Phasing alleles improves network inference with allopolyploids

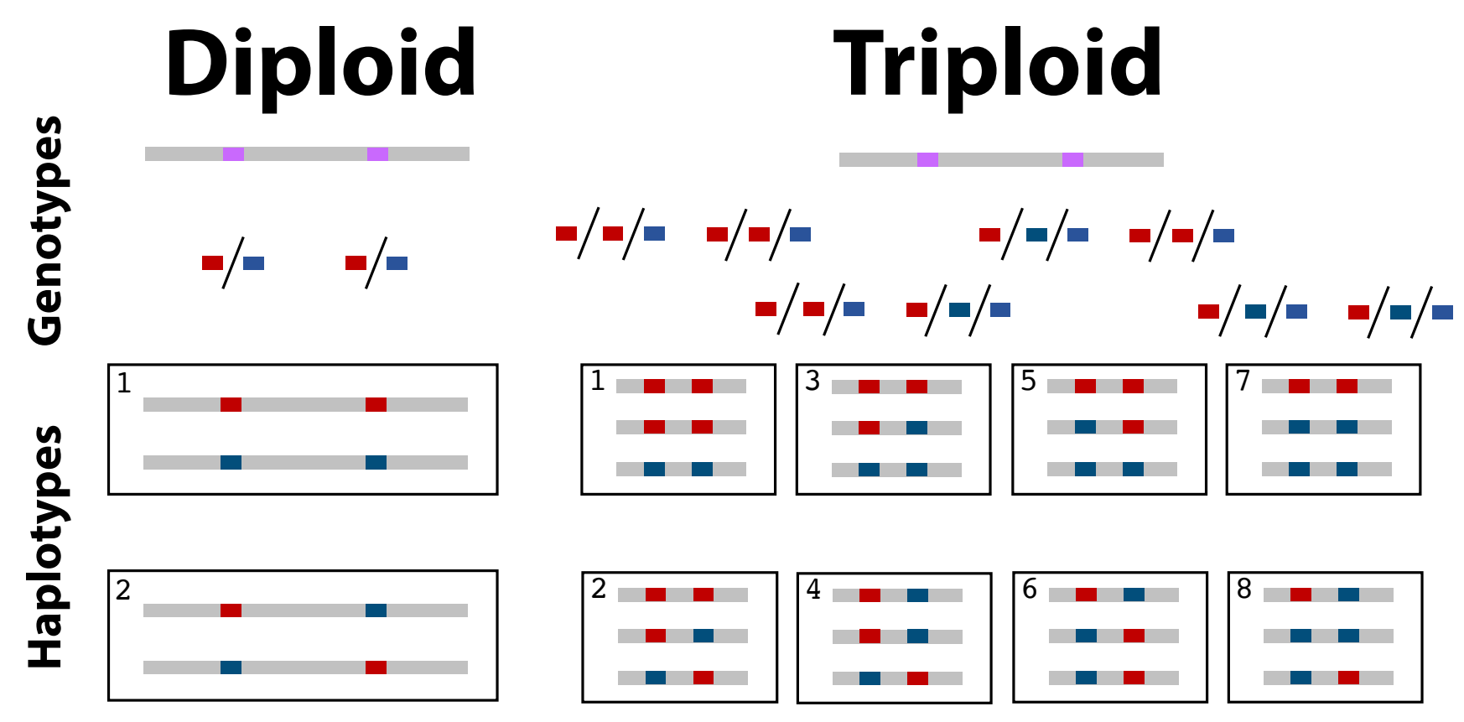
George P. Tiley[1,†], Andrew A. Crowl[2], Paul S. Manos[2], Emily B. Sessa[3], Claudia Solís-Lemus[4], Anne D. Yoder[2], J. Gordon Burleigh[3]

[1]Herbairum, Royal Botanic Gardens Kew, Richmond TW9 3AE, UK †g.tiley@kew.org
[2]Department of Biology, Duke University, Durham NC 27708, USA
[3]Department of Biology, University of Florida, Gainesville FL 32611, USA
[4]Department of Plant Pathology, University of Wisconsin – Madison, Madison WI, 53706, USA

Funded by the European Union

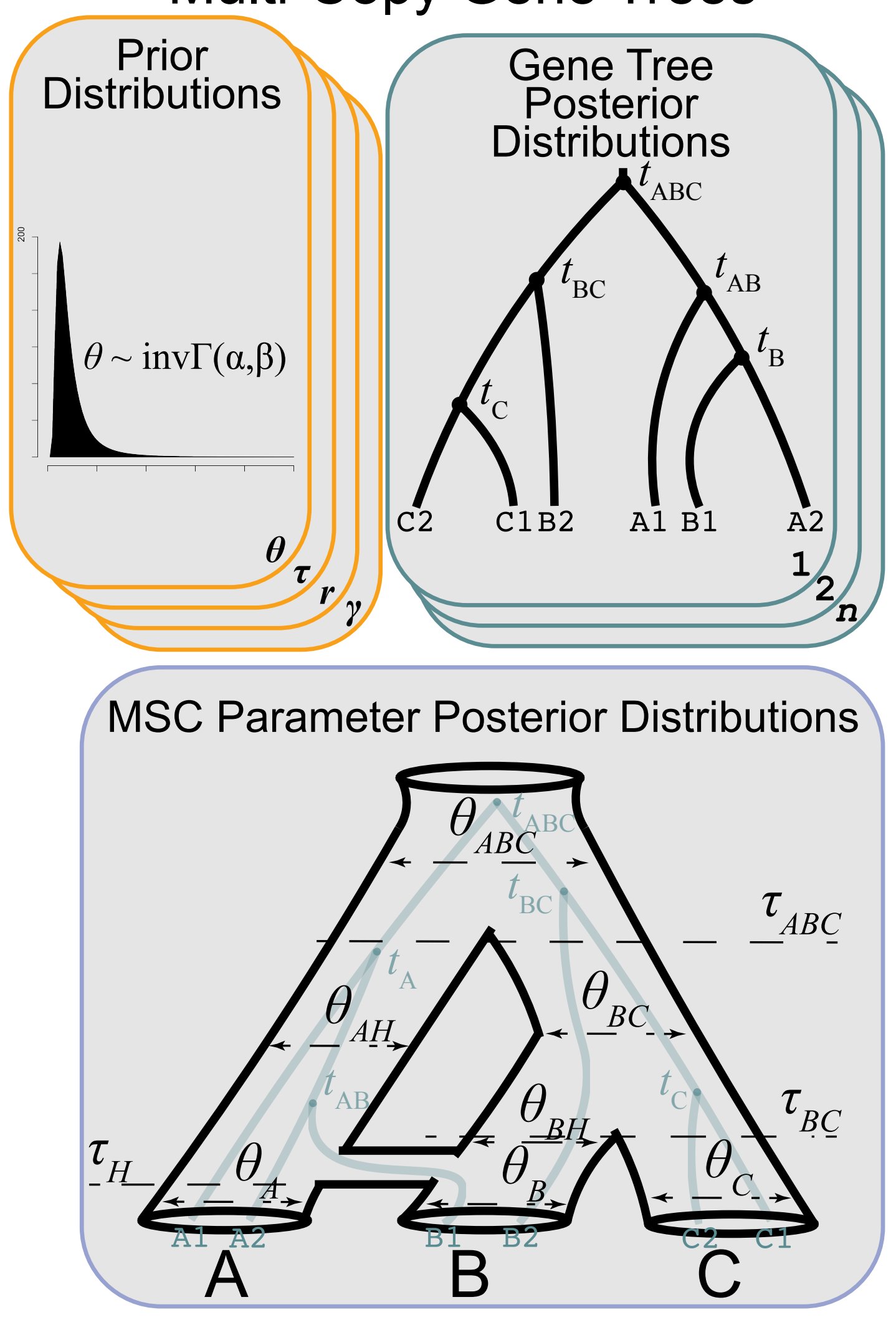
Reticulate evolution is common

We want to estimate networks from sequence data. Polyploids are complicated unless we can recover haplotype sequences. The number of haplotypes given n biallelic variants at k ploidy is $2^{n-1}(k-1)^n$



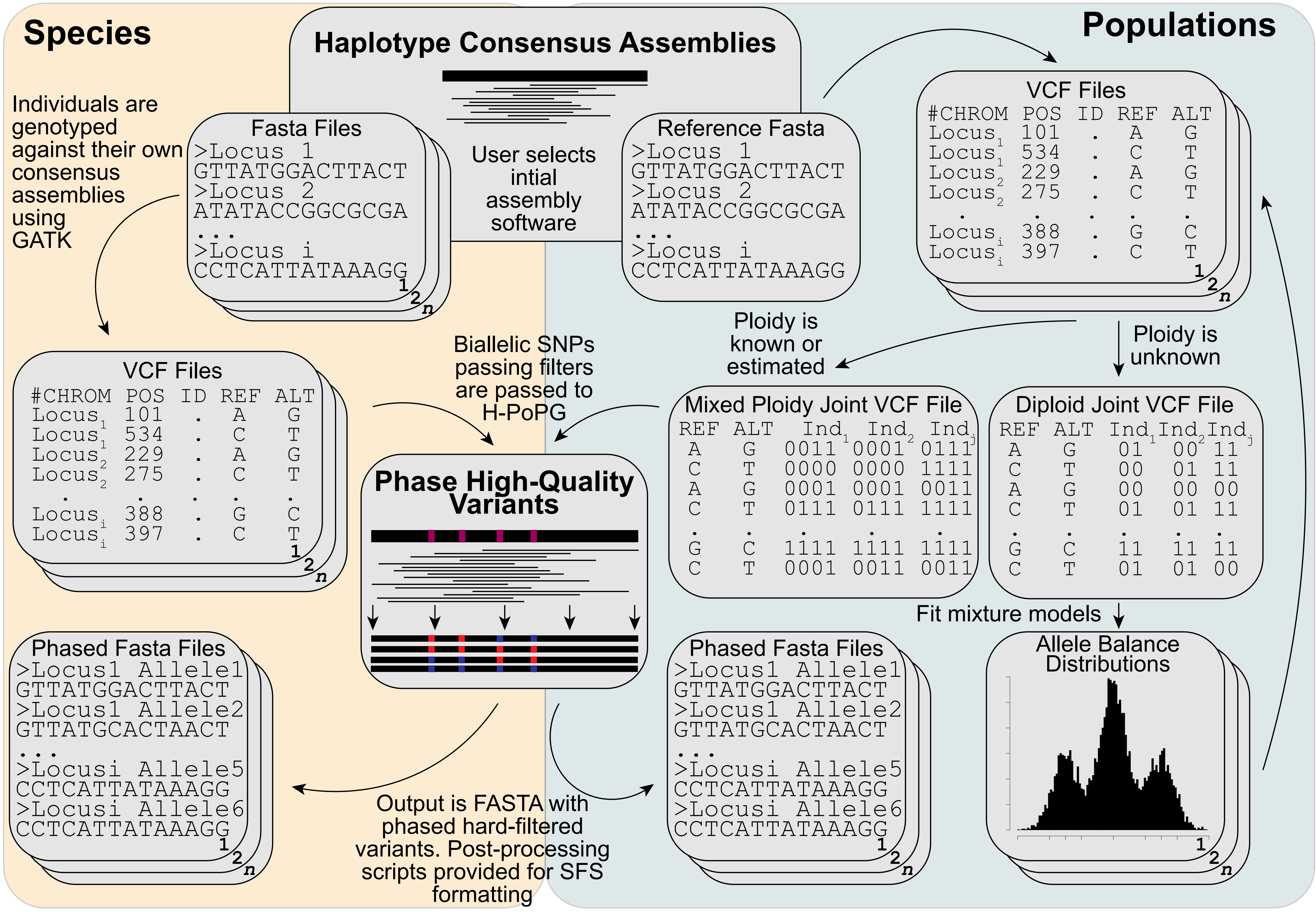
PATÉ can bridge the gap between observed data and model assumptions

Estimating Species Networks with Multi-Copy Gene Trees



PATÉ: Phased Alleles from Target Enrichment data

A low-dependencey pipeline for genotyping and phasing



Species

Populations

Haplotype Consensus Assemblies

Individuals are genotyped against their own consensus assemblies using GATK

User selects intial assembly software

Fasta Files

```
>Locus 1
GTTATGGACTTACT
>Locus 2
ATATACCGGCGCGA
...
>Locus i
CCTCATTATAAAGG
```

Reference Fasta

```
>Locus 1
GTTATGGACTTACT
>Locus 2
ATATACCGGCGCGA
...
>Locus i
CCTCATTATAAAGG
```

VCF Files

```
#CHROM  POS  ID REF ALT
Locus1  101  .  A   G
Locus1  534  .  C   T
Locus2  229  .  A   G
Locus2  275  .  C   T
  .      .   .  .   .
Locusi  388  .  G   C
Locusi  397  .  C   T
```

Biallelic SNPs passing filters are passed to H-PoPG

Ploidy is known or estimated

Ploidy is unknown

Mixed Ploidy Joint VCF File

```
REF ALT  Ind1 Ind2 Ind3
 A   G   0011 0001 0111
 C   T   0000 0000 1111
 A   G   0001 0001 0011
 C   T   0111 0111 1111
 .   .    .    .    .
 G   C   1111 1111 1111
 C   T   0001 0011 0011
```

Diploid Joint VCF File

```
REF ALT  Ind1 Ind2 Ind3
 A   G   01   00   11
 C   T   00   01   11
 A   G   00   00   00
 C   T   01   01   11
 .   .   .    .    .
 G   C   11   11   11
 C   T   01   01   00
```

Fit mixture models

Allele Balance Distributions

Phase High-Quality Variants

Phased Fasta Files

```
>Locus1 Allele1
GTTATGGACTTACT
>Locus1 Allele2
GTTATGCACTAACT
...
>Locusi Allele5
CCTCATTATAAAGG
>Locusi Allele6
CCTCATTATAAAGG
```

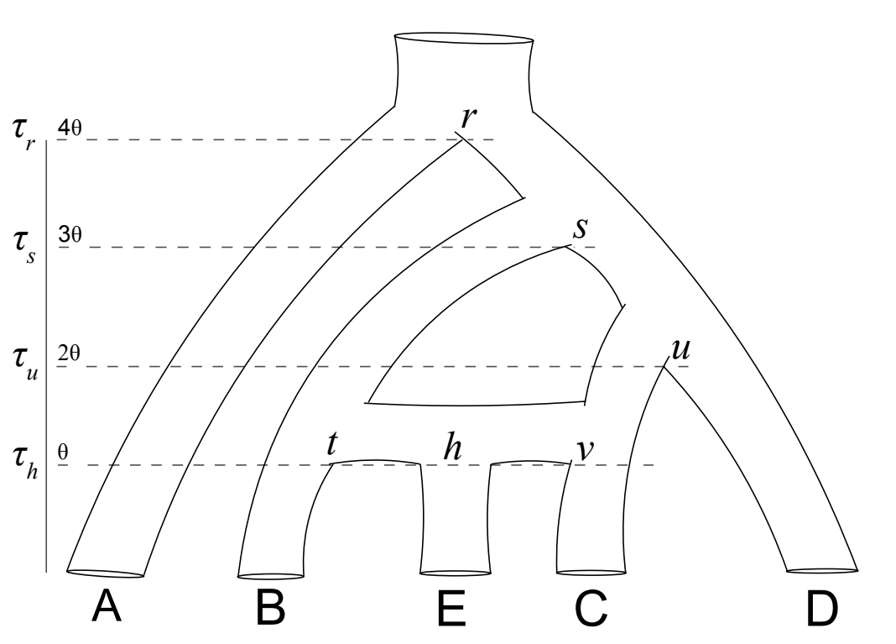
Output is FASTA with phased hard-filtered variants. Post-processing scripts provided for SFS formatting

Species– appropriate for cases where no reference outgroup sequence is available. This maximizes the amount of data retained per individual and is useful when there is enough information for gene tree estimation. The ploidy level must be known *a priori*.

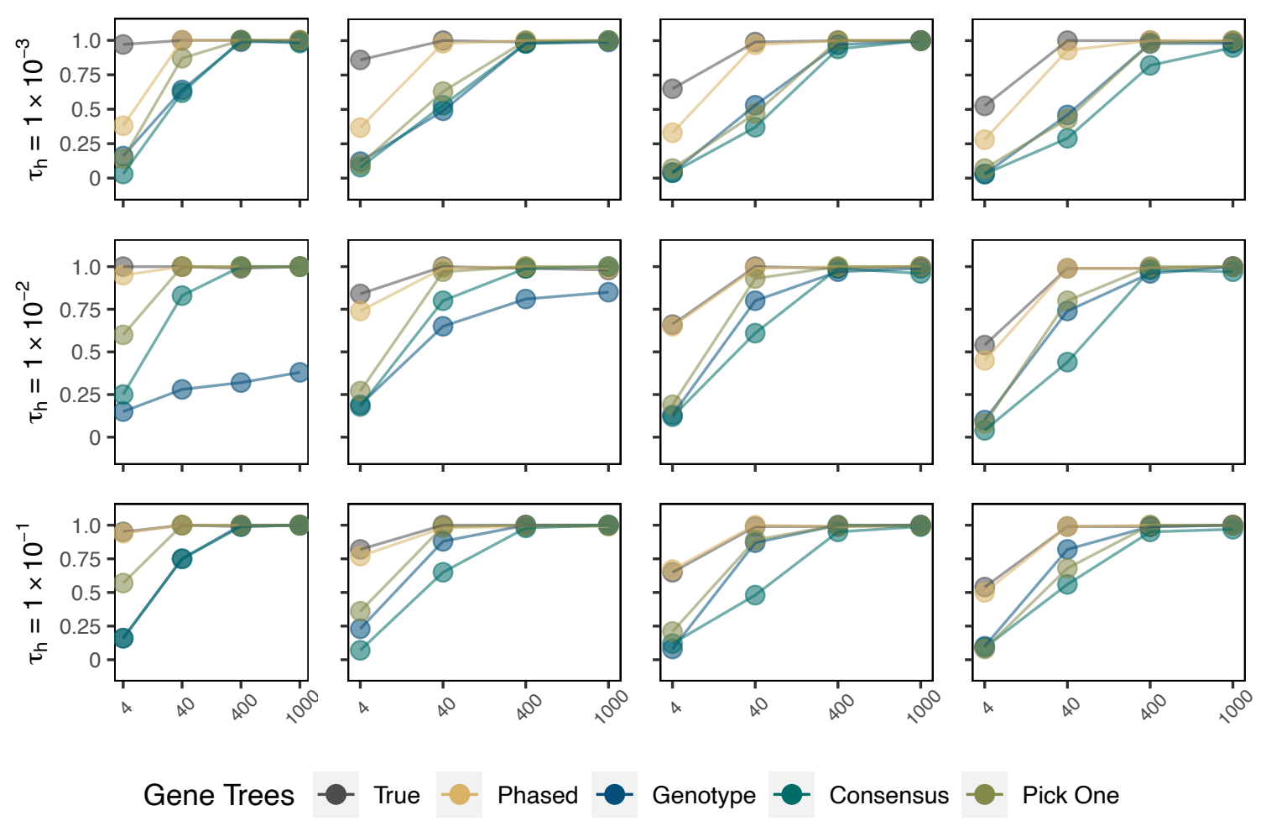
Populations– appropriate when a reference is available and takes advantage of joint genotyping. Results can be used for population genetic analyses or post-processing tools can extract SNP sets. It is possible to estimate ploidy directly from the data, but this requires many sites and a decent outgroup for polarizing alternate alleles.

Benefits of phasing are evident through simulation. Phased data can recover a correct network with less data across a wide range of conditions. Even with the correct network, using the genotype or haplotype consensus sequences can cause bias in divergence time estimates.

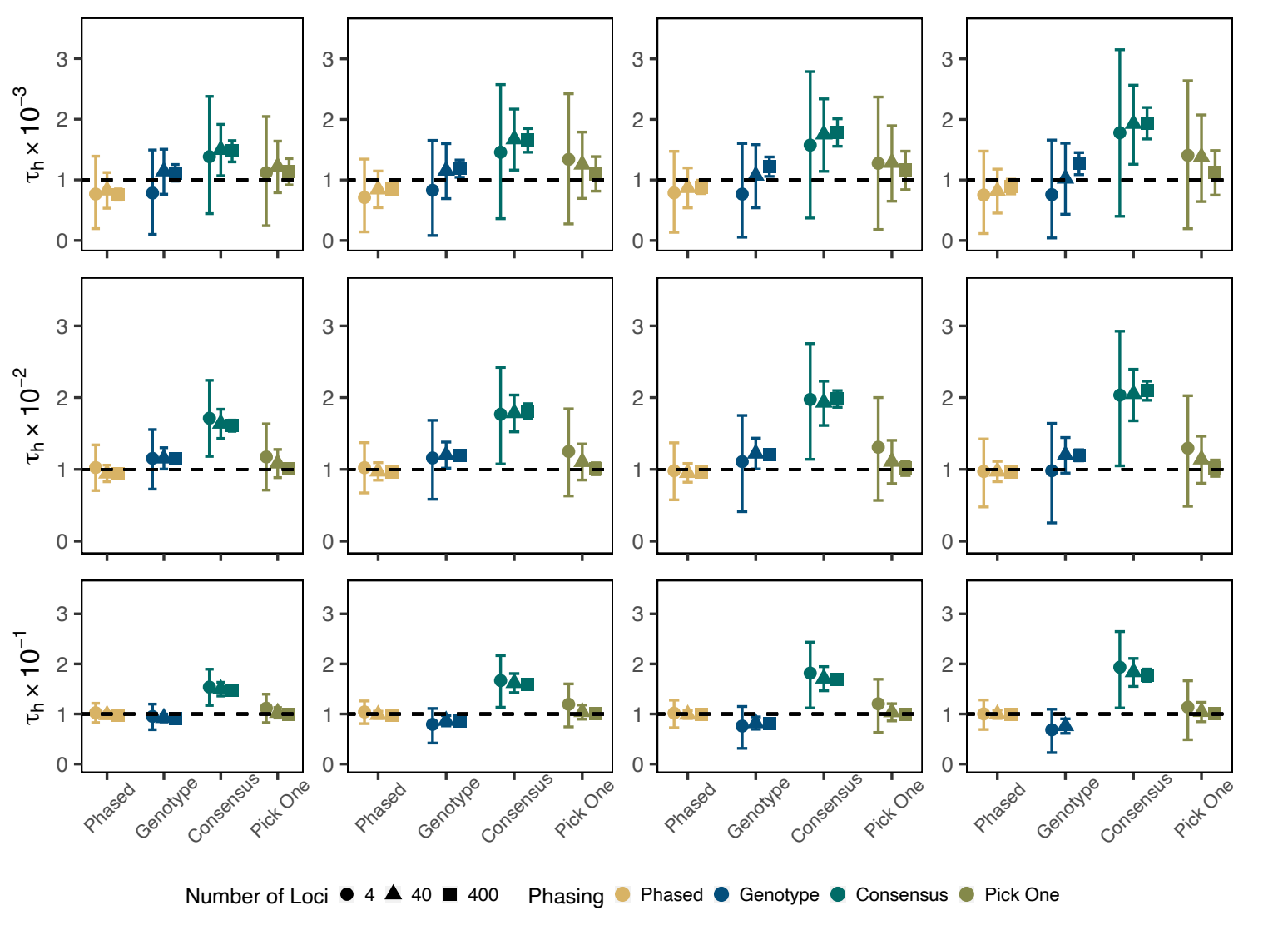
Tree for simulation – We can incorporate different levels of sequence divergence and ILS by changing θ



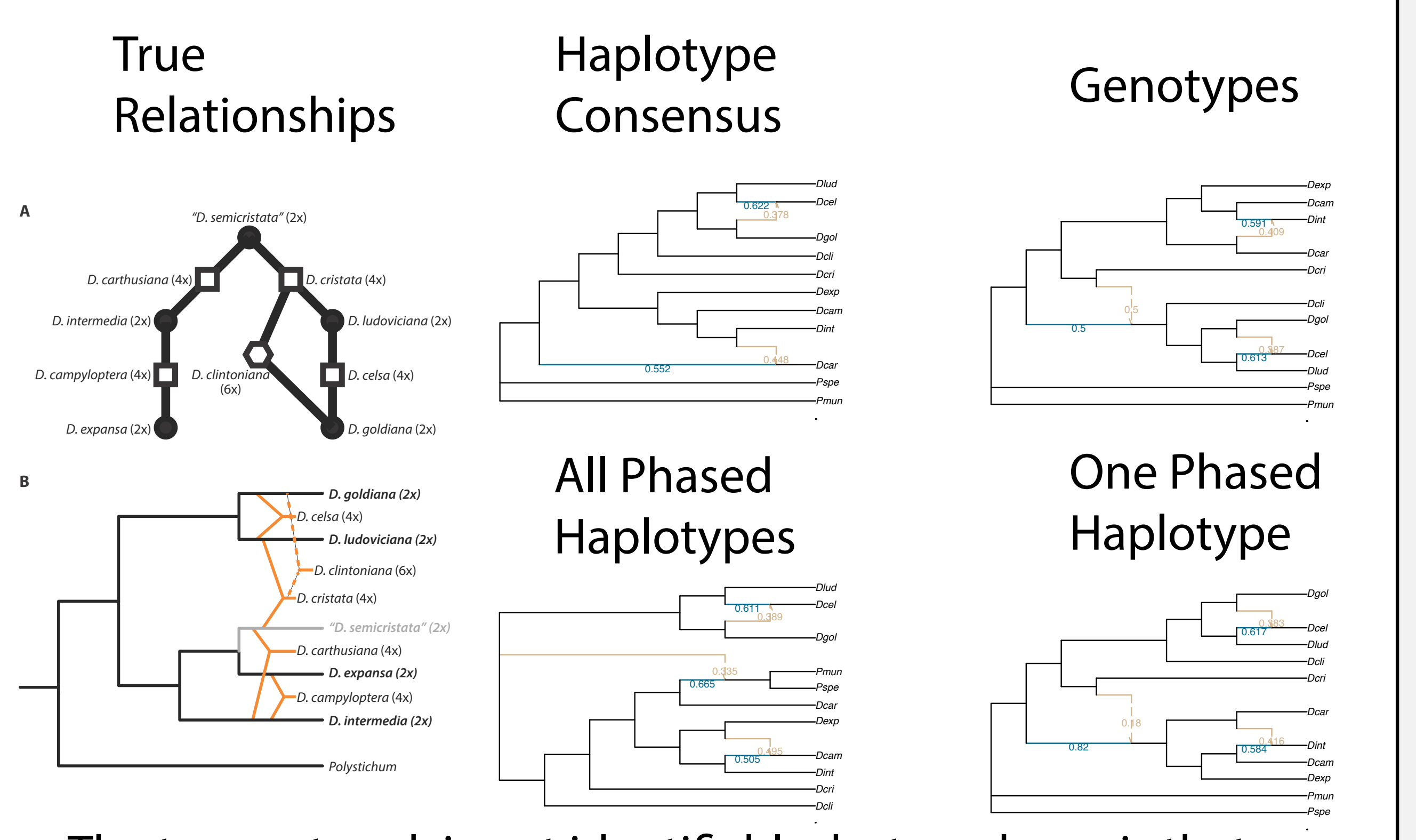
Probability of recovering correct network over 100 replicates with maximum pseudolikelihood



Bias and uncertainty in estimating the time of hybridization using full-likelihood on the correct network



Some evidence that phased data recovers more accurate networks for complex empirical systems.



The true network is not identifiable, but our hope is that methods recover some of the true relationships. Using all phased haplotypes at least gets the two allotetraploids with sampled diploid parents correct.

Phasing target enrichment data can provide additional information for studies of resticulate evolution in non-model groups. Strategies should be applicable to other data types (e.g. RADseq) too, but the value largely remains unexplored. SFS estimation is planned soon.

Try PATÉ! https://gtiley.github.com/Phasing

Preprint! https://doi.org/10.1101/2021.05.04.442457

Methods - Gene trees are estimated with IQ-TREE[1]. All network estimation was done with the SNaQ[2] function as part of the PhyloNetworks[3] Julia package. Full-likelihood estimation of multispecies network coalescent model parameters was done wth BPP[4]. PATÉ[5] is used to prepare data for these analyses. Genotyping is done with GATK[6] and phasing uses H-PoPG[7]. Empirical data was generated with the GoFlag flagellate plant probe set[8]. **Ask me why I like SNaQ and take a sticker.**

References - [1]Nguyen et al. 2015. Mol. Biol. Evol. 32:268-274; [2]Solís-Lemus and Ané 2016. PLoS Genet. 12:e1005896; [3]Solís-Lemus et al. 2017. Mol. Biol. Evol. 34:3292-3298; [4]Flouri et al. 2020. Mol. Biol. Evol. 37:1211-1223; [5]Tiley et al. 2021. bioRxiv https://doi.org/10.1101/2021.05.04.442457; [6]McKenna et al. 2010. Genome Res. 20:1297-1303; [7]Xie et al. 2016. Bioinformatics 32:3735-3744. [8]Breinholt et al. 2021. Appl. Plant Sci. 9:e11406.

Acknowledgements - This project has received funding from the European Union's Horizon 2020research and innovation programme under the Marie Skłodowska-Curie grant agreement No MSCA-IF-EF-ST 101026923. This funding supports GPT and the development of population genetic methods and applications for polyploids.