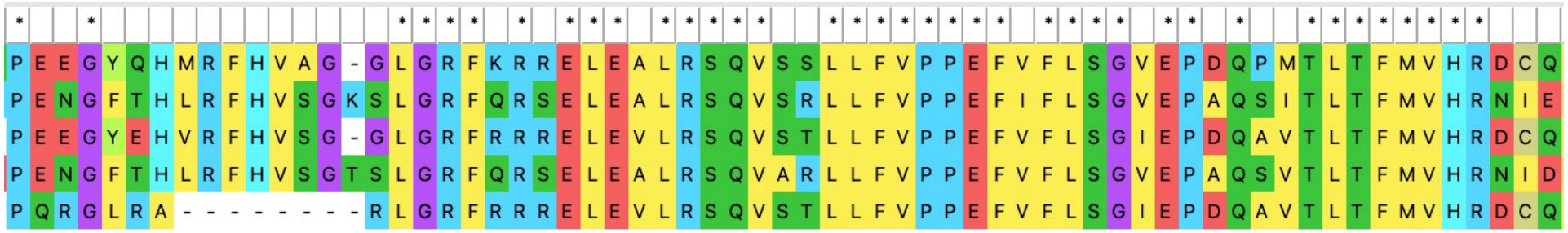
Multiple Sequence Alignment: a practical lesson



BIOL 435/535: Bioinformatics

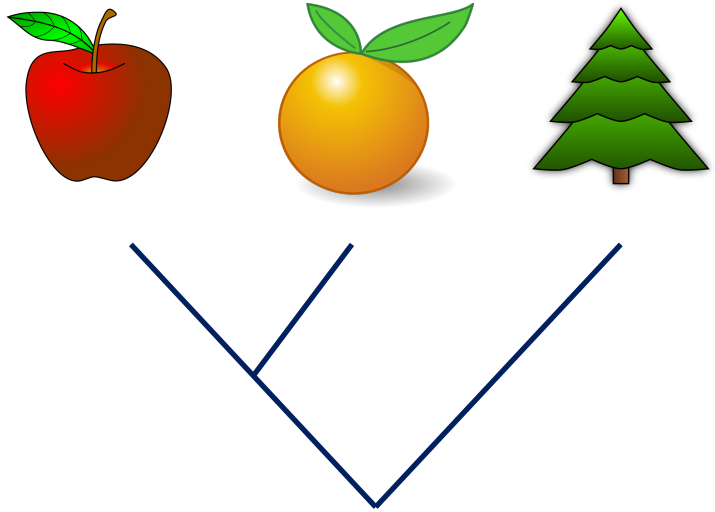
January 27, 2022

Any luck/roadblocks with pairwise matrices?

What can multiple sequence alignments get you that pairwise alignments can't?

What can multiple sequence alignments get you that pairwise alignments can't?

- Population genetics (any application in which N needs to be >> 2)
- Ancestral state reconstruction
- Estimating rates/patterns of natural selection
- Inferring species relationships
- Population structure/haplotype networks
- Identifying adaptive radiations
- Estimating mutation rates
- Variant analysis



Important parameters/considerations

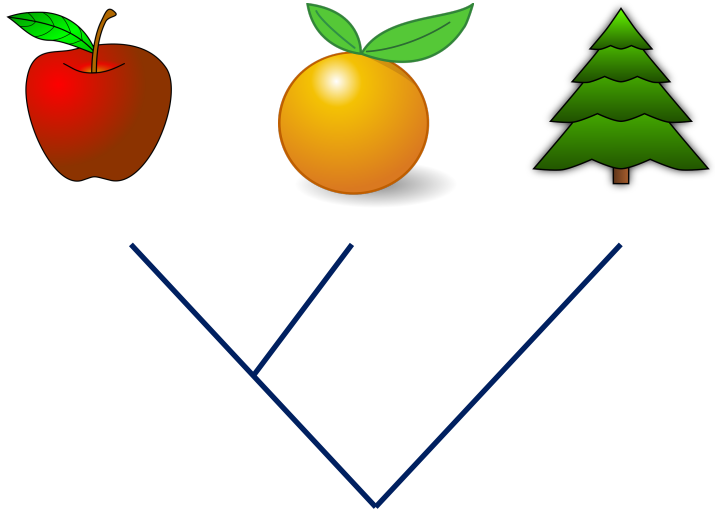
- Match, mismatch, gap scoring system
- Algorithm (e.g., clustal, MUSCLE, global, local, long gaps)
- Nucleotide/codon/protein

Post alignment inspection

- If fewer than 100 genes/1000 species, alignments need to be manually inspected
- Re-align gappy/difficult regions
  - **Be careful to not alter reading frame if working in codon space!**
  - Trim out if no easy fix
- Bigger datasets, look at **synonymous** rates of evolution – can tell you whether you have a bad alignment

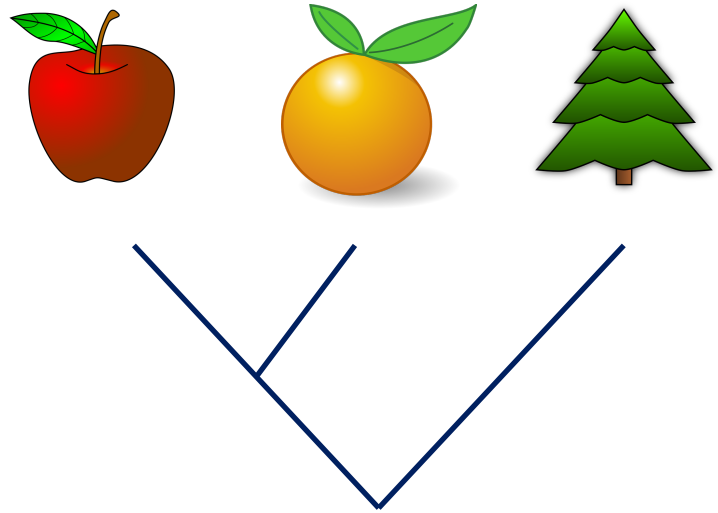
Useful tools for performing MSAs

- Mafft
- MEGA
- MUSCLE
- Clustal



Useful tools for trimming MSAs

- GBlocks
- TrimAL
- Clipkit



Let’s give it a shot!

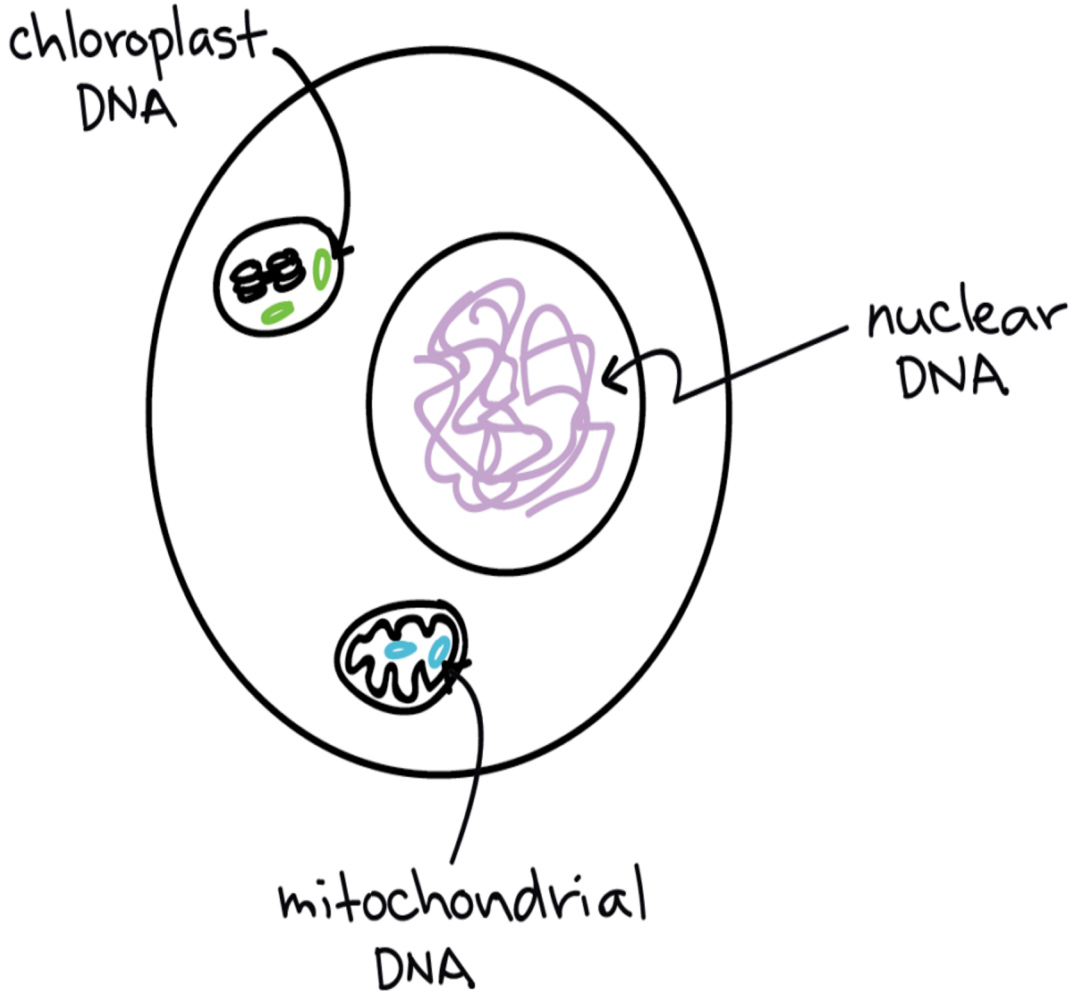
Download from GitHub Activities folder:

- MSA.nuc.fasta
- MSA.prot.fasta

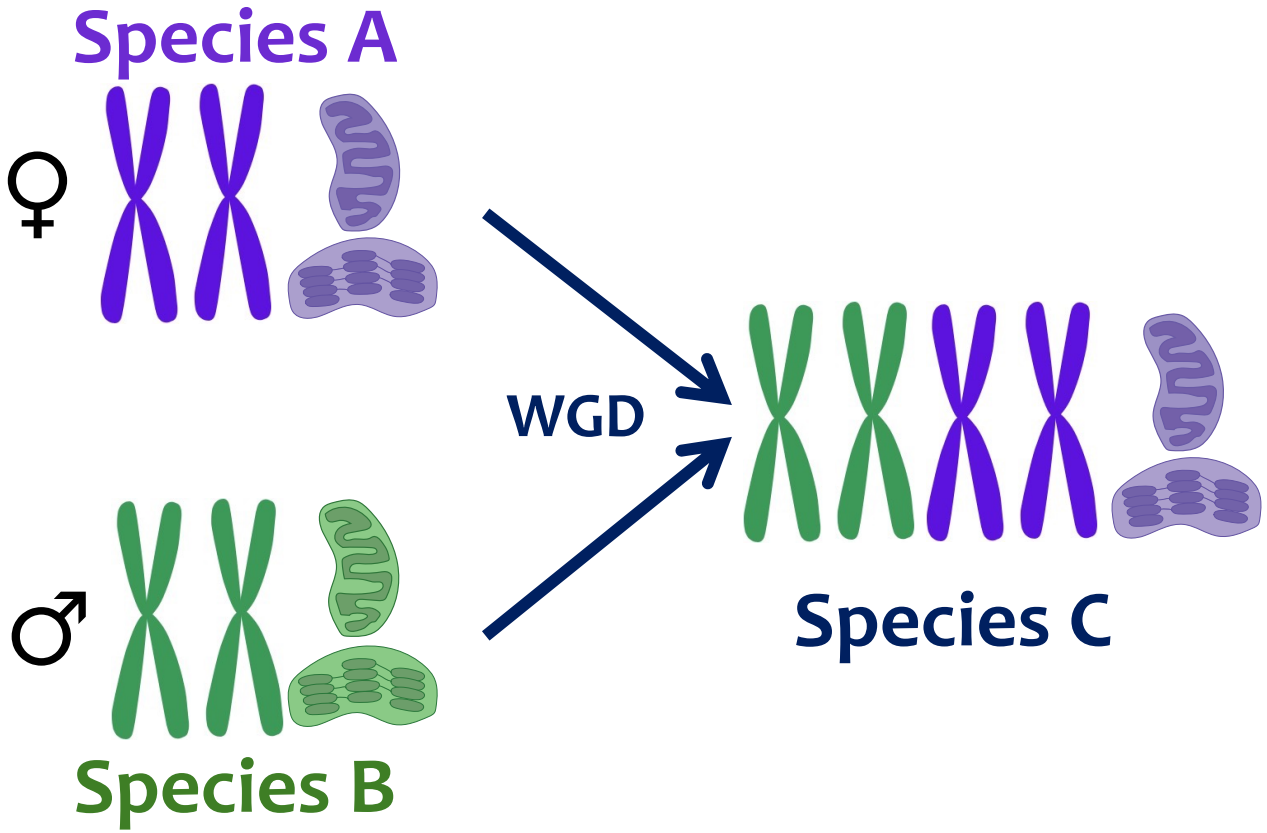
MEGA (or use online Mafft tool)

Trim using Gblocks

Bad alignments: a cautionary tail

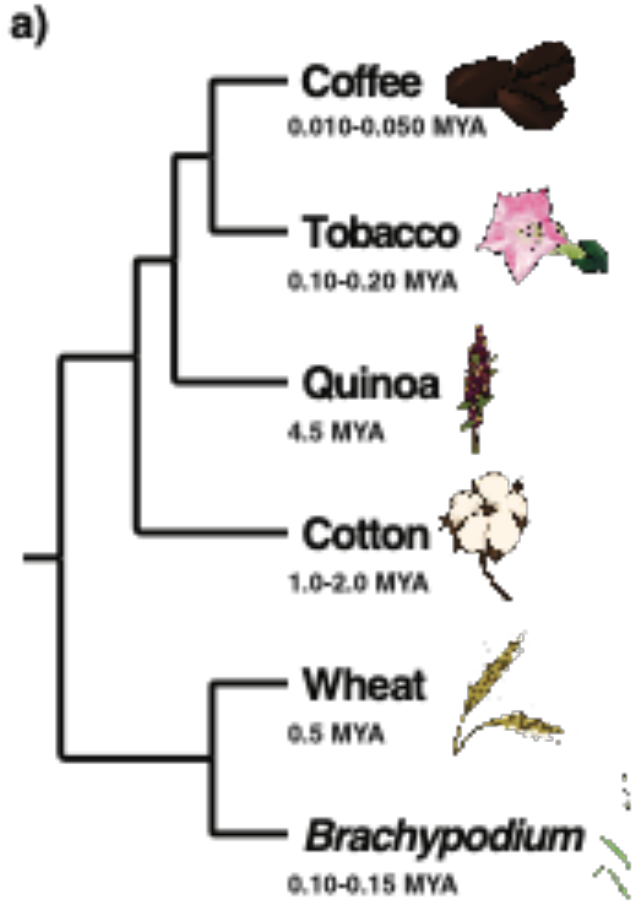


Many polyploids are the result of hybridization – "genome merger"



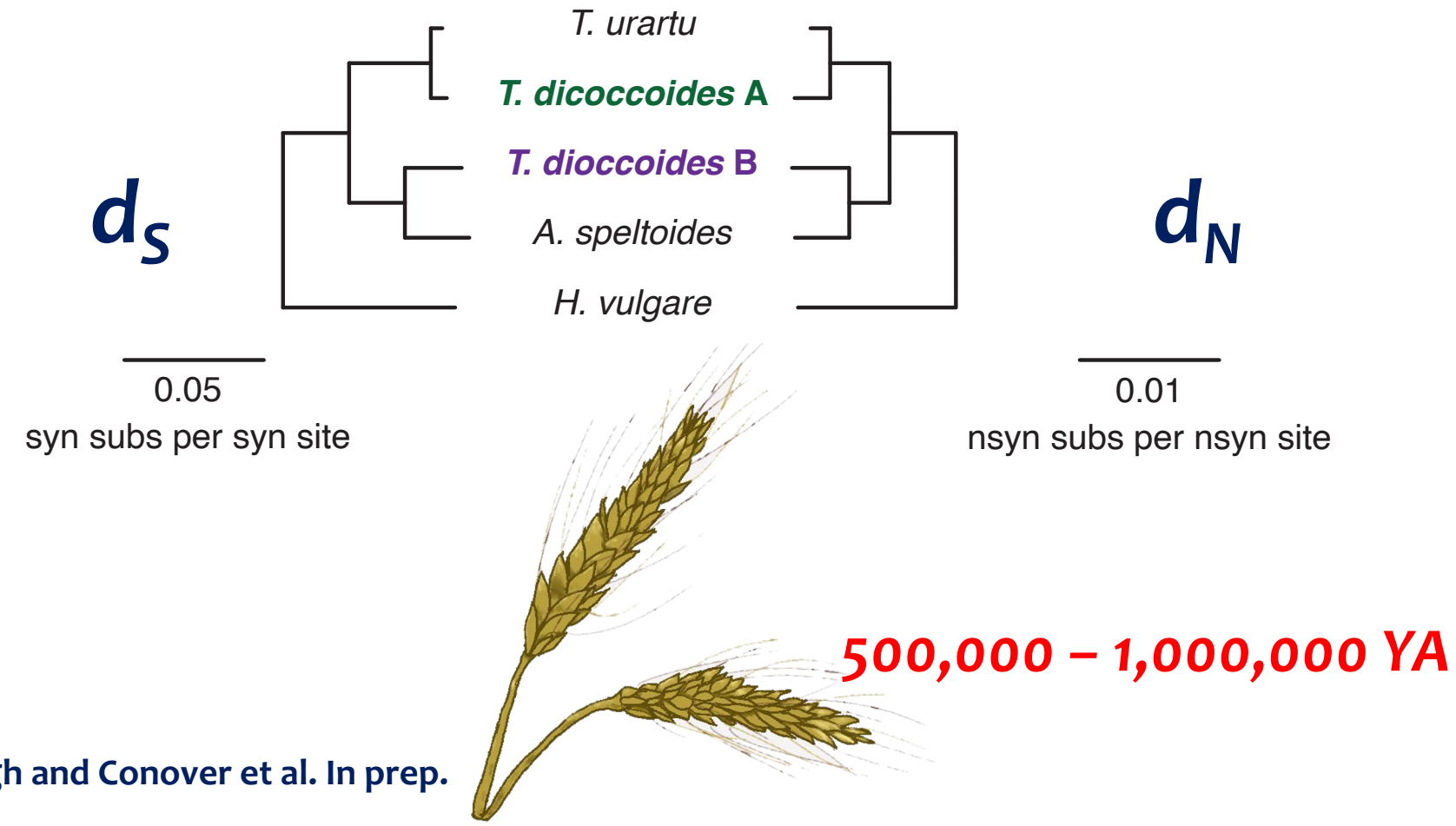
Sharbrough et al. 2017. Am. J. Bot.

Genome-wide effects of hybridization-induced polyploidy in a diverse set of allopolyploid plants



Sharbrough and Conover et al. In prep.

Genome-wide effects of hybridization-induced polyploidy in a diverse set of allopolyploid plants

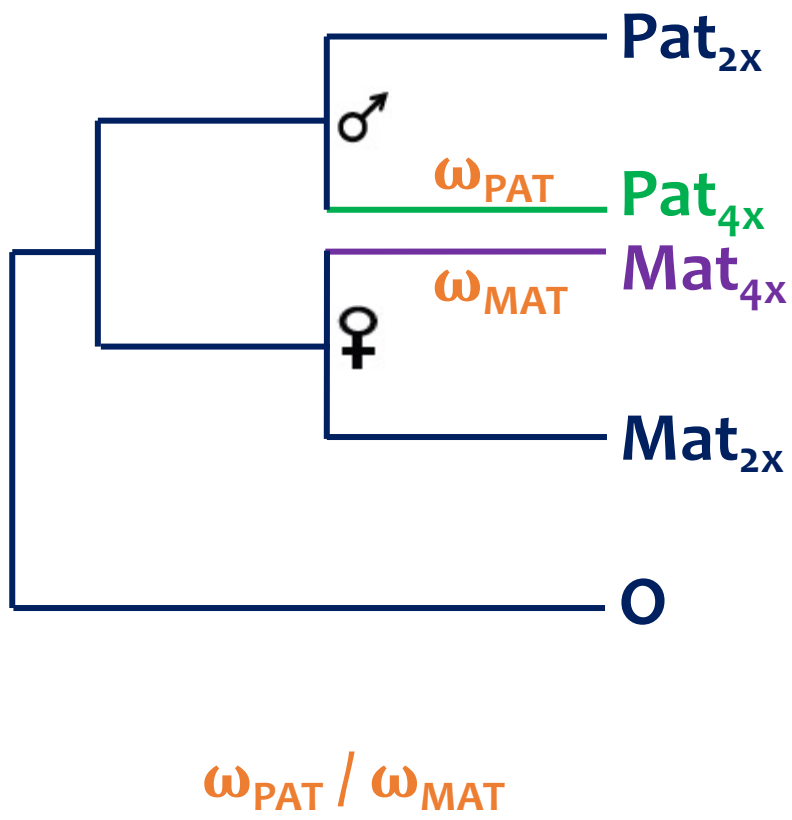


Sharbrough and Conover et al. In prep.

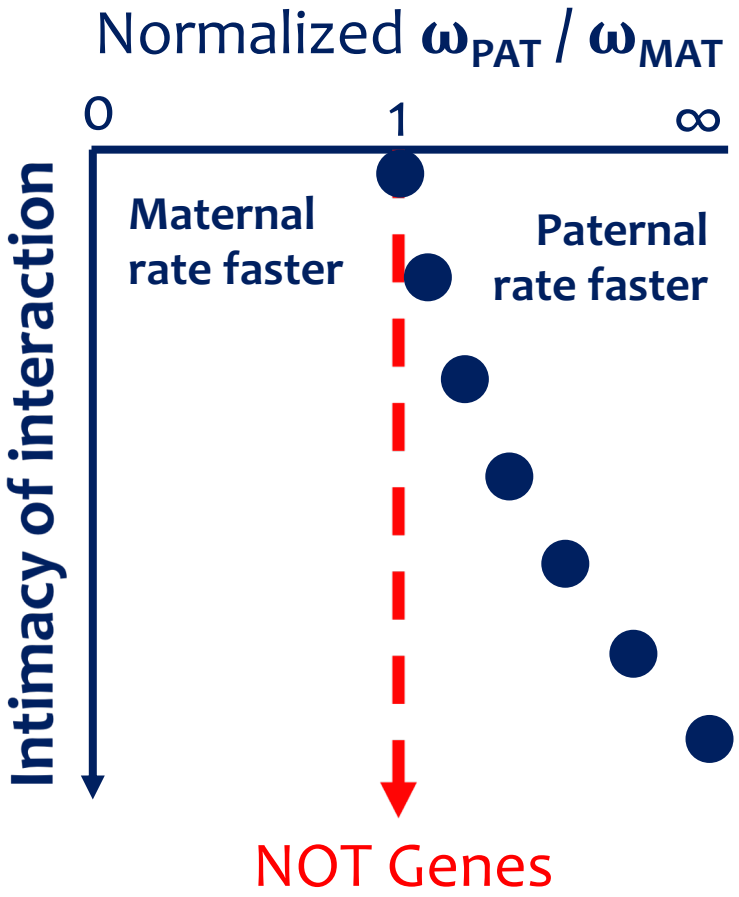
Evolutionary mismatches between paternal subgenome and cytoplasmic genomes give rise to cytonuclear incompatibilities in hybrid polyploids

- Relaxed selection in paternal copy
- Compensatory co-evolution "fixing" paternal copy

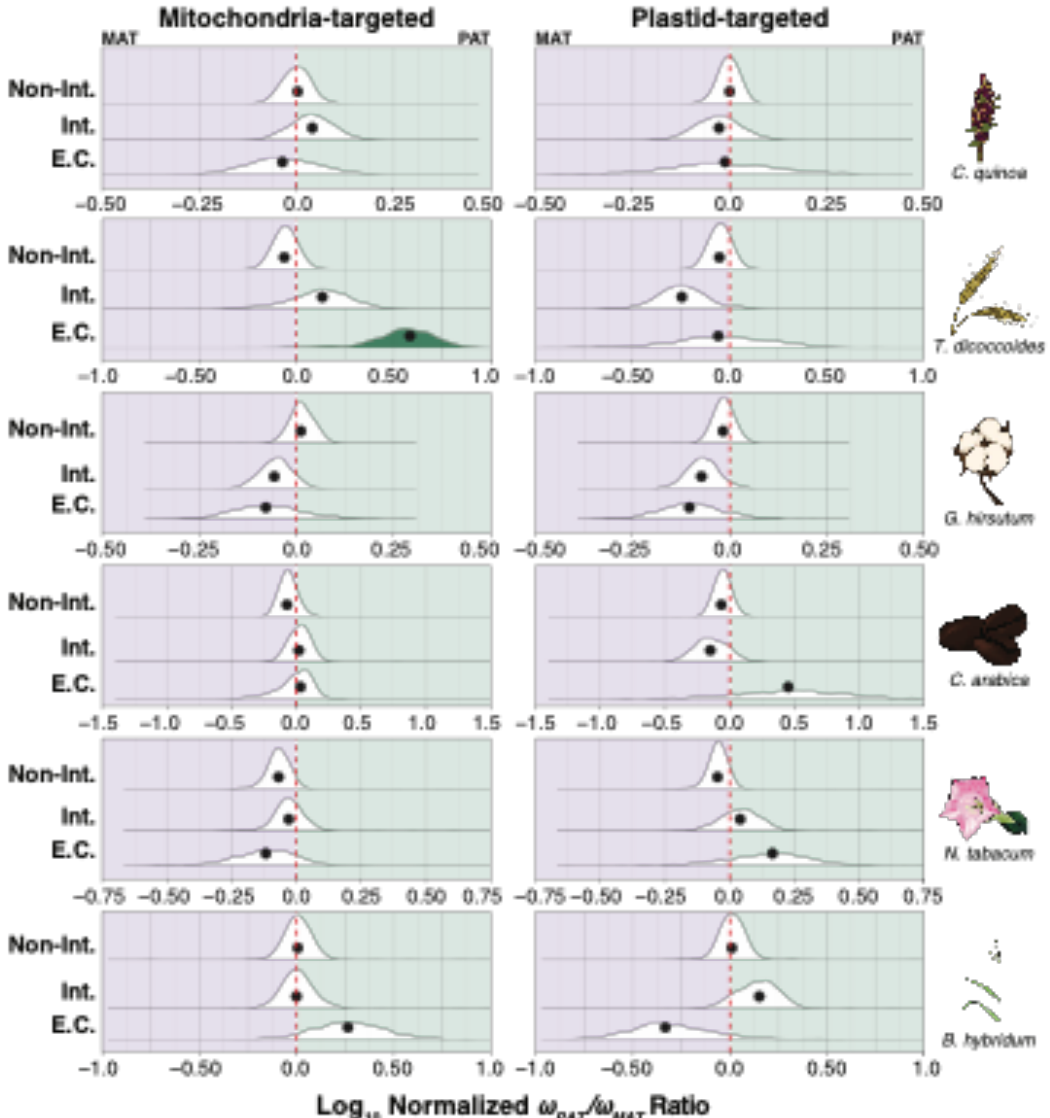
Accelerated rate of protein sequence evolution (ω) in paternal copies of organelle-interacting genes



Evolutionary rate across subgenomes in genes targeted to the organelles



Evolutionary rate across subgenomes in genes targeted to the organelles

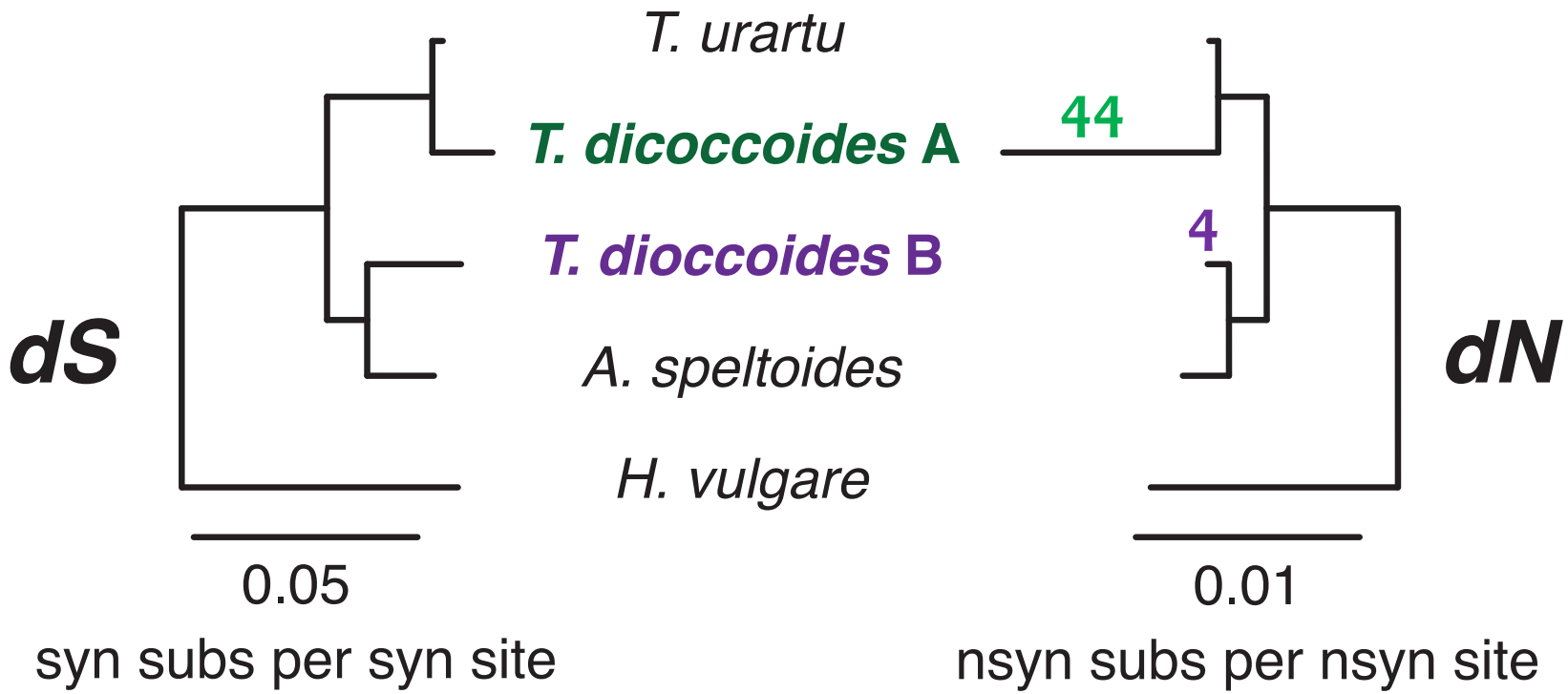


ω_{PAT} = rate of protein sequence evolution in paternal subgenome

ω_{MAT} = rate of protein sequence evolution in maternal subgenome

Bootstrap distributions with <2.5% overlap over 1 shaded

Paternal subgenome appears to have many more amino acid changes than maternal subgenome



Numbers above branches reflect the # of derived amino acids at conserved positions

Except… it was all due to a bad alignment

Except… it was all due to a bad alignment

Signatures of bad alignments

- Multiple amino acid changes in a row from a single sequence
- Higher rate of synonymous change than other genes
- Lots of gaps/frameshifts if working in nucleotide space
- Disagreement between nucleotide/protein alignments

Next up: Gene Architecture & Gene Discovery

Please Read: Jordan & Goldman 2012 (Introduction)