Homology and orthology inference

Ya Yang

yangya@umn.edu

Transcriptome workshop, Botany 2018

Herbarium and Department of Plant Biology

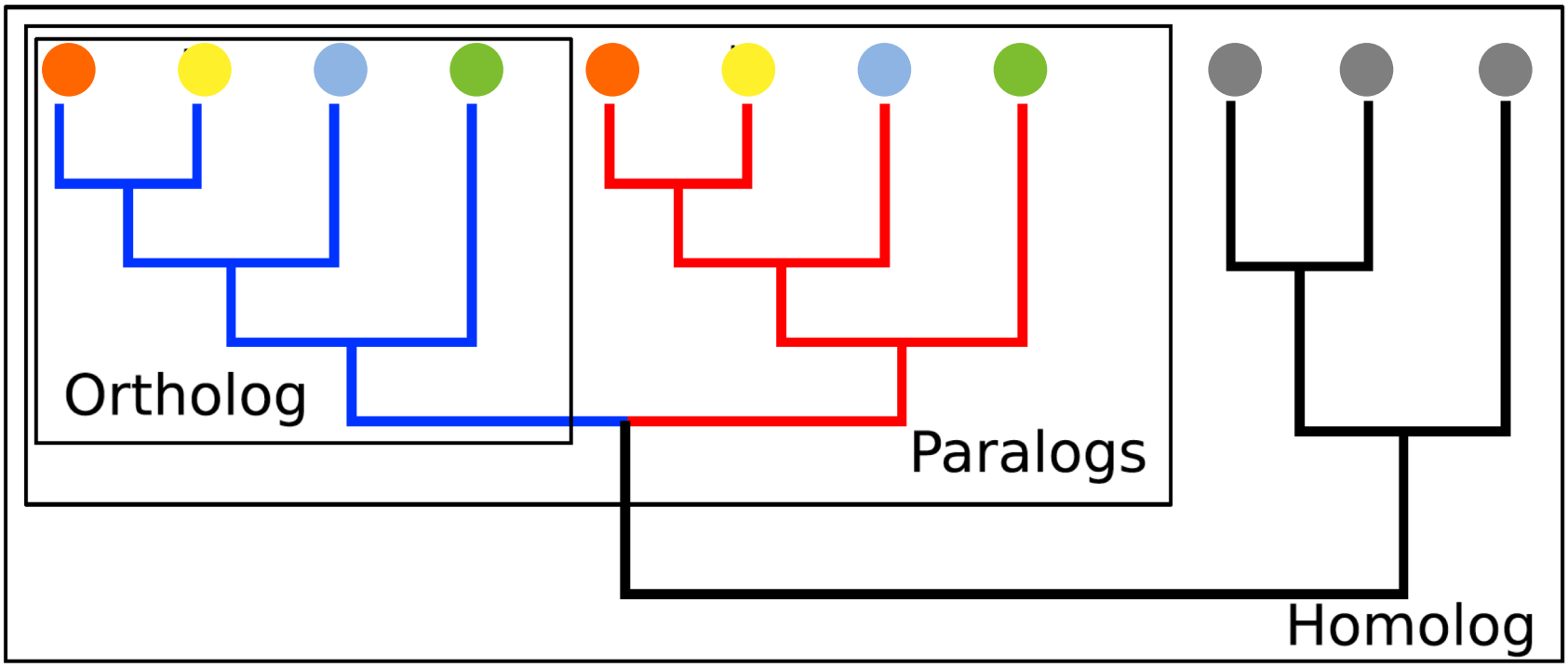
University of Minnesota-Twin Cities

Homology and orthology inference methods

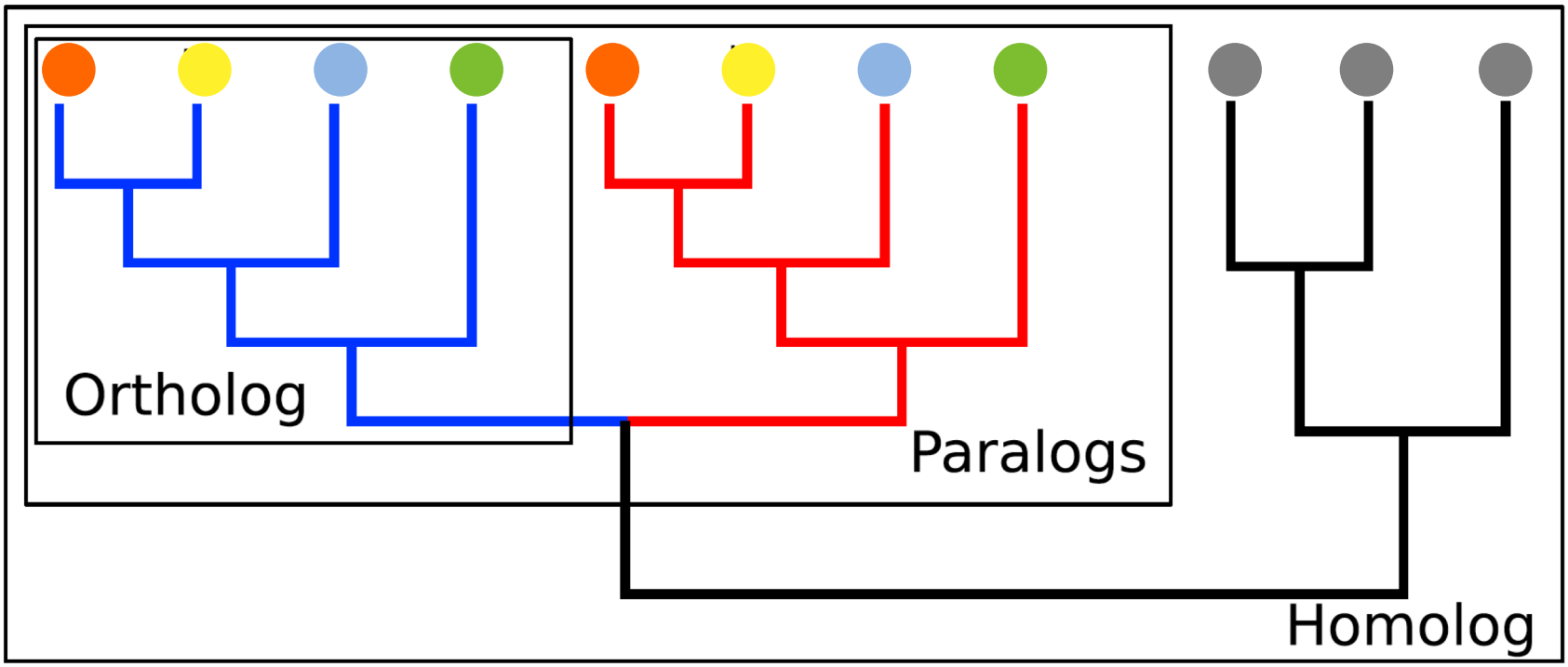
- Based on all-by-all homology search
  - Yang and Smith, 2014, MBE
  - orthoMCL and orthoFinder
- Homology search using a reference gene set or clusters from annotated genomes
  - HMM or Phylome
- Hierarchical clustering

Homology and orthology inference methods

- Based on all-by-all homology search
  - Yang and Smith, 2014, MBE
  - orthoMCL and orthoFinder
- Homology search using a reference gene set or clusters from annotated genomes
  - HMM or Phylome
- Hierarchical clustering


Ortholog
Paralogs
Homolog

Homolog → ortholog



Homology Inference

All-by-all blast

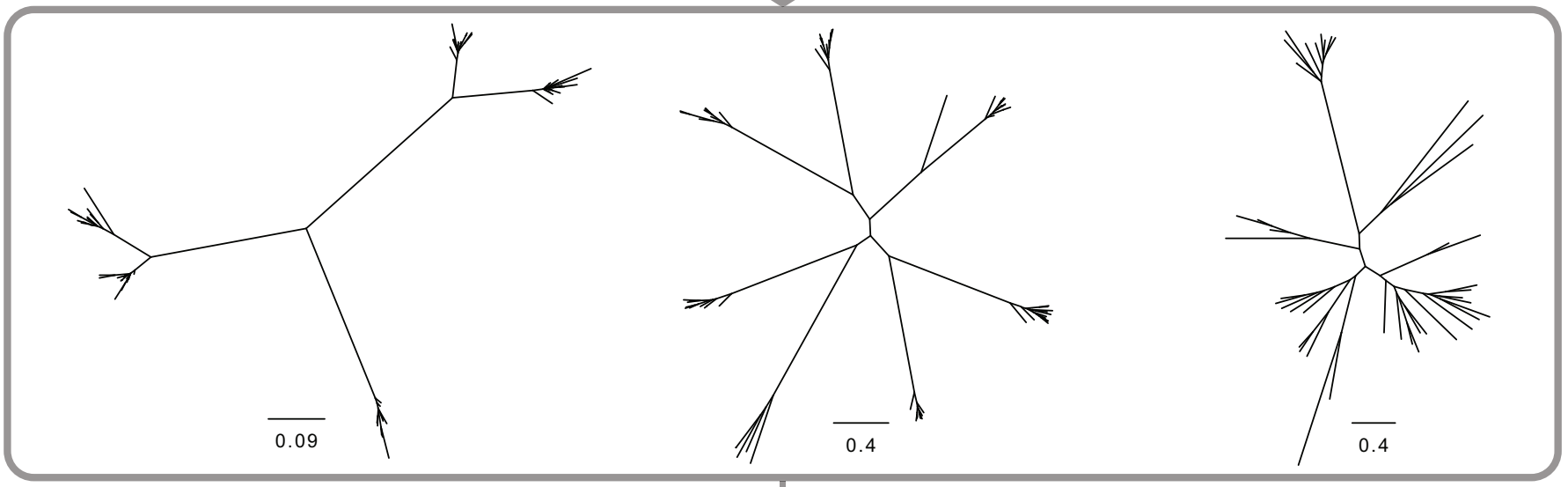
Filter blast results by hit fraction

Markov Clustering

Bait SWIPE against each proteome

Filter top hits within each proteome

Alignment and tree inference



Cut spurious tips and mask tips

Separate long internal branches

Baited homology Inference

All-by-all blast

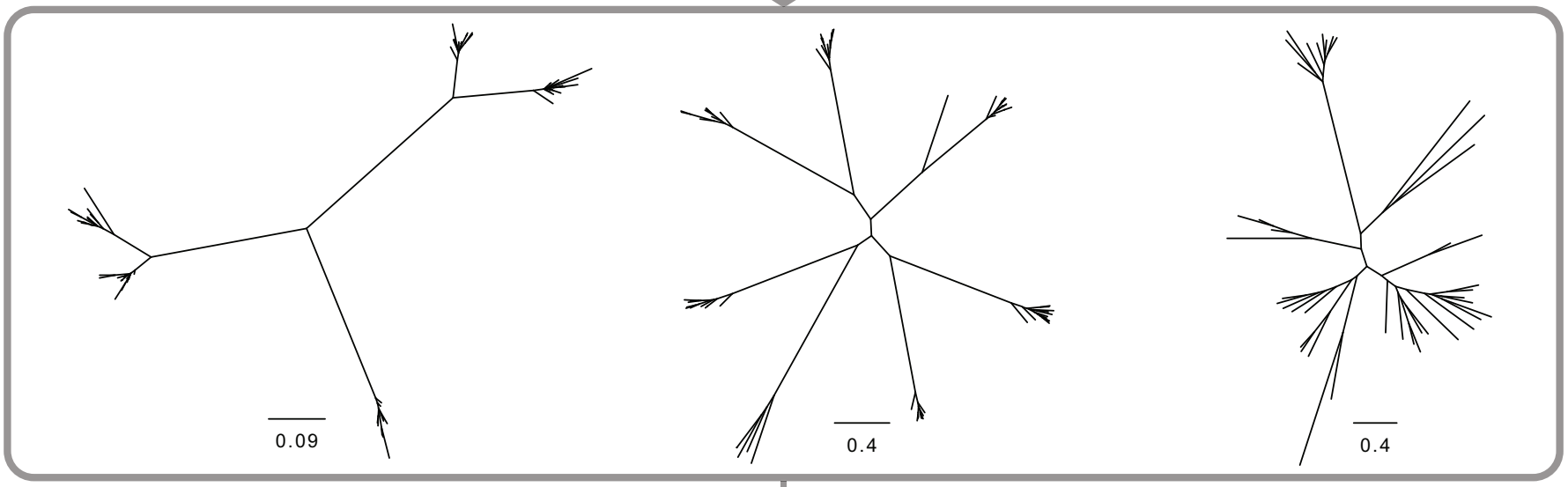
Filter blast results by hit fraction

Markov Clustering

Bait SWIPE against each proteome

Filter top hits within each proteome

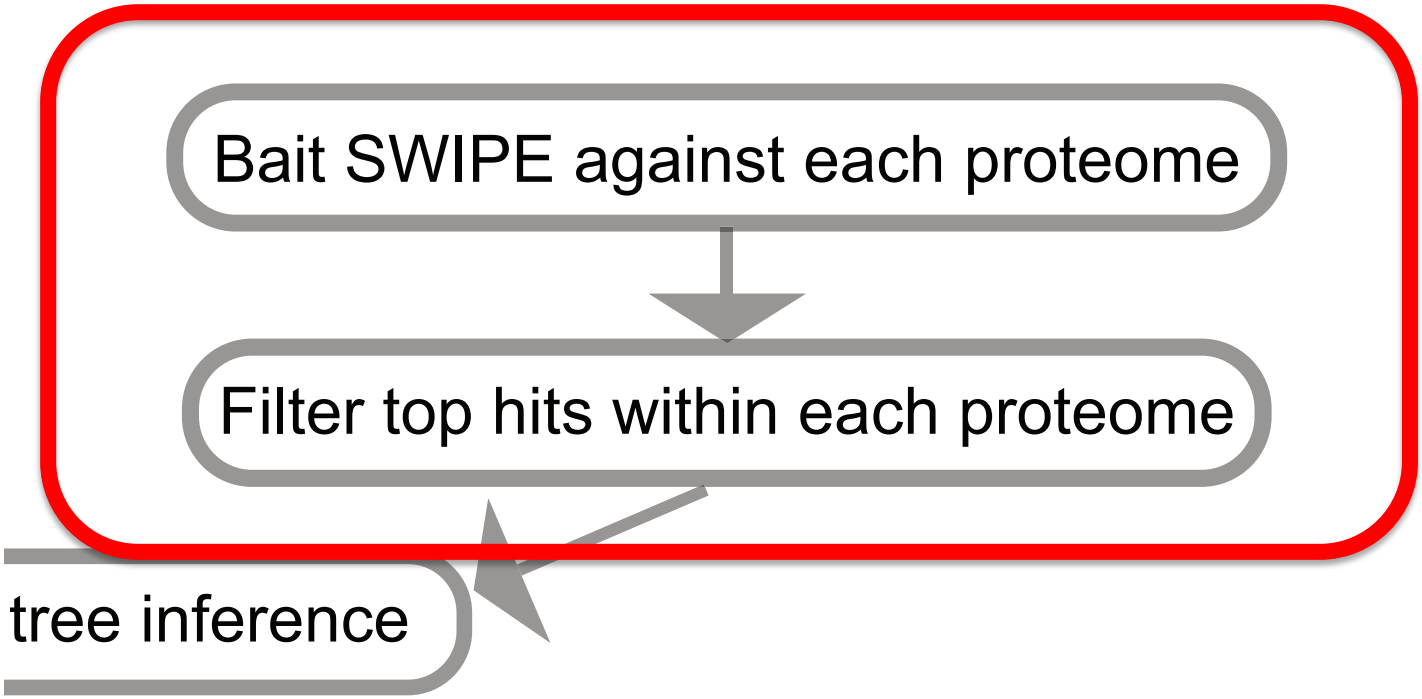
Alignment and tree inference



Cut spurious tips and mask tips

Separate long internal branches

Baited homology Inference



- Use Smith-Waterman (SWIPE), not BLAST
- Search against each species separately and take the top 5–20 hits, instead of searching against all species at once

All-by-all homology Inference

All-by-all blast

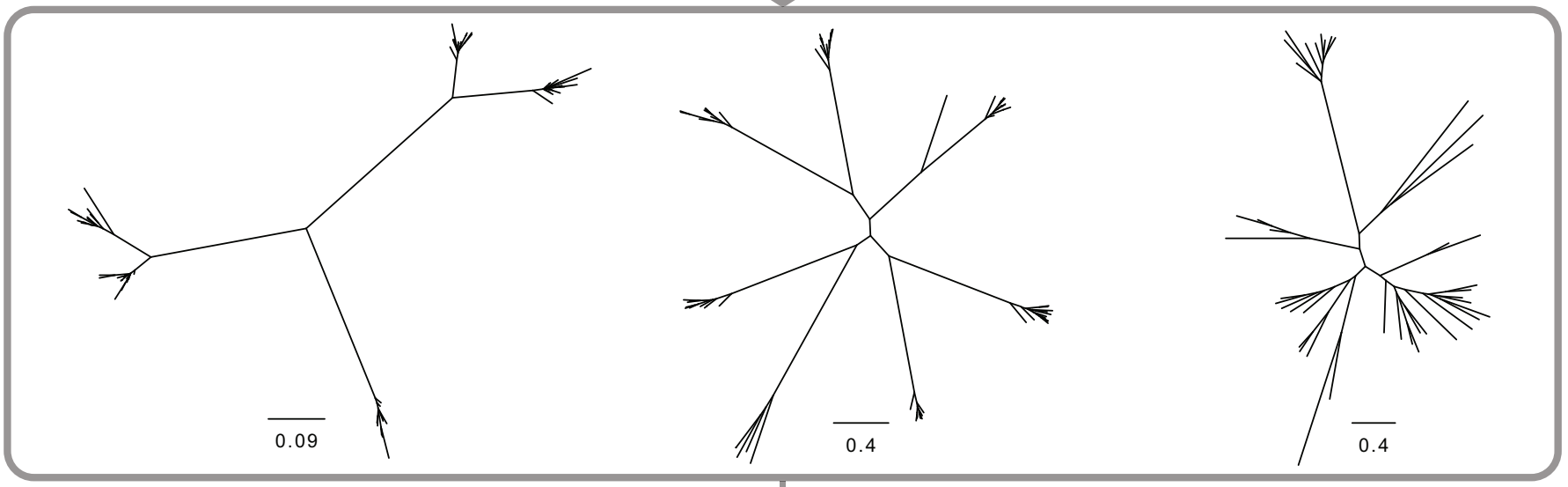
Filter blast results by hit fraction

Markov Clustering

Bait SWIPE against each proteome

Filter top hits within each proteome

Alignment and tree inference



Cut spurious tips and mask tips

Separate long internal branches

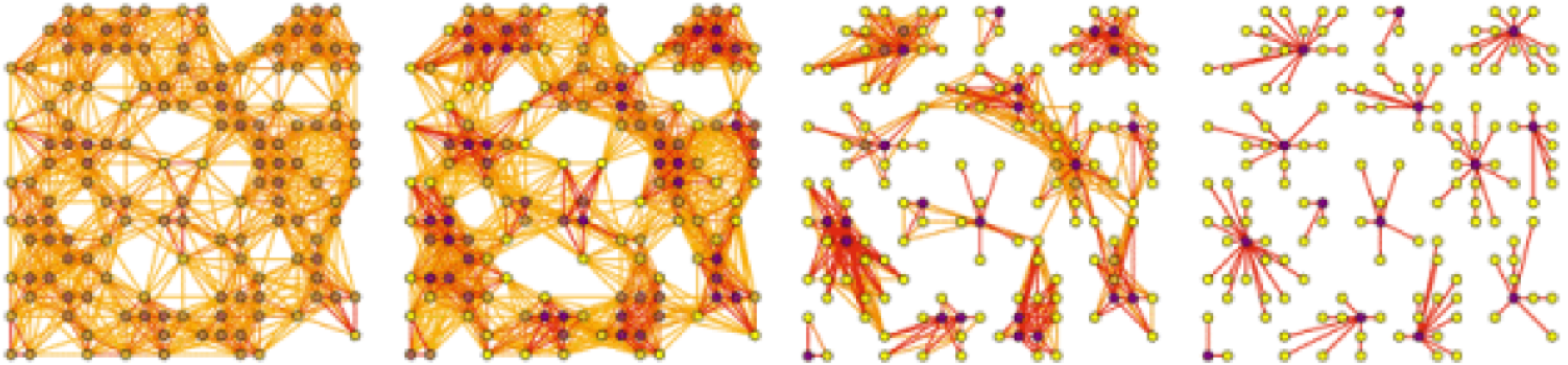
All-by-all homology Inference

All-by-all BLAST search

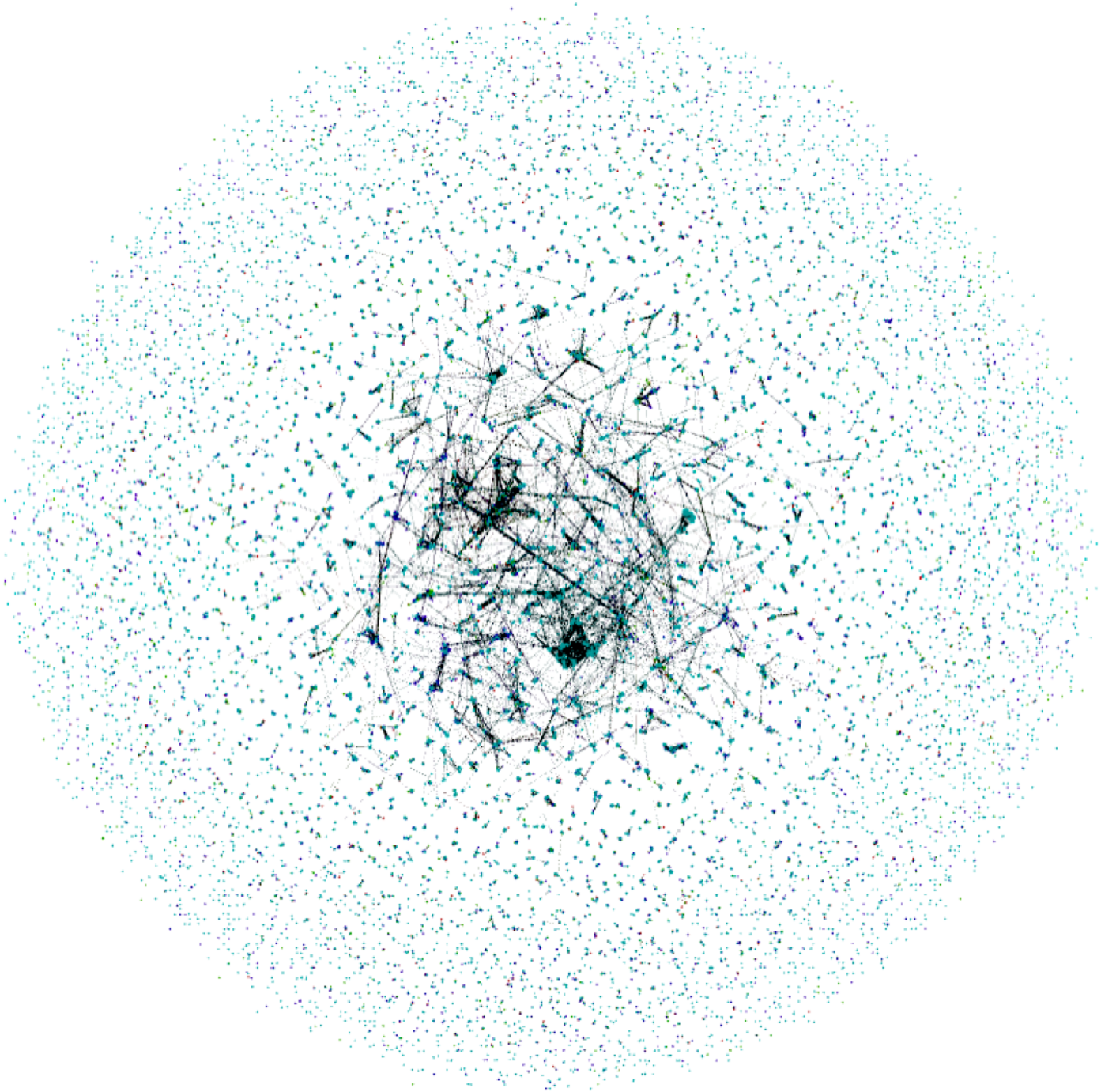
Filter BLAST results

- Yang and Smith, 2014, MBE: minimal filtering
- orthoFinder (improved upon orthoMCL): normalize BLAST scores by gene length and phylogenetic distance; keeping the reciprocal best hit pairs only

Markov Cluster Algorithm (MCL; van Dongen 2000)



http://micans.org/mcl/



All-by-all blast

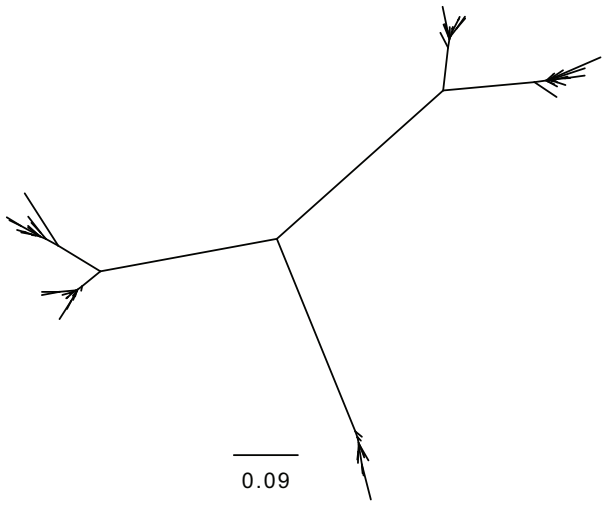
Filter blast results by hit fraction

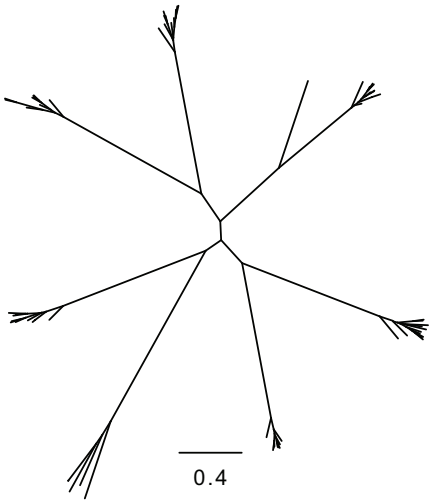
Markov Clustering

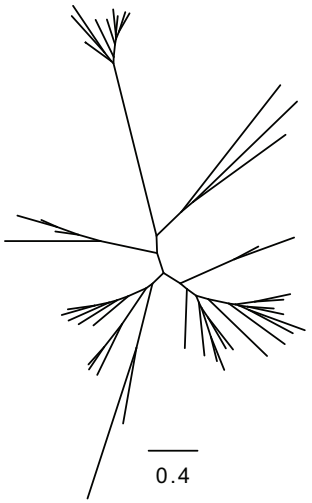
Bait SWIPE against each proteome

Filter top hits within each proteome

Alignment and tree inference





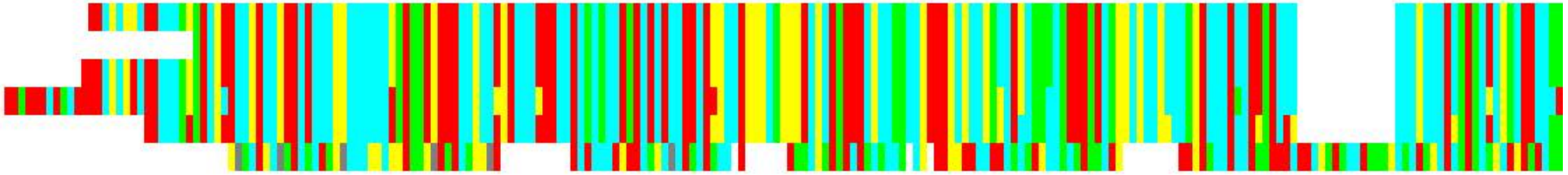


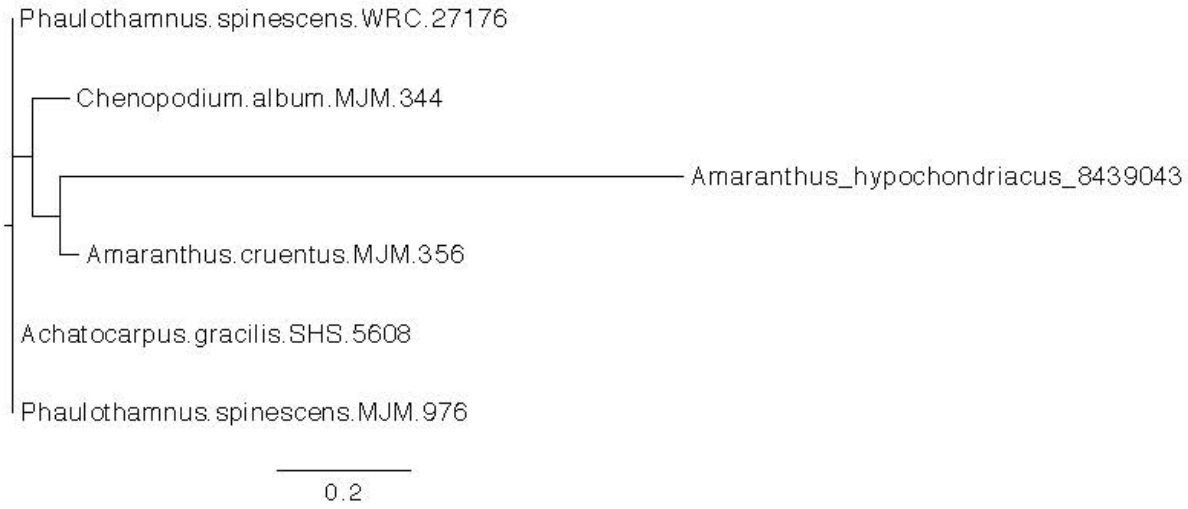
Cut spurious tips and mask tips

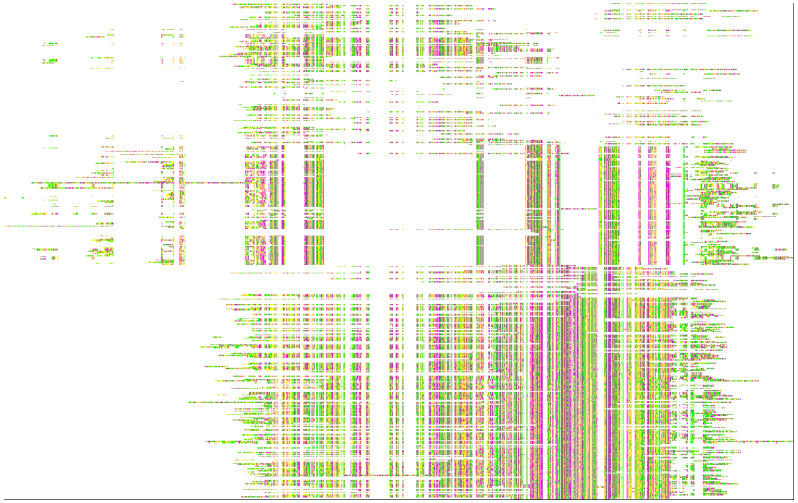
Separate long internal branches

Refining homologs

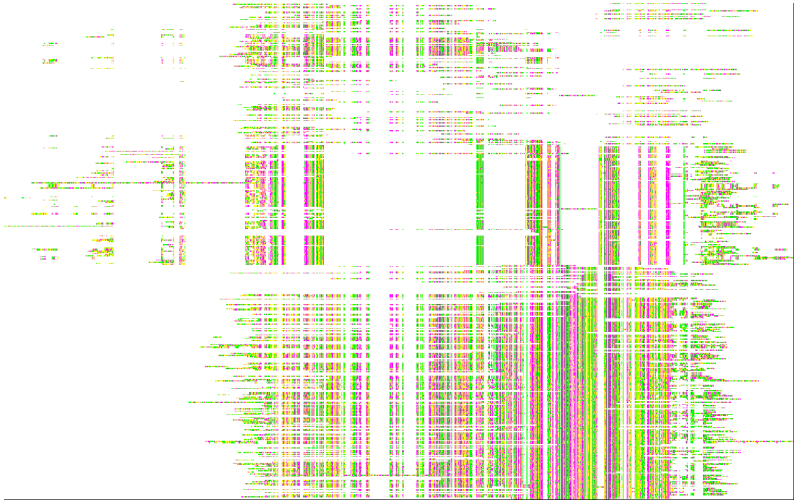
Trim spurious tips with TreeShrink

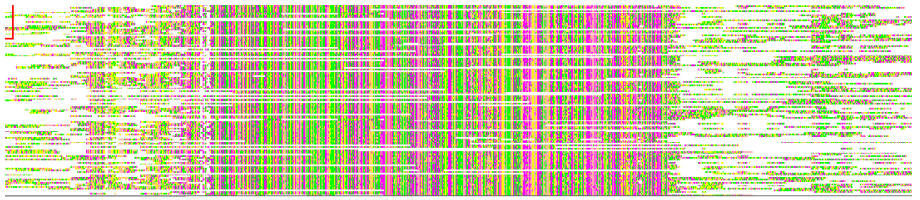


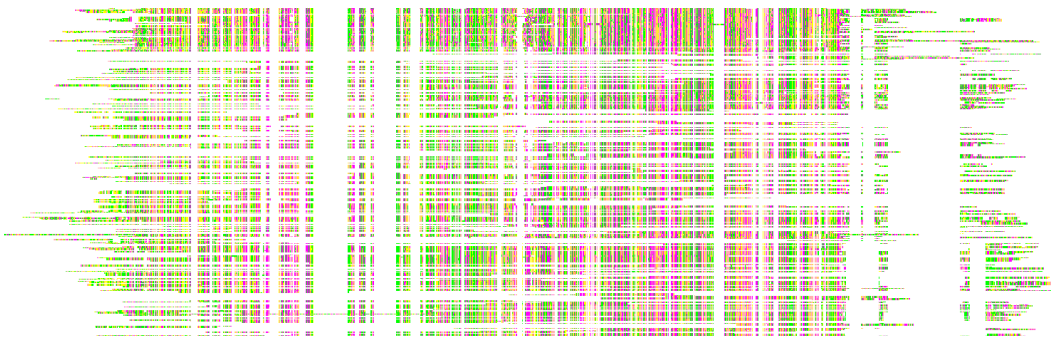




Cut long internal branches







Homology Inference

All-by-all blast

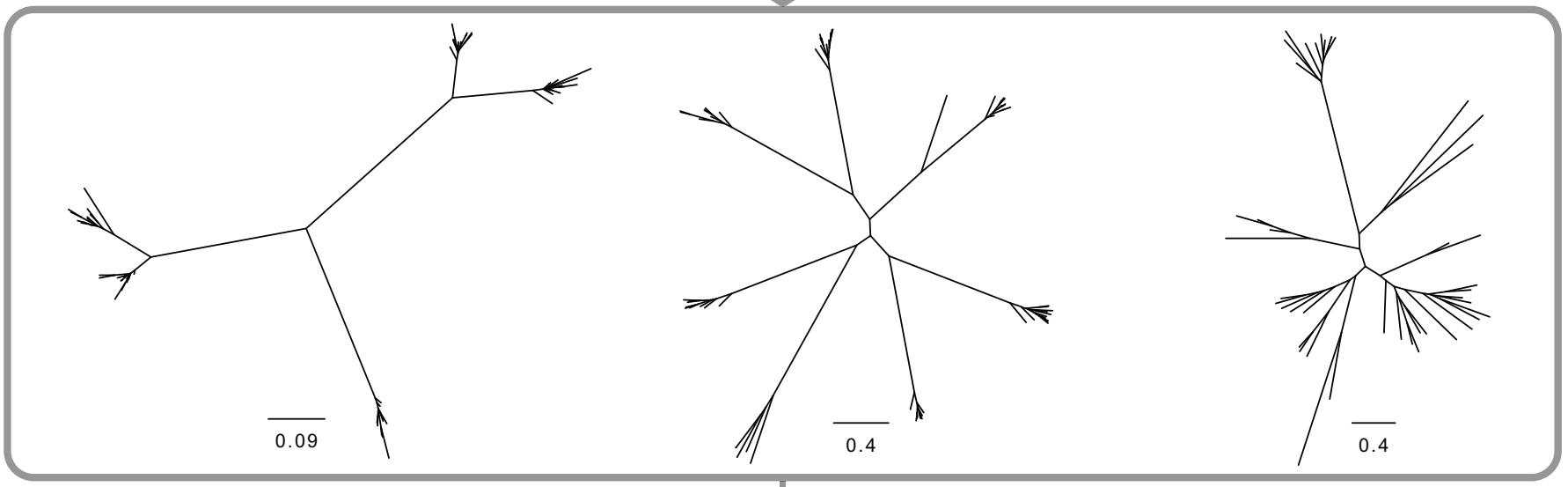
Filter blast results by hit fraction

Markov Clustering

Bait SWIPE against each proteome

Filter top hits within each proteome

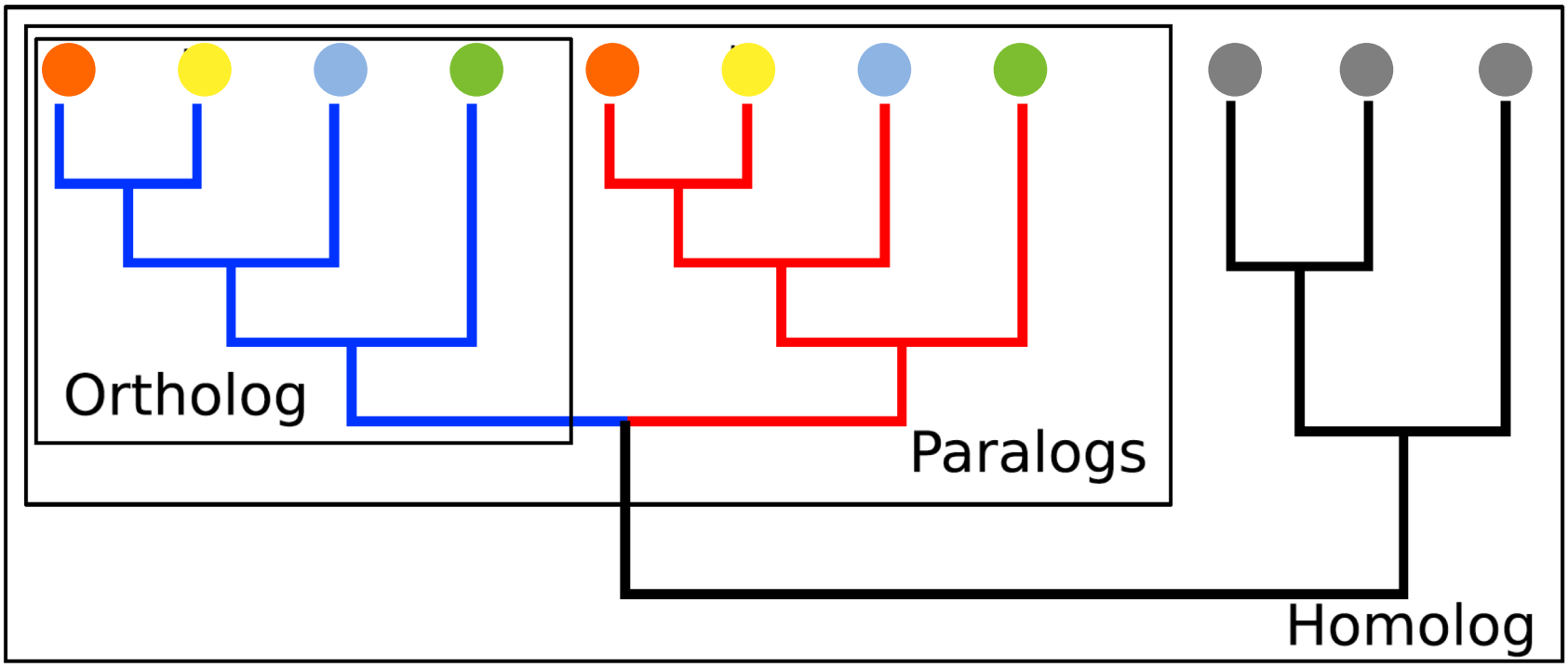
Alignment and tree inference



Cut spurious tips and mask tips

Separate long internal branches

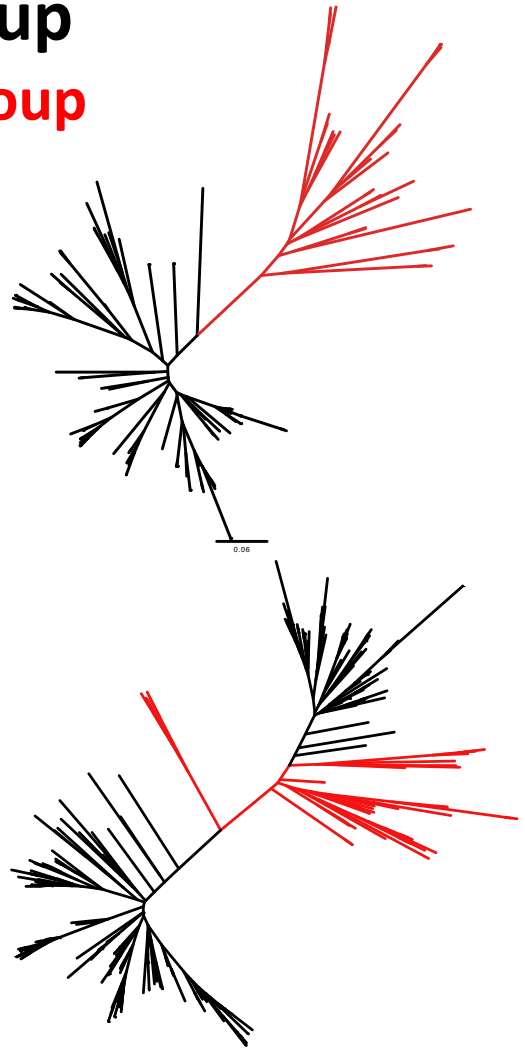
Homolog → ortholog



Tree-based approach maximizes information retained

Ingroup

Outgroup

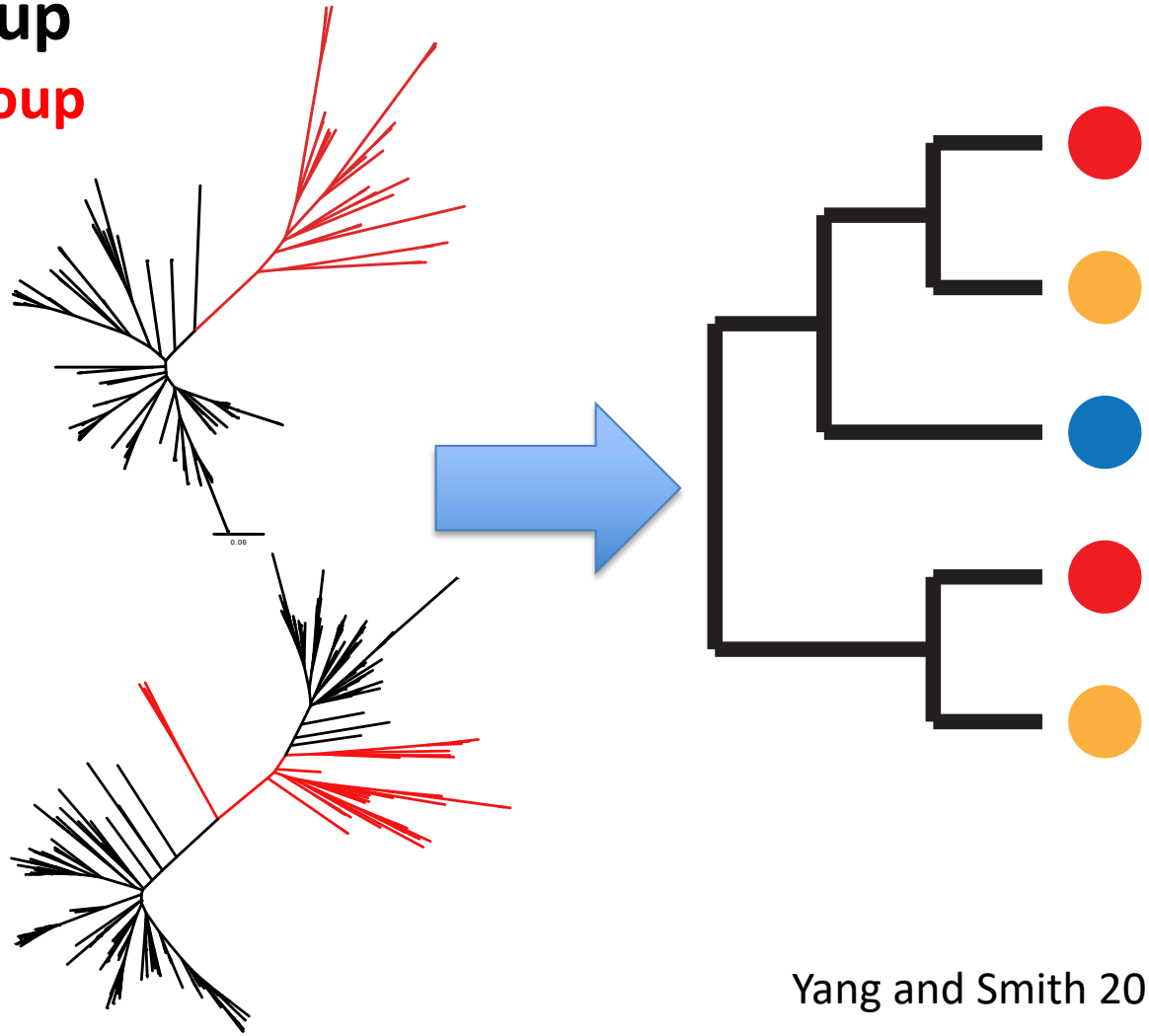


Yang and Smith 2014 *Mol Biol Evol*

Tree-based approach maximizes information retained

Ingroup

Outgroup

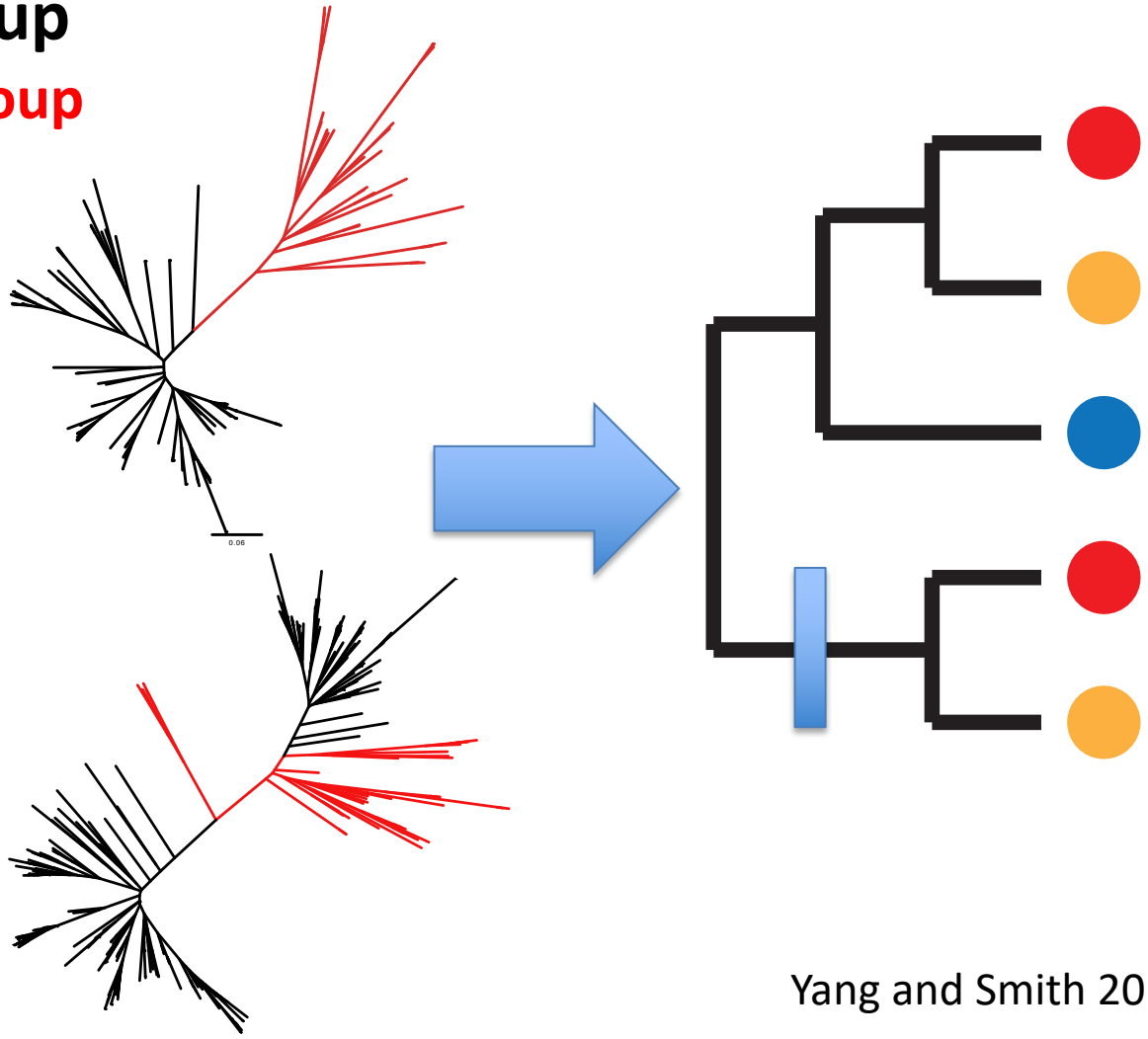


Yang and Smith 2014 *Mol Biol Evol*

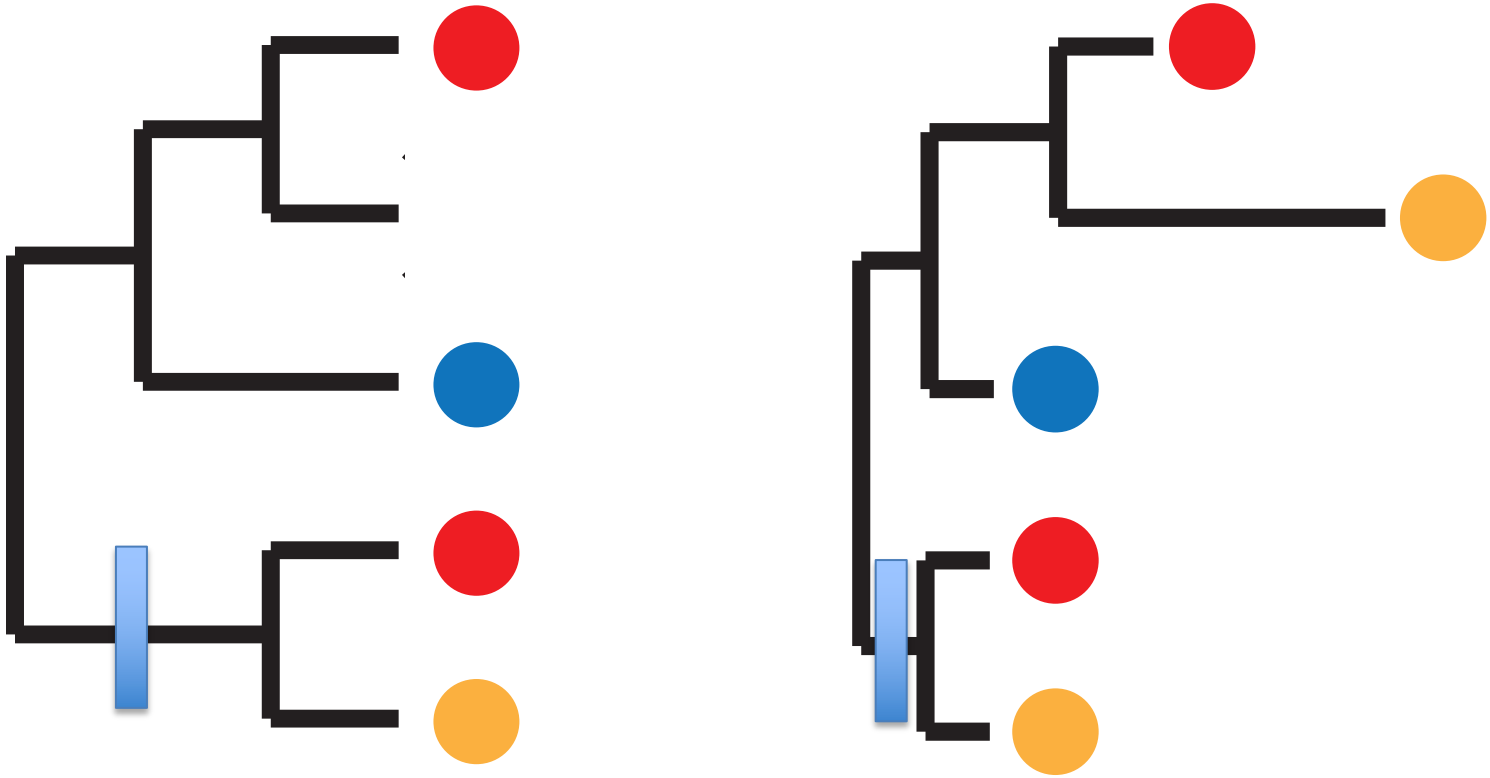
Tree-based approach maximizes information retained

Ingroup

Outgroup



Yang and Smith 2014 *Mol Biol Evol*

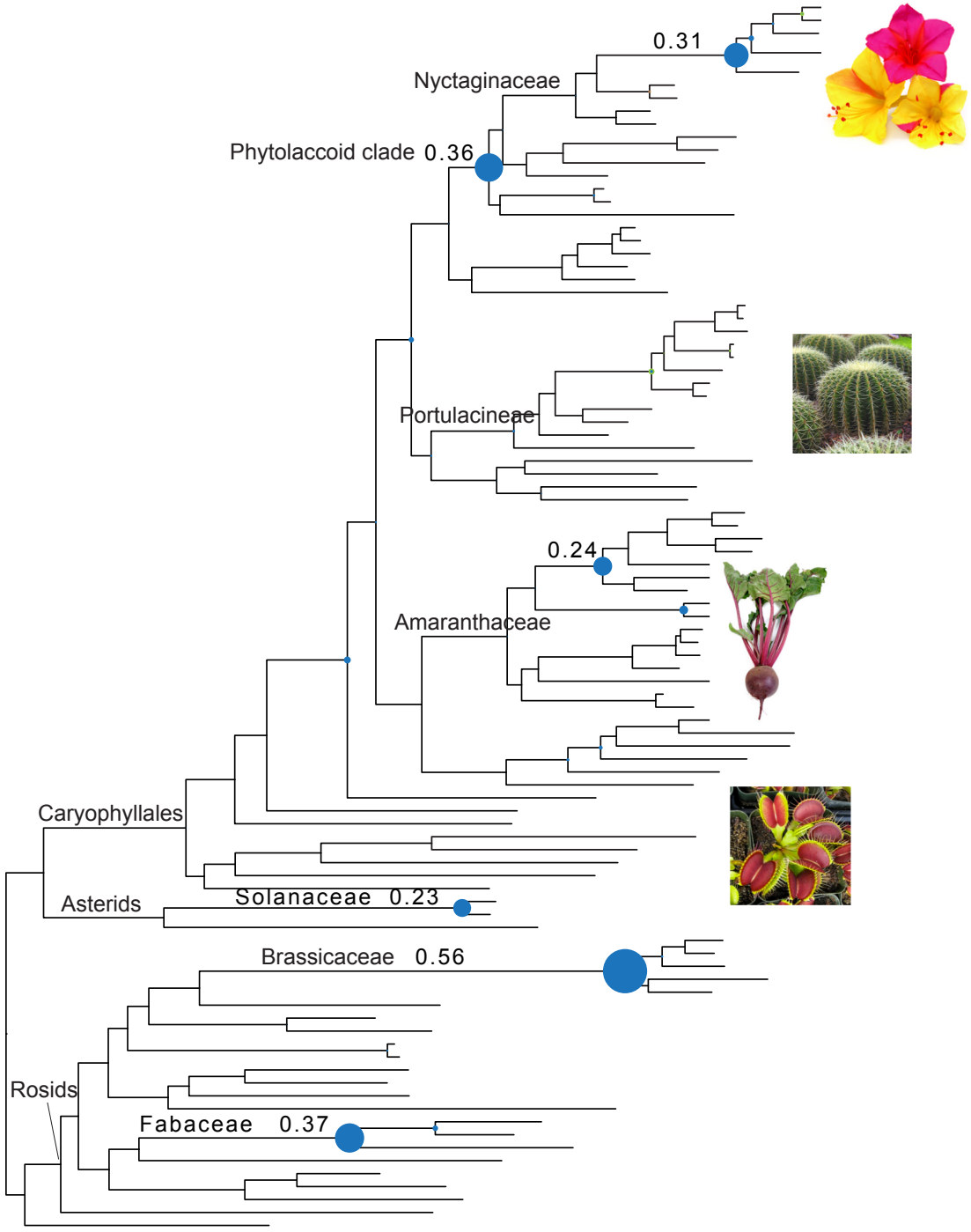


Open source python code, tutorials and test data are available from bitbucket

bitbucket.org/yangya/phylogenomic_dataset_construction

Yang and Smith 2014 *Mol Biol Evol*

Detecting genome duplication events

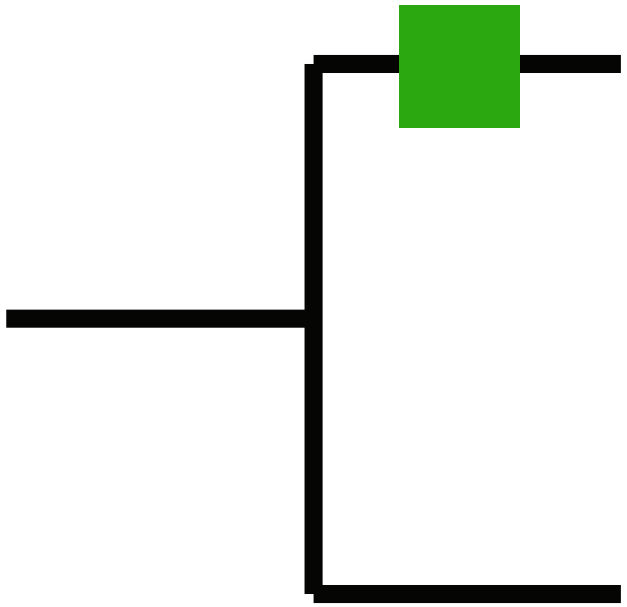


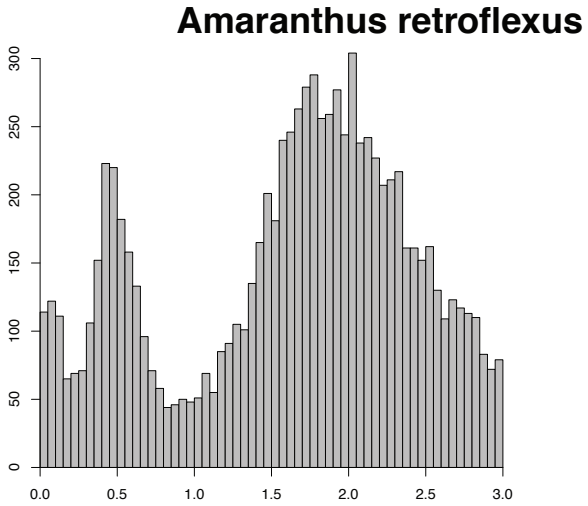
Six phylogenetic hotspots of gene duplications

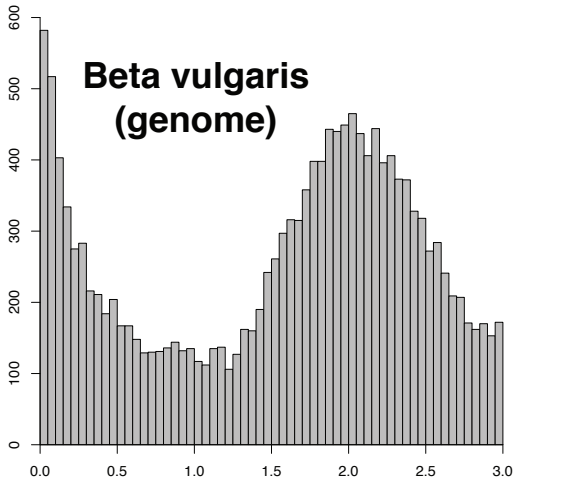
Yang et al. 2015 *Mol Biol Evol*

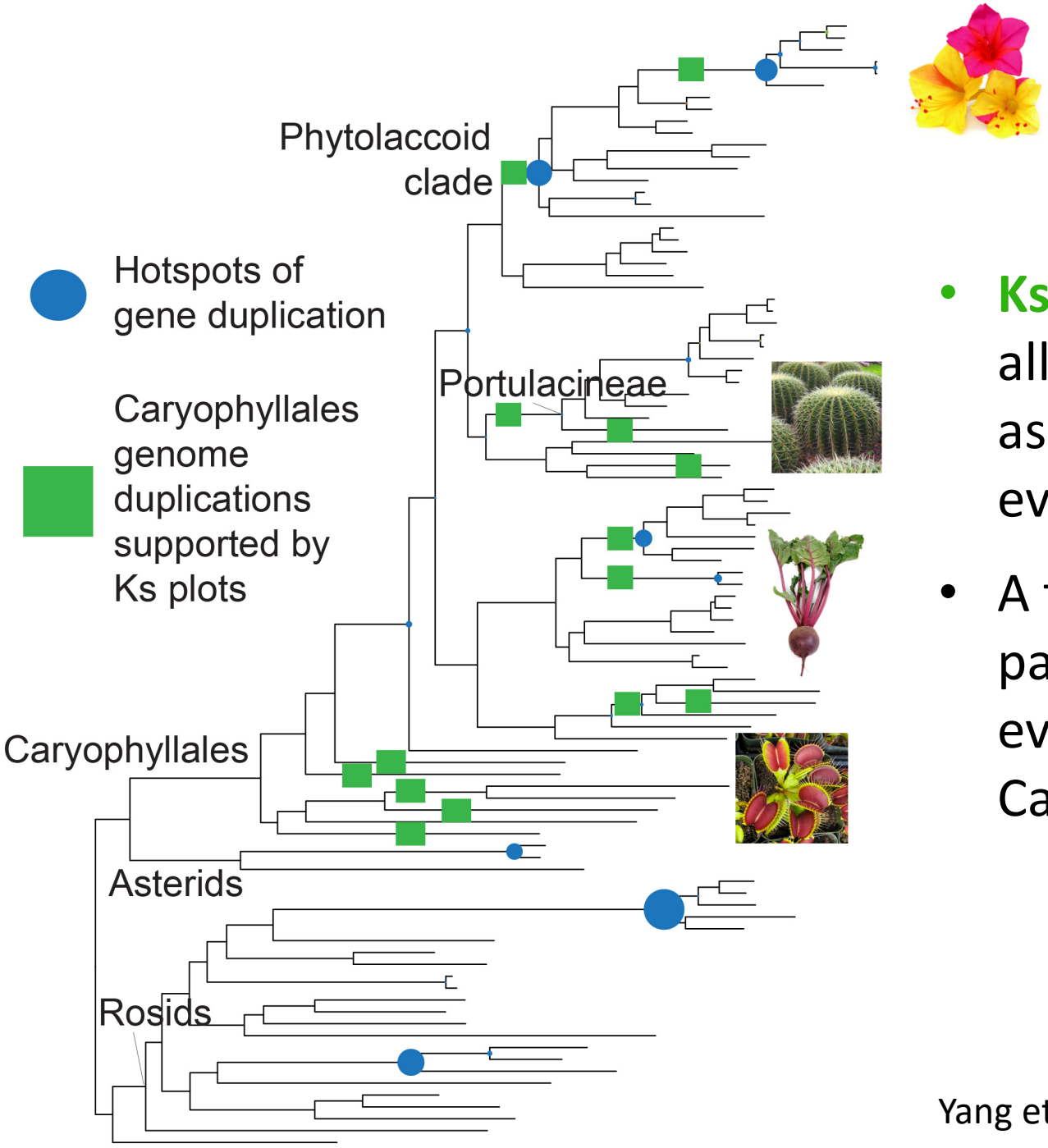
Ks plots are combined with tree topologies to locate WGDs

WGD duplication









- **Ks plots** confirmed all hotspots mapped as paleopolyploidy events
- A total of 13 paleopolyploidy events in Caryophyllales

Yang et al. 2015 *Mol Biol Evol*

Why we did not write a standalone software for homology and orthology inference

Homology and orthology inference methods

- Based on all-by-all homology search
  - Yang and Smith, 2014, MBE: **gene family**
  - orthoMCL and orthoFinder: **low-copy genes**
- Homology search using a reference gene set or clusters from annotated genomes
  - HMM or Phylome **> 50 species**
- Hierarchical clustering: **>200 species**