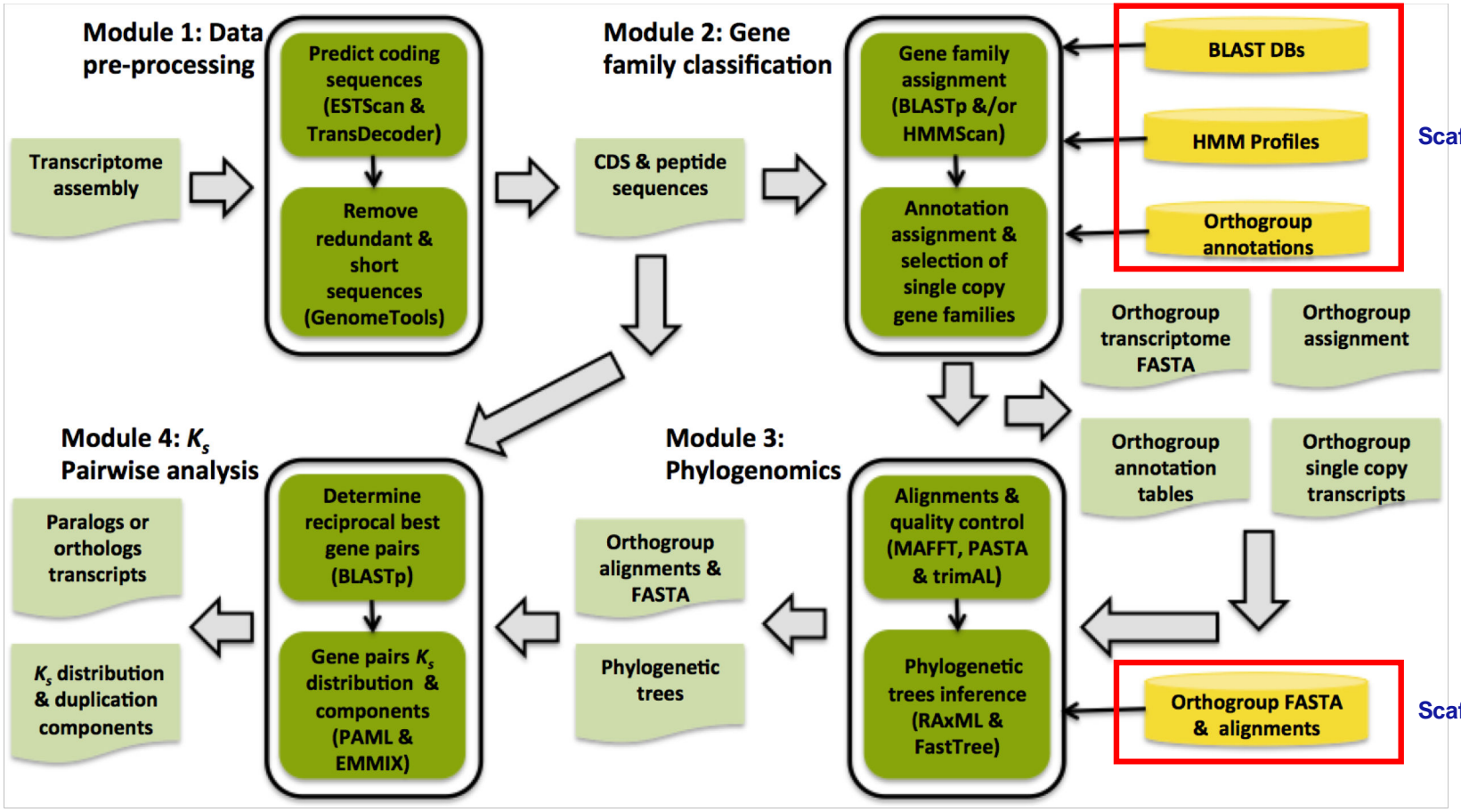
Expression shifts

Comparative Transcriptome Analyses Reveal Core Parasitism Genes and Suggest Gene Duplication and Repurposing as Sources of Structural Novelty

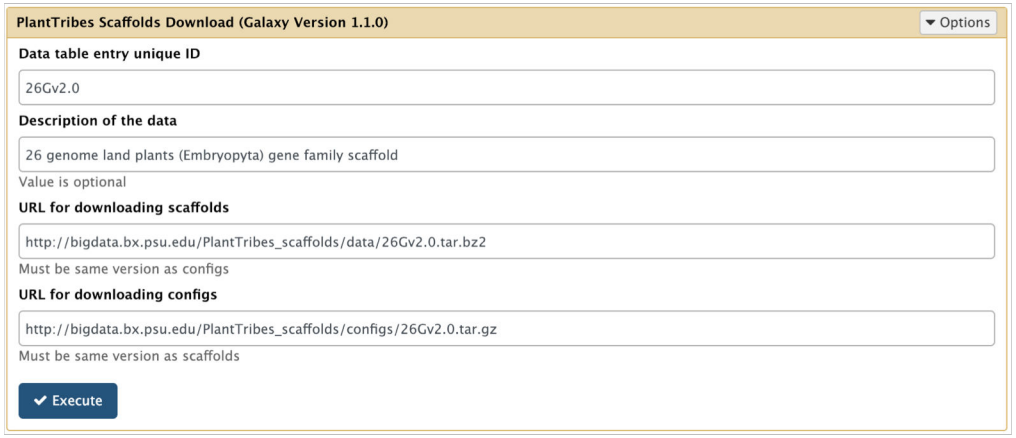
Functional genomics of a generalist parasitic plant: Laser microdissection of host-parasite interface reveals host-specific patterns of parasite gene expression

Transcriptomes of the Parasitic Plant Family Orobanchaceae Reveal Surprising Conservation of Chlorophyll Synthesis

PlantTribes workflow



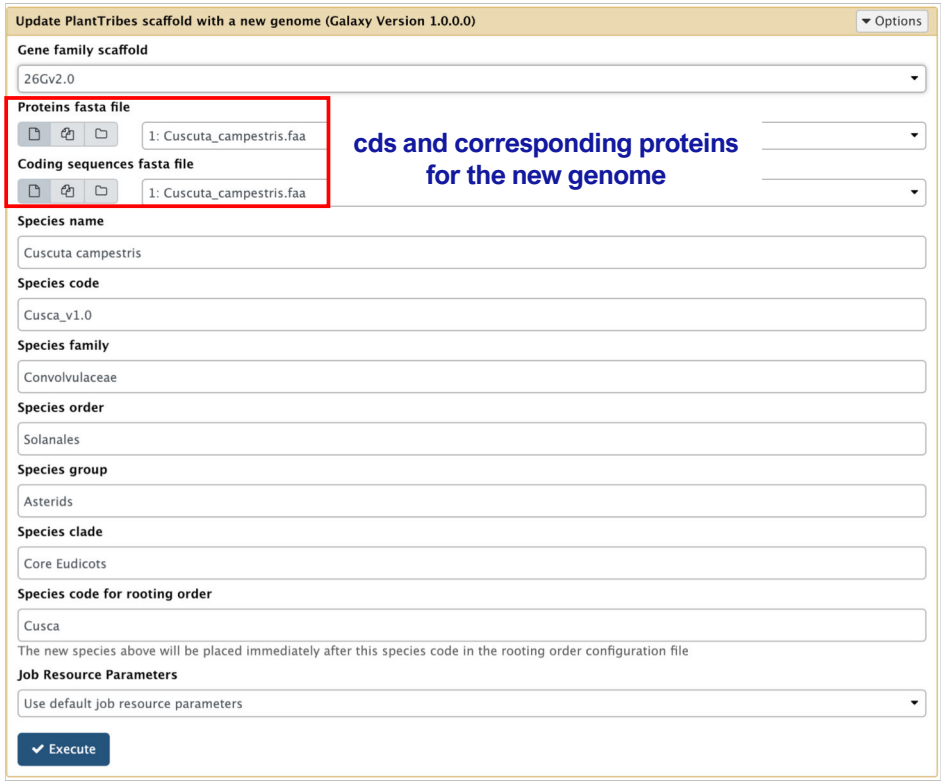
Scaffold installation



Installing gene family scaffold

http://bigdata.bx.psu.edu/PlantTribes_scaffolds/

https://github.com/dePamphilis/PlantTribes

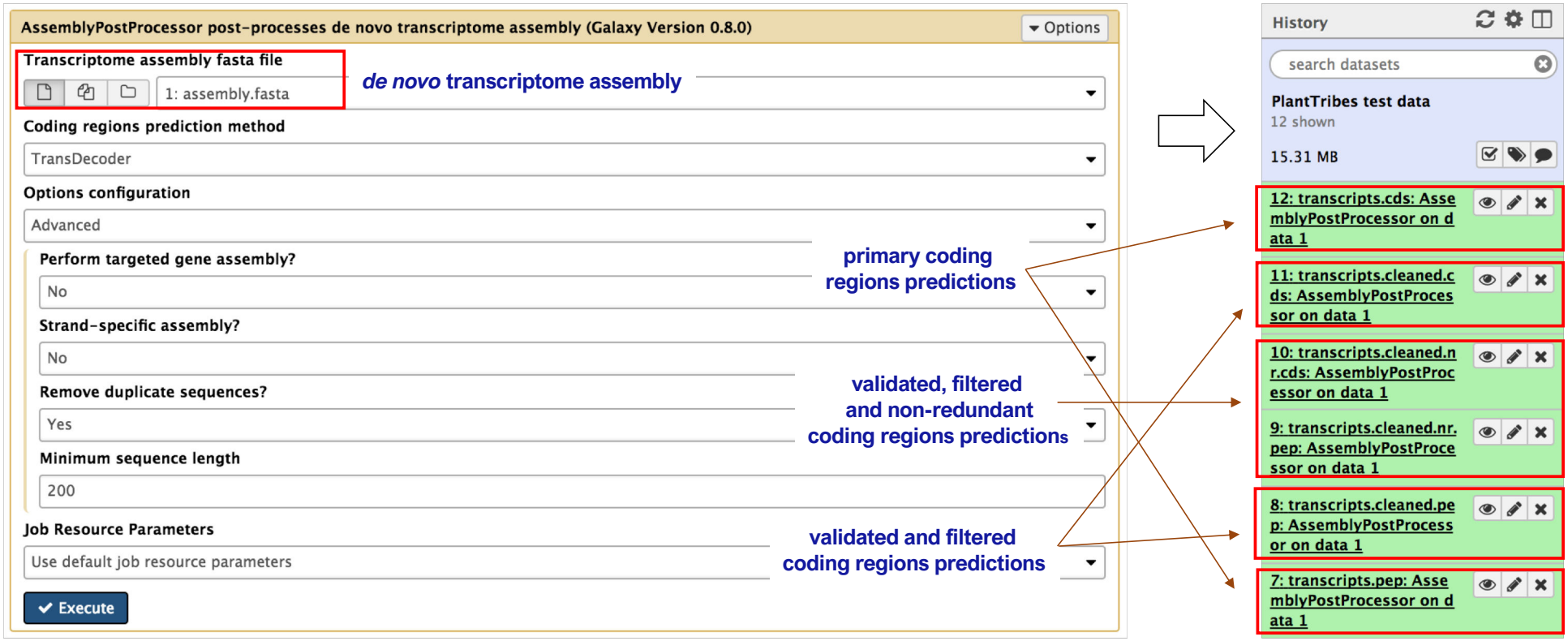


Updating Installed gene family scaffold with a new sequenced genome

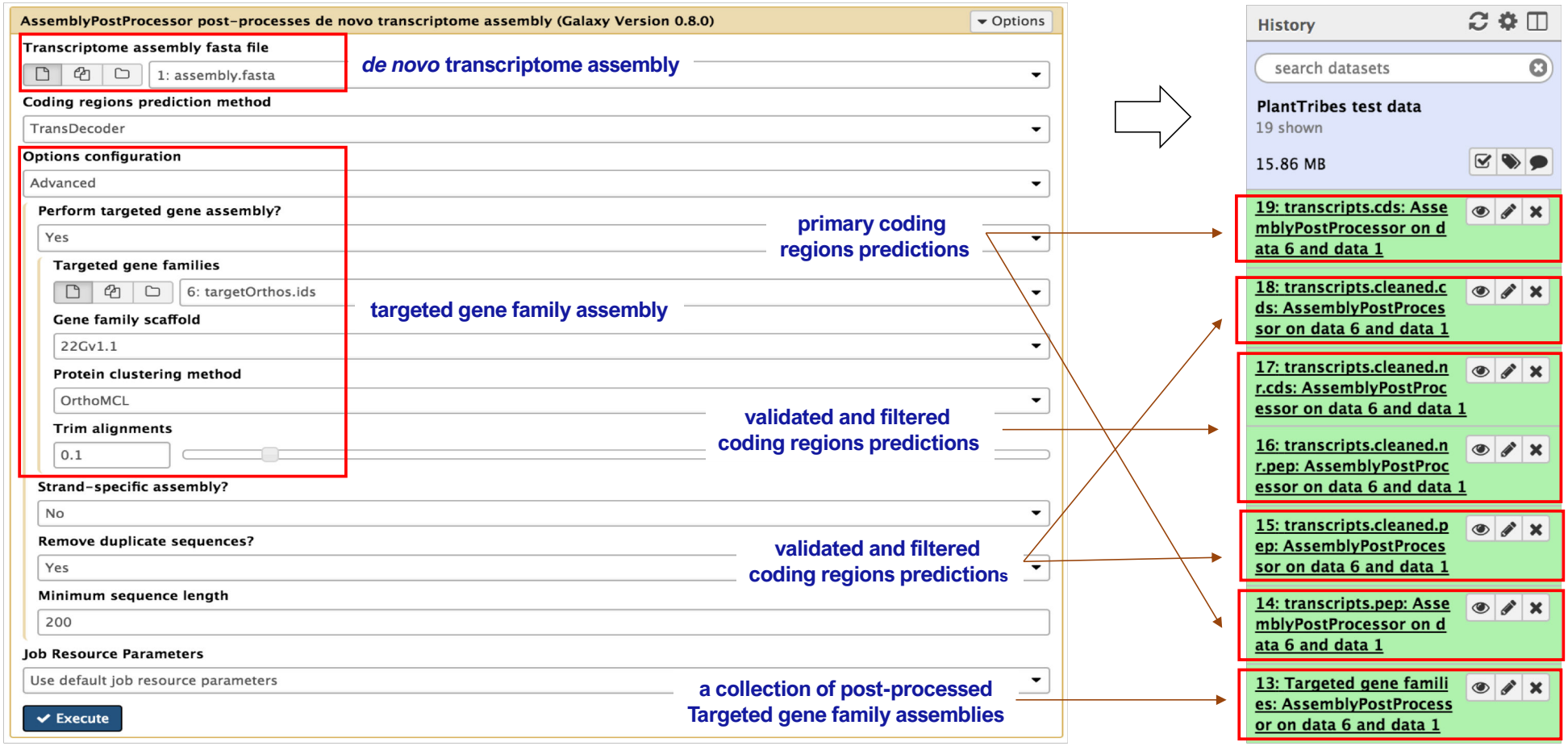
Analysis tools

1. Post-assembly QC
2. Assembly sorting
3. Alignment and QC
4. Phylogenetic inference
5. Selection and WGD analysis

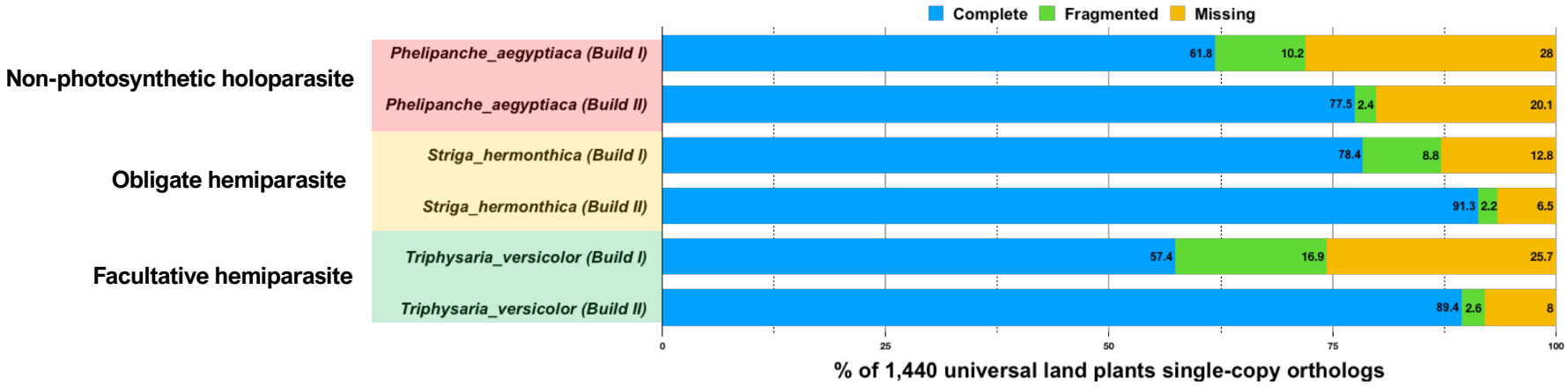
1. Post-assembly QC
Assembly post-processing



1. Post-assembly QC
Targeted gene family assembly

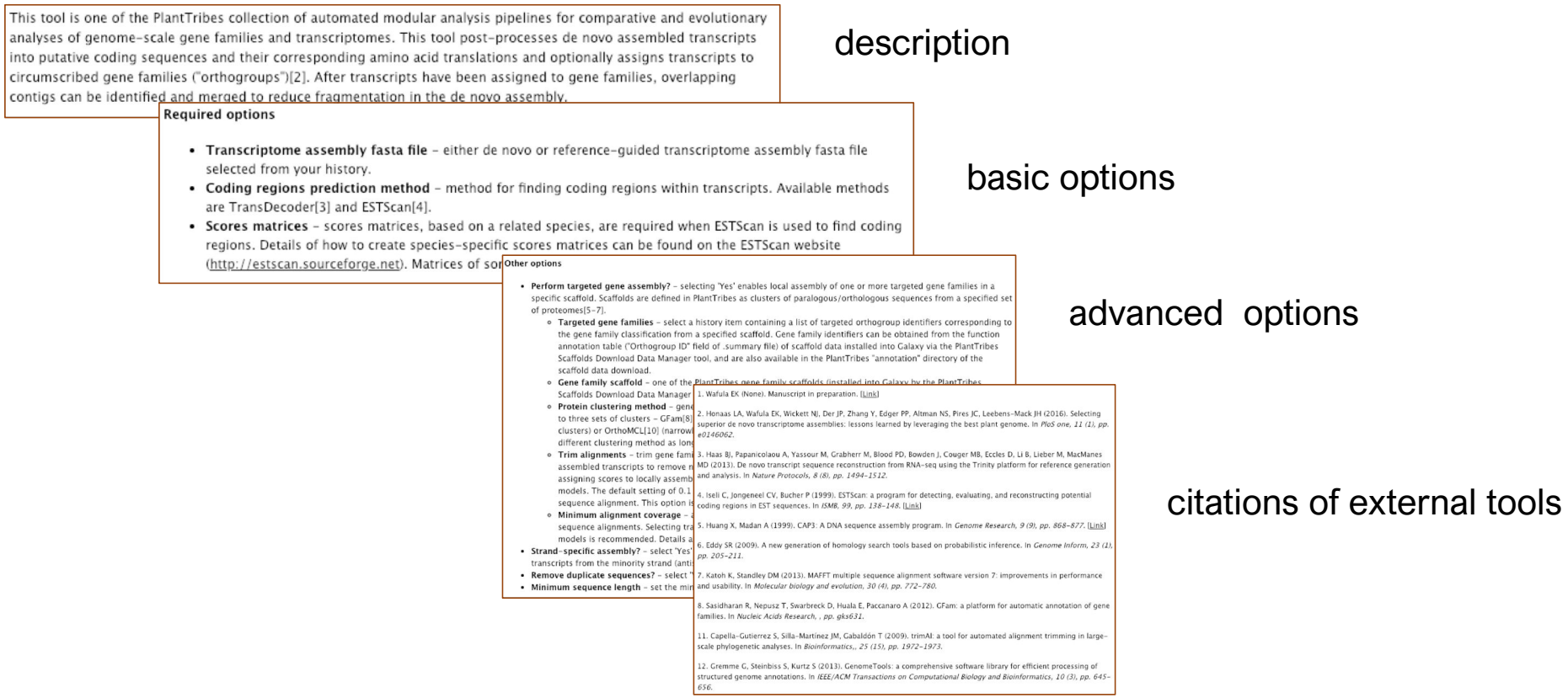


1. Post-assembly QC
Assembly improvement



Assemblies from Parasite Plant Genome Project (PPGP)
(*Orobancheceae* family)

1. Post-assembly QC
Tool documentation



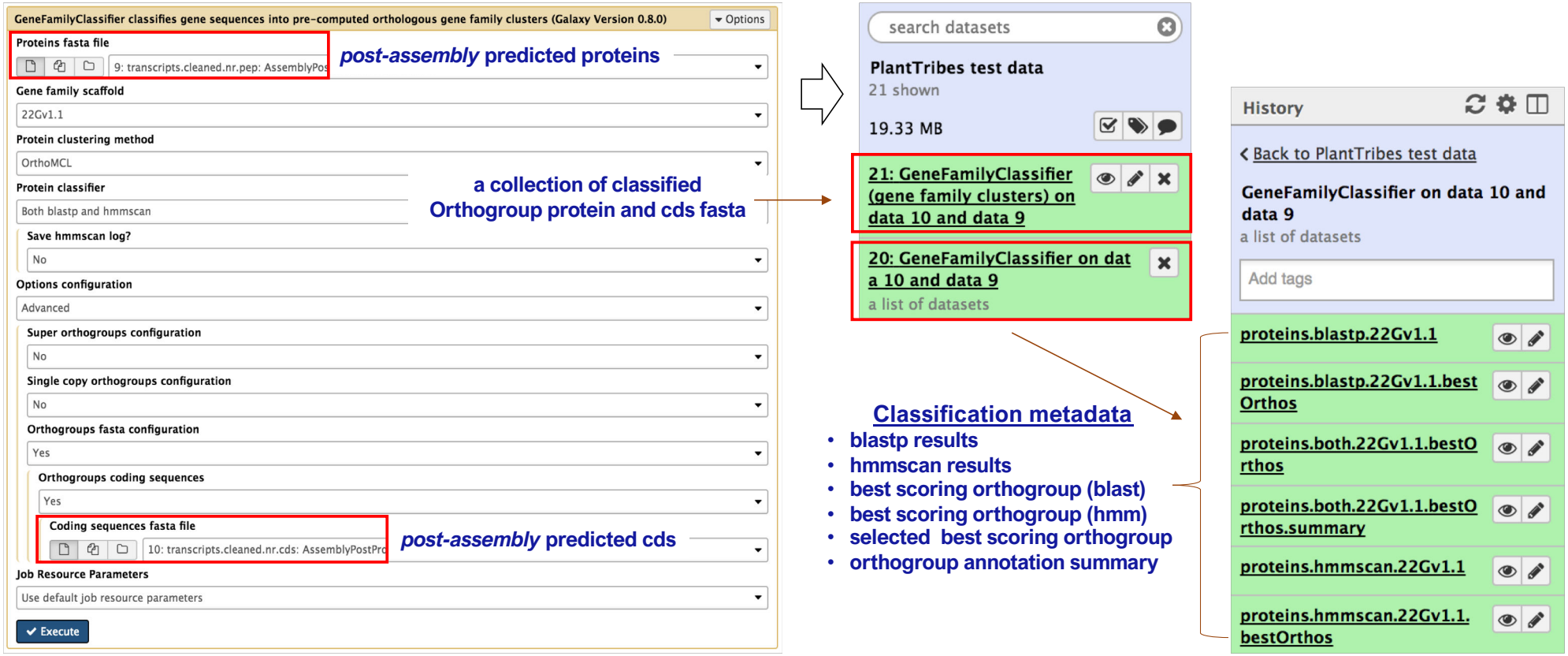
description

basic options

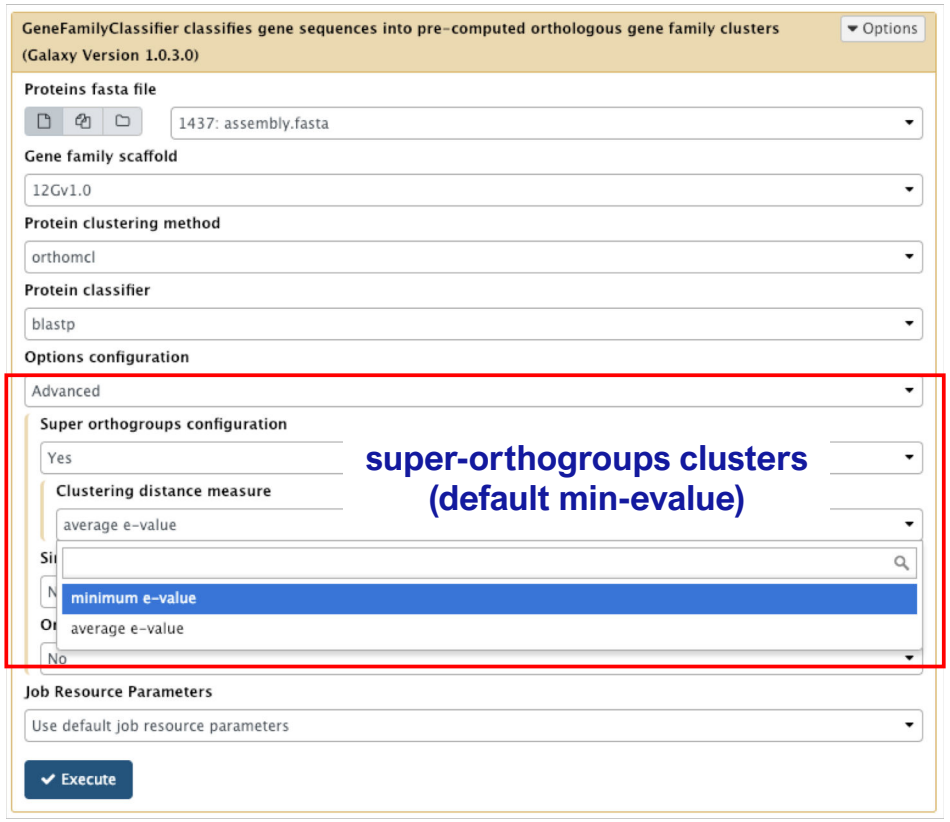
advanced options

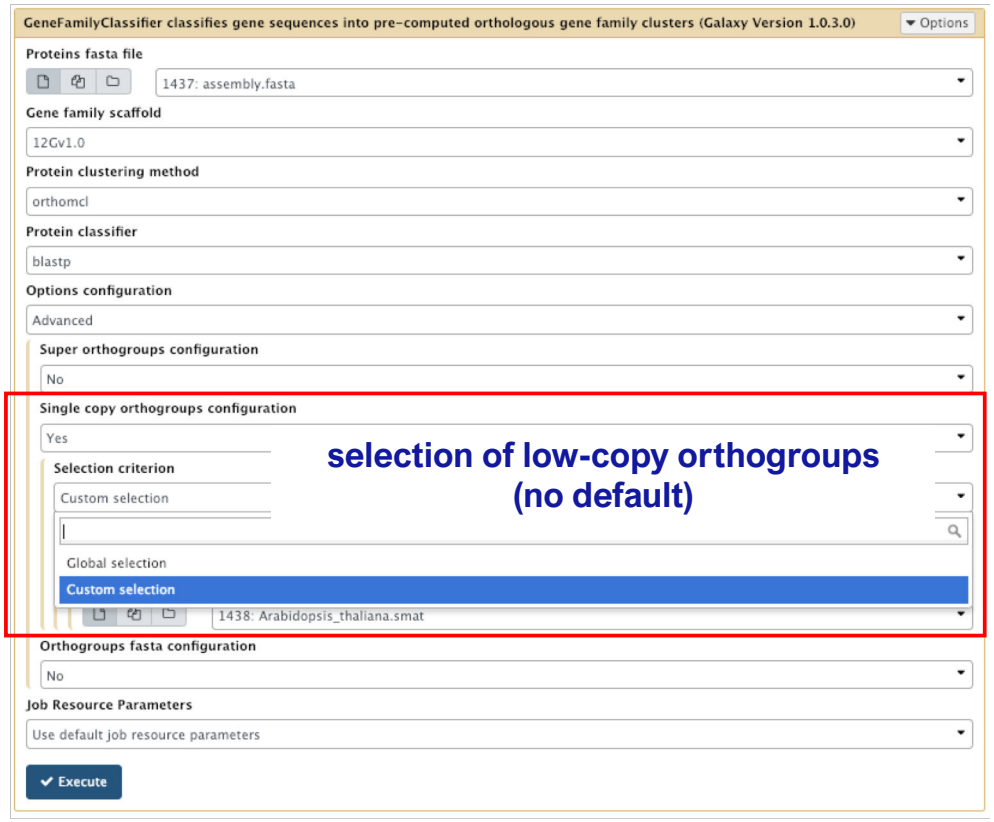
citations of external tools

2. Assembly sorting
Gene family classification

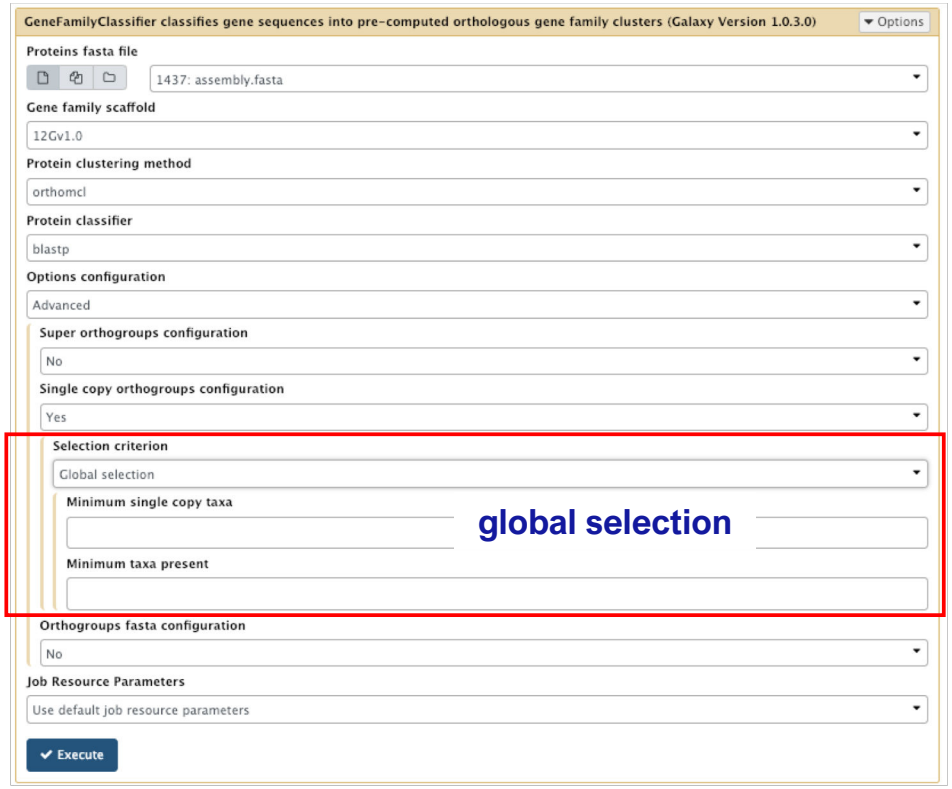


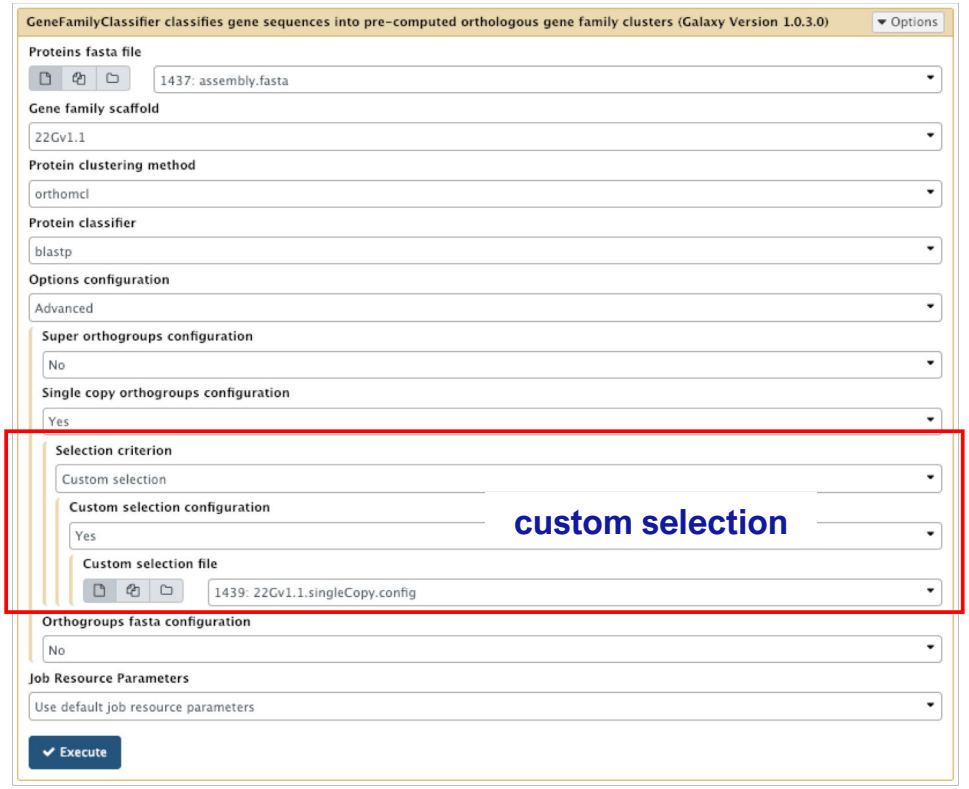
2. Assembly sorting
Advanced analyses



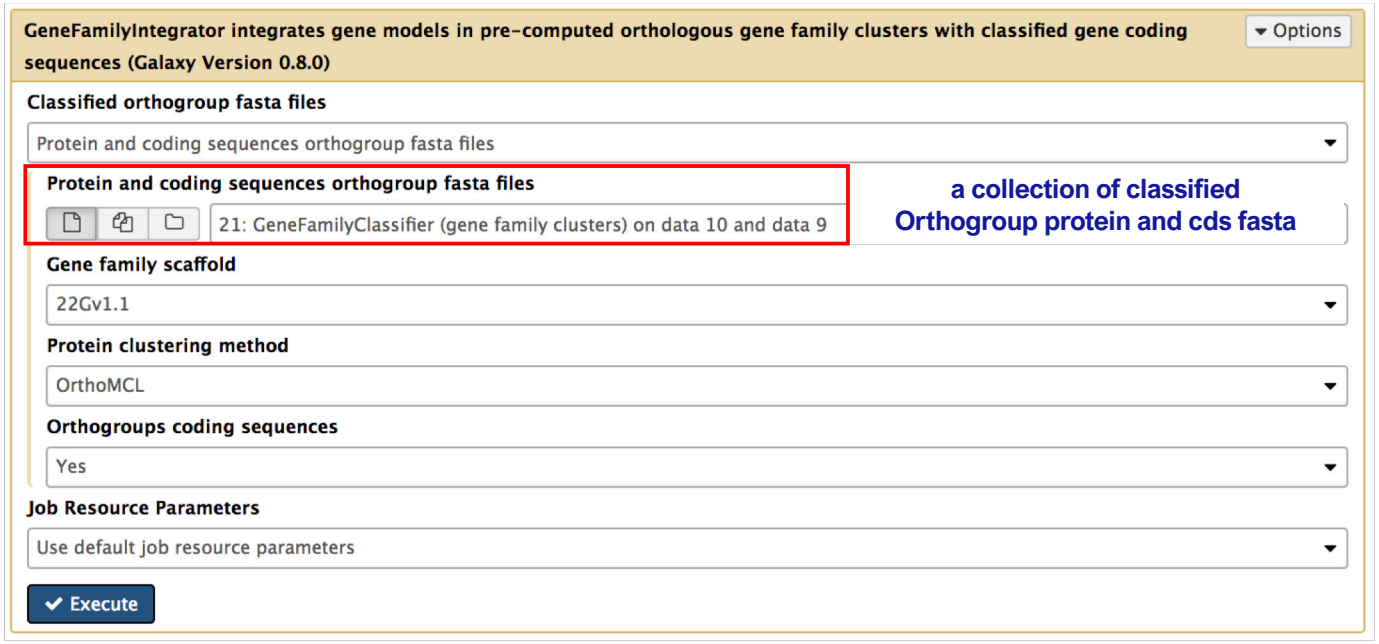


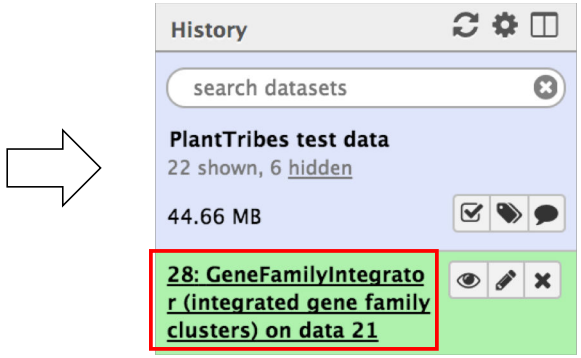
2. Assembly sorting
Selection low-copy gene families



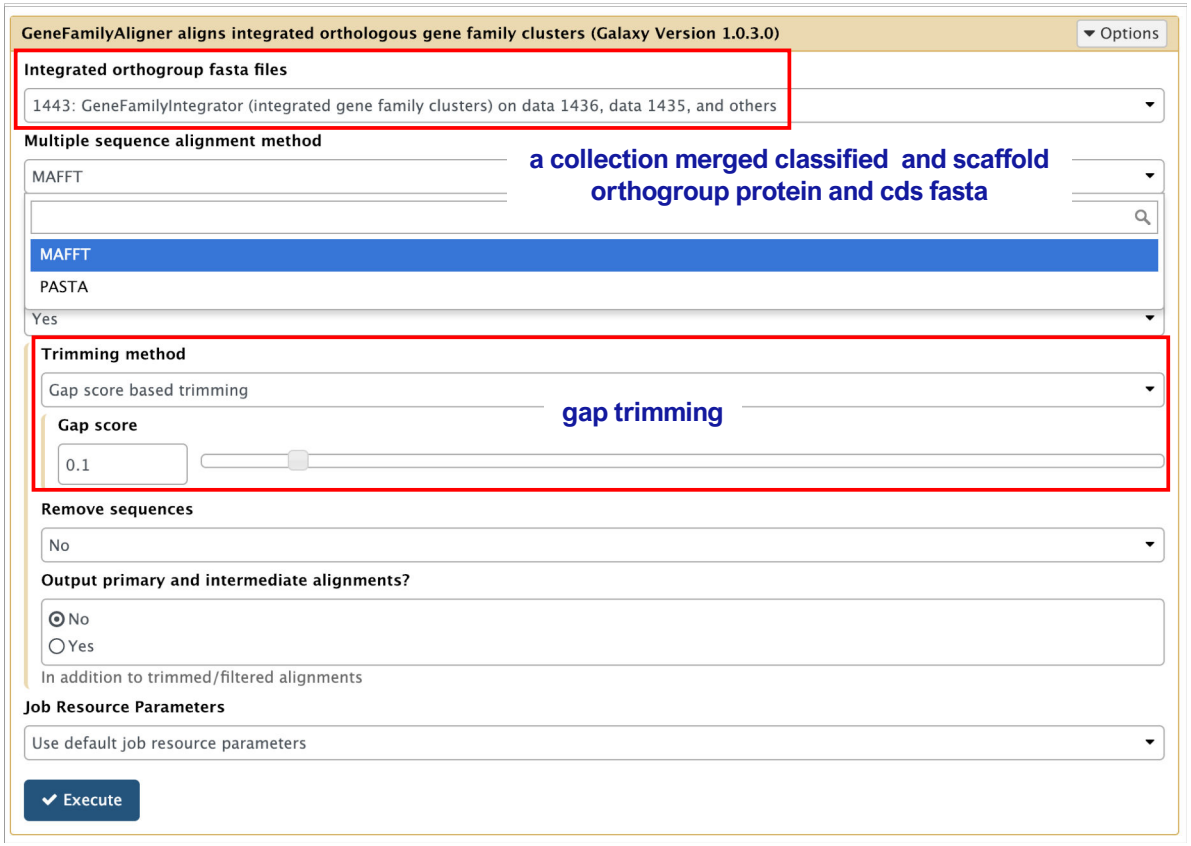


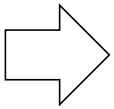
2. Assembly sorting
Gene family integration

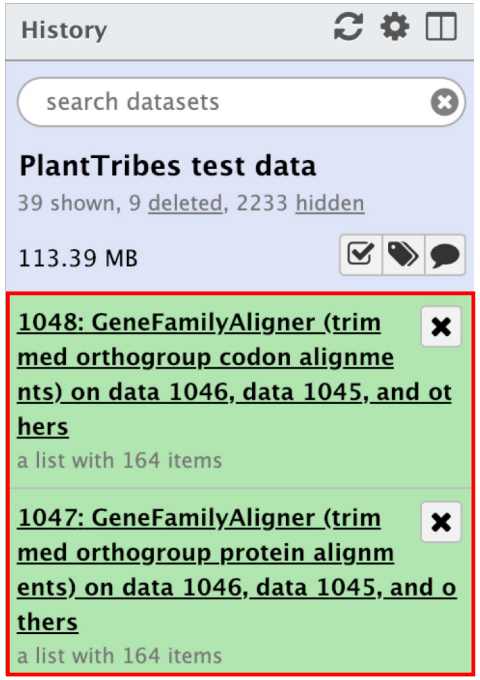




3. Alignment and QC
Gene family alignments

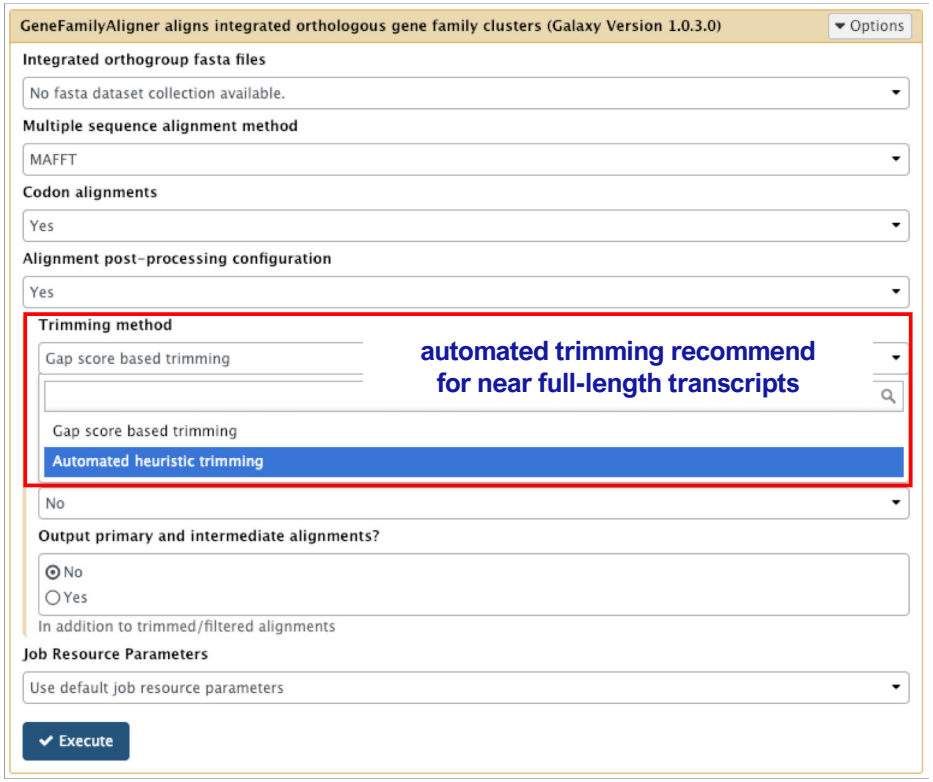


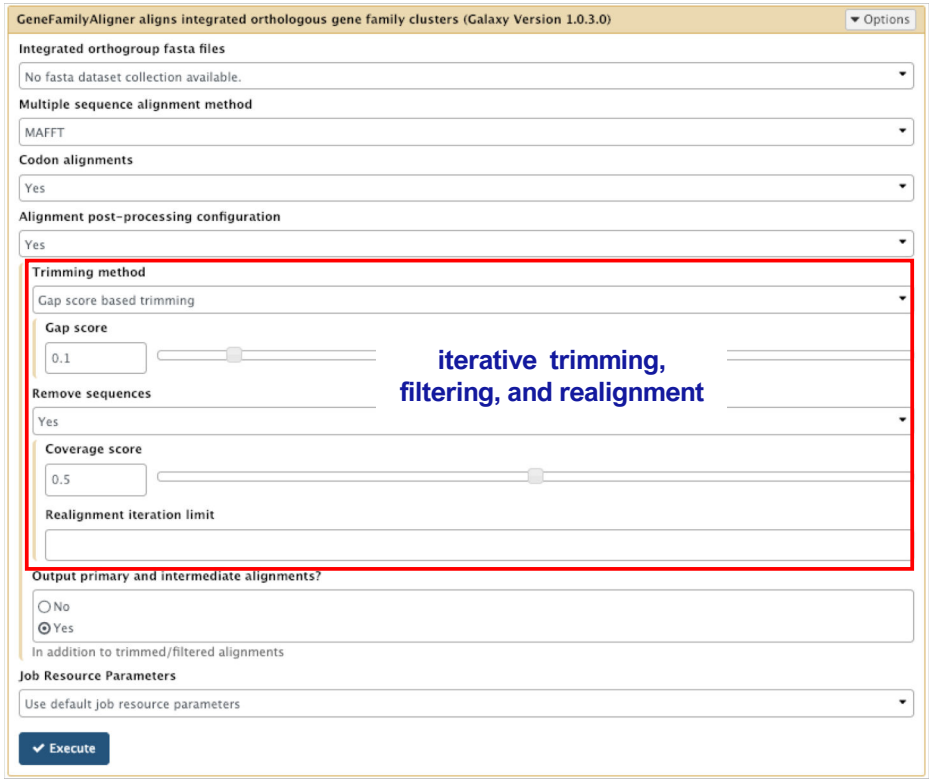




collection of estimated orthogroup codon and protein multiple sequence alignments

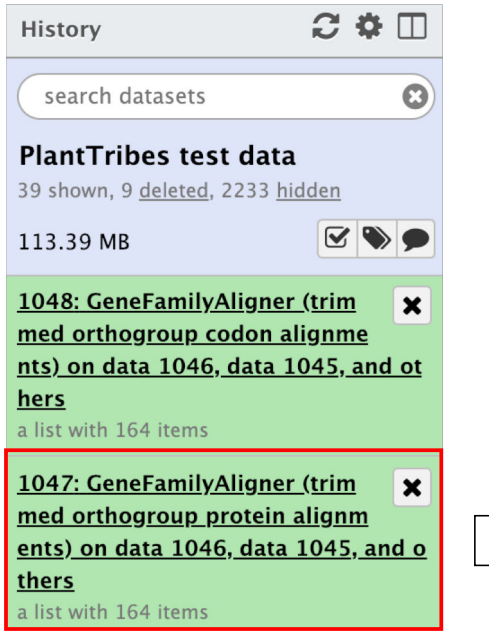
3. Alignment and QC
Alignment post-processing



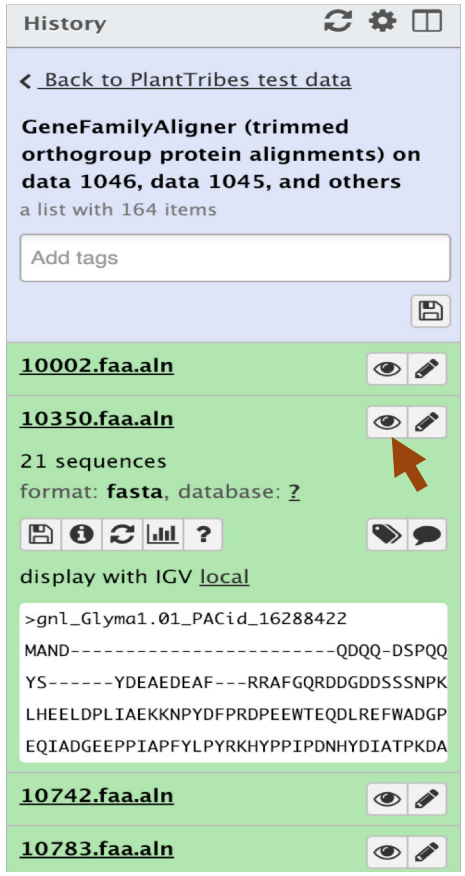


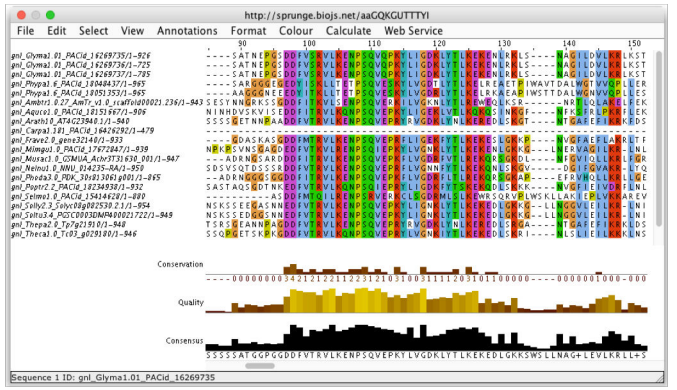
3. Alignment and QC

Alignment visualization

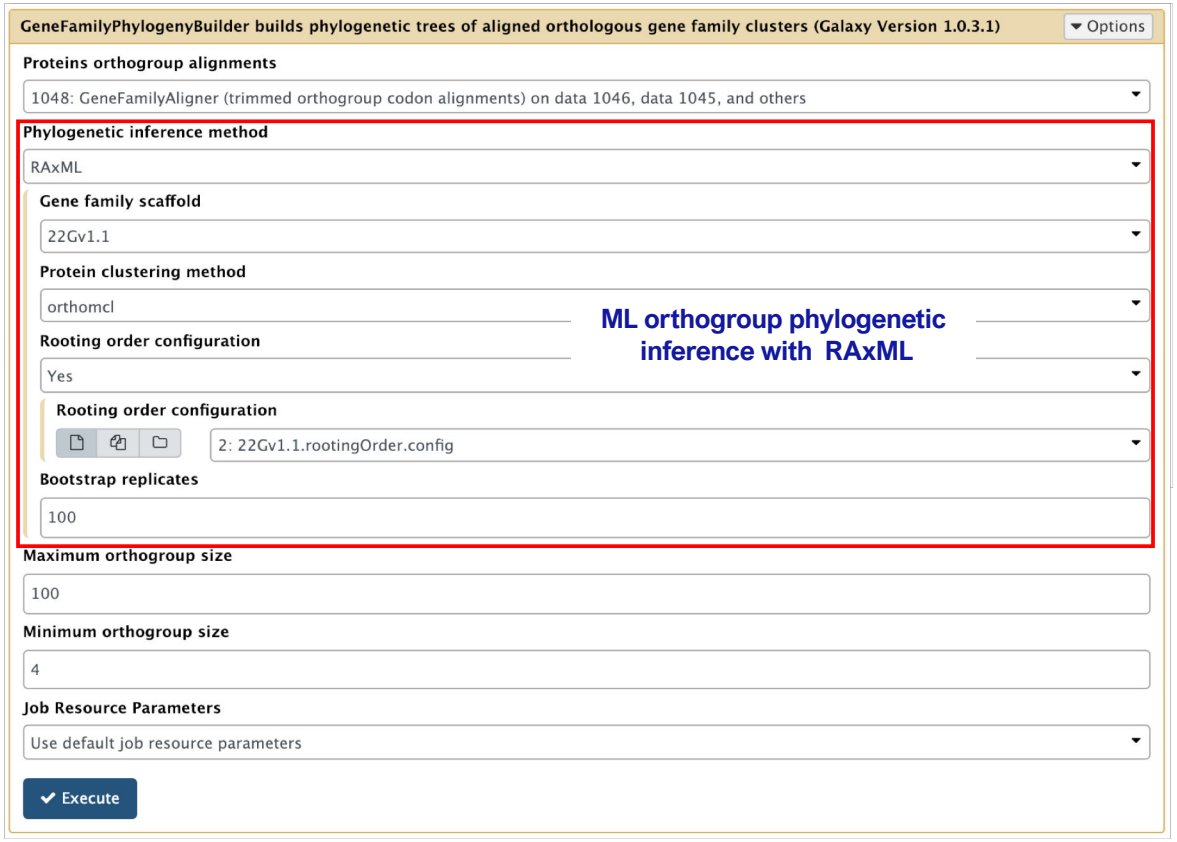


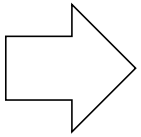
collection of estimated orthogroup protein multiple sequence alignments

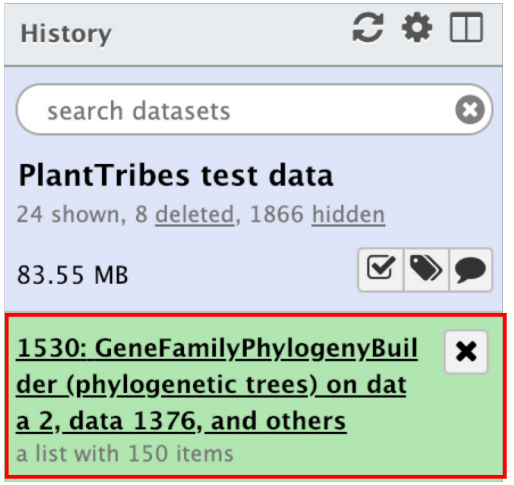




4. Phylogenetic inference
Gene family phylogenetic trees

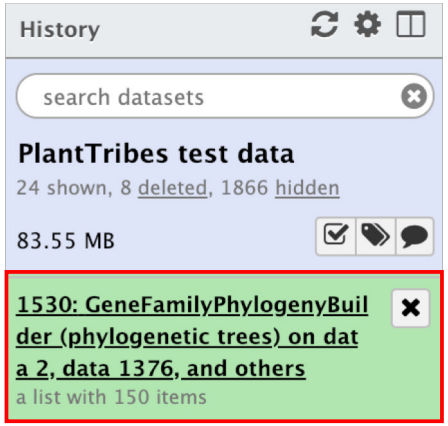




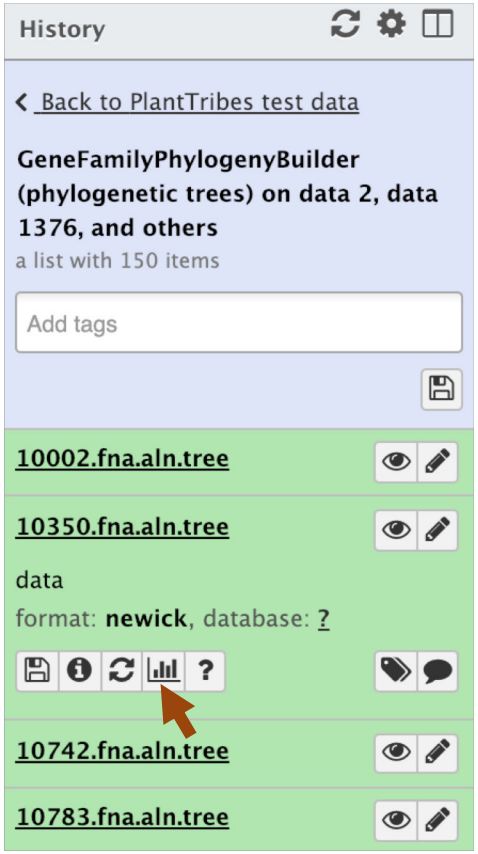


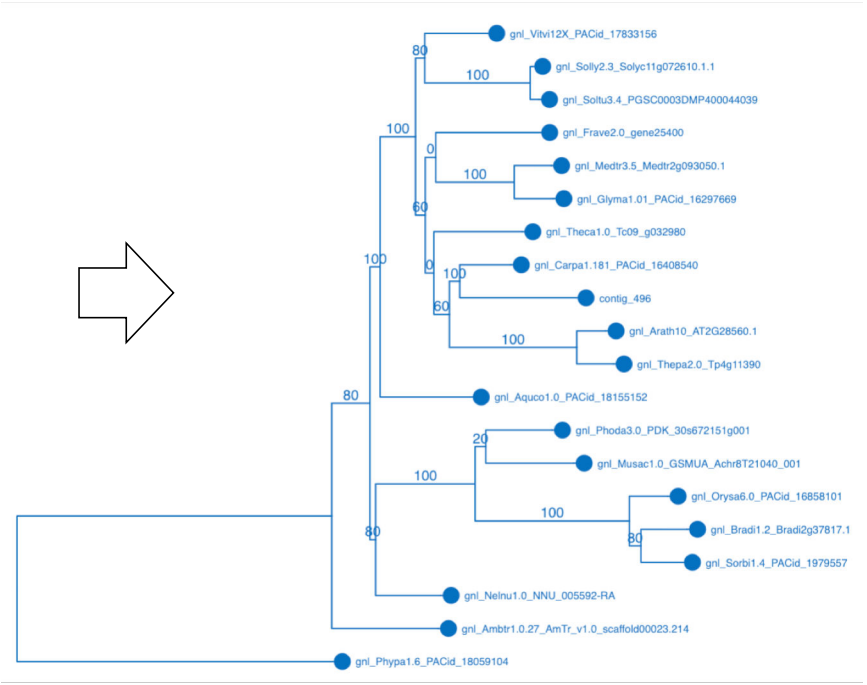
collection of inferred orthogroup Phylogenetic trees

4. Phylogenetic inference
Phylogenetic tree visualization



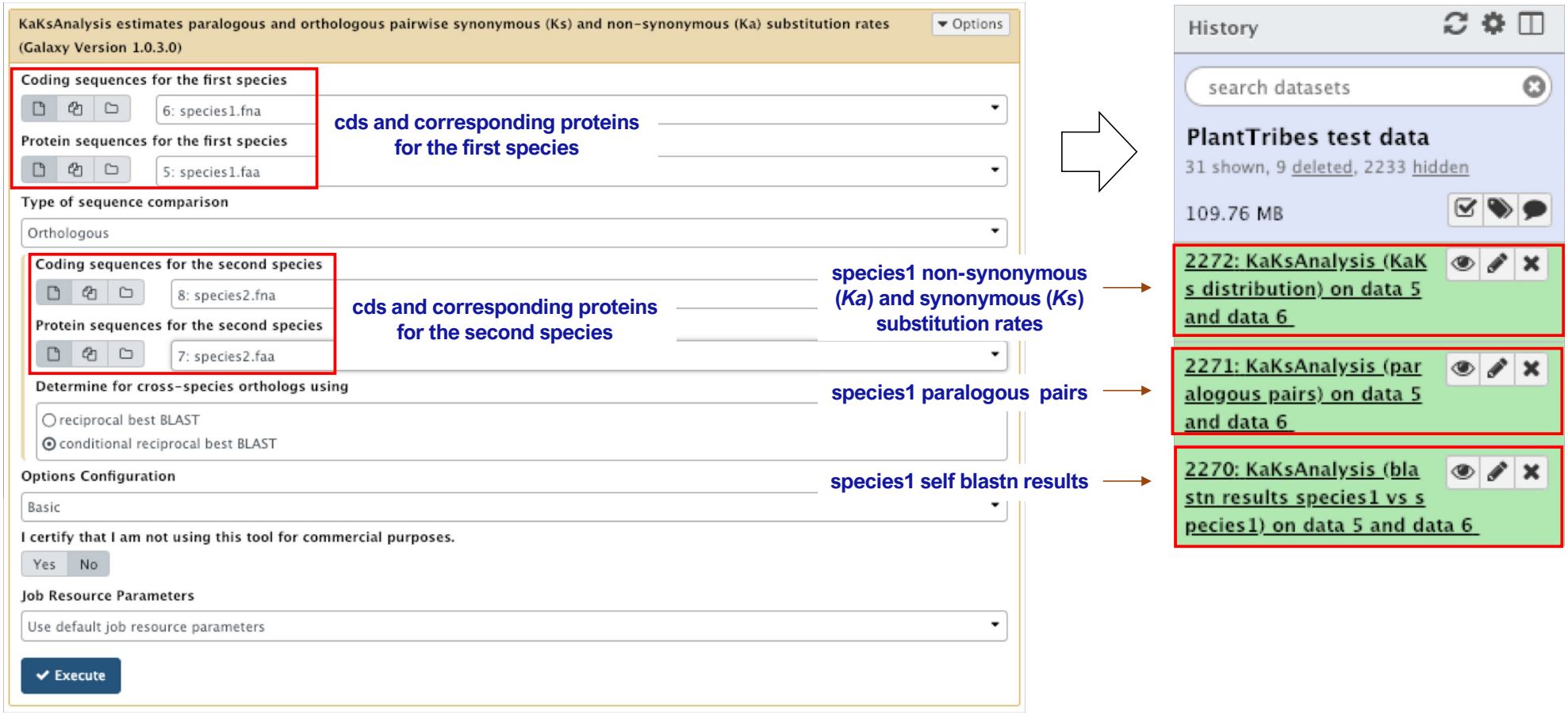
collection of inferred orthogroup Phylogenetic trees





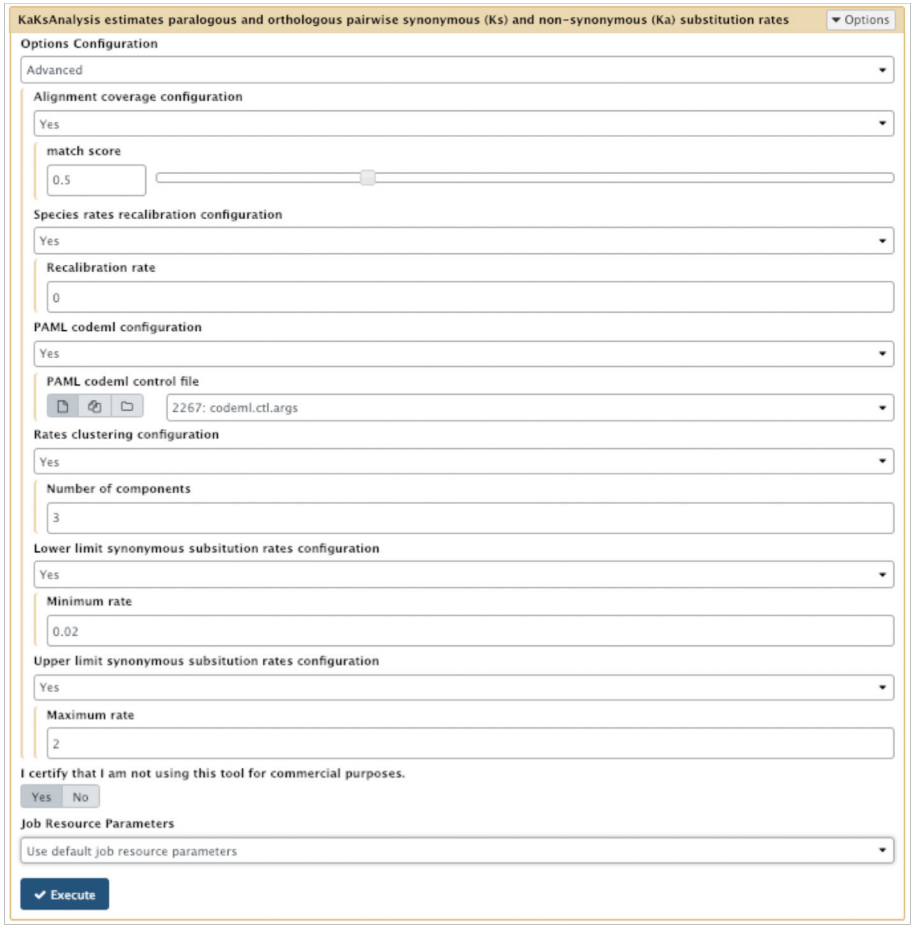
5. Selection and WGD analysis

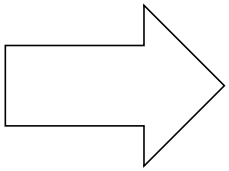
Nucleotide substitution rates

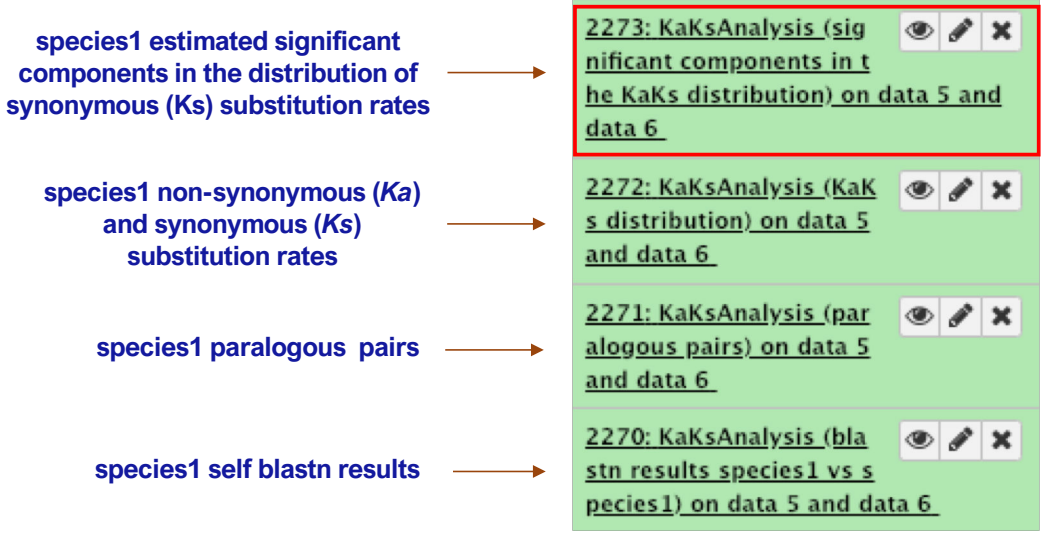


5. Selection and WGD analysis

Genome duplication inference

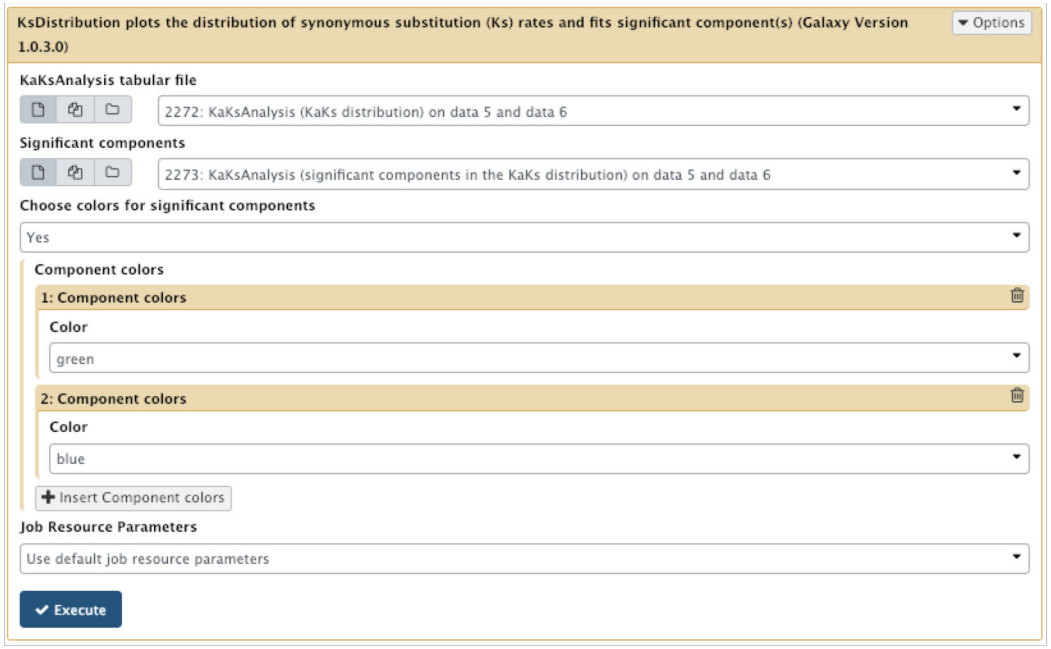


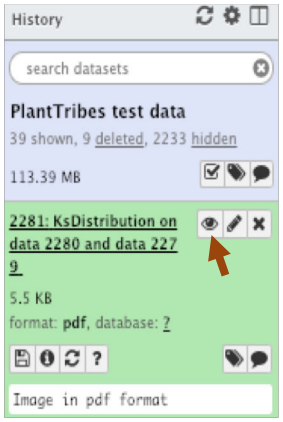


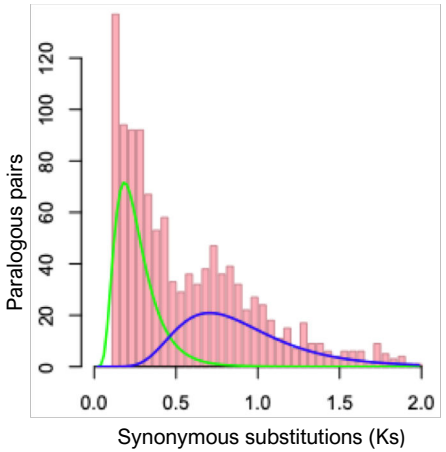


5. Selection and WGD analysis

Genome duplication synonymous substitution rates (Ks) distribution plot







Applications

- Improving genome annotation quality
- Transcriptome coverage and functional annotation
- Species tree inference using single copy genes
- Gene tree – species tree reconciliation
- Timing of gene duplications and polyploidy
- Ancestral gene content reconstruction
- Gene family expansions/contractions
- Timing of new gene function evolution
- Studies of horizontal transfers among species
- Many others…

Availability

Tools:

- Stand-alone: https://github.com/dePamphilis/PlantTribes
- Galaxy: https://usegalaxy.org/
- Galaxy Tool Shed: https://toolshed.g2.bx.psu.edu

Tutorials:

- Stand-alone: https://github.com/dePamphilis/PlantTribes/blob/master/docs/Tutorial.md
- Galaxy: https://galaxyproject.org/tutorials/pt_gfam/

Acknowledgments

Claude dePamphilis
Penn State

James Leebens-Mack
UGA

Raj Ayyampalayam
UGA

Kerr Wall
BASF

Greg Von Kuster
Penn State

Norman Wickett
Chicago Botanic Garden

Joshua Der
CSU Fullerton

Loren Honaas
USDA ARS

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.