Allopolyploid Phylogenetics

Annotated Bibliography

Ixchel S. González-Ramírez, Keir M. Wefferling, and Shirley Zhang

Principal papers

1) **Oxelman, B., Brysting, A. K., Jones, G. R., Marcussen, T., Oberprieler, C., & Pfeil, B. E. (2017). Phylogenetics of allopolyploids.** ***Annual Review of Ecology, Evolution, and Systematics*****, *48*, 543–557.**

This review paper outlines some of the challenges in inferring phylogenies of allopolyploids, including some of the technical challenges of obtaining sequence data, as well as the difficulty in phasing alleles and homoeologues (i.e., loci of different subgenomes harbored within an allopolyploid genome). The authors define some terms (e.g., auto- vs. allopolyploidy) and outline the assumptions inherent to most of the below approaches (e.g., non-recombination between parental subgenomes; sampling of all homoeologous copies of nuclear genes; populations with no selection; random mating; non-overlapping generations; no migration after divergence). A brief history of polyploid phylogenetics is provided, setting the stage for new approaches. Finally, the current state of the problem is summarized, with an outline of the limitations and caveats of these different methods.

Oxelman et al. (2017) outline several approaches to inferring a species network (i.e., a tree showing both divergences and reticulations) using molecular sequence data. These approaches fall into several categories:

1. ***Ad hoc:*** e.g., **multi-labeled trees in MUL-trees** (Thomas et al. 2017), **translated onto species tree using PADRE** (Lott et al. 2009)—a maximum parsimony approach that minimizes the number of hybridizations to reconcile the multi-labeled tree with a network. Oxelman et al. (2017) describe several allopolyploid systems, *Galeopsis, Cerastium, Silene, Viola,* and *Fumaria*. The work on *Galeopsis* was part of some pioneering work on synthetic allotetraploids via a triploid "bridge". The studied *Cerastium* belong to a complex high-ploid system with allooctoploids (8*x*) and allododecaploids (12*x*) and no known closely related diploids. A worthwhile quote: "*The fact that gene loss, pseudogenization, and possible lineage sorting are working independently in different parts of the polyploid genome may hamper the interpretation of reticulate evolution even in relatively young plant groups such as the C. alpinum complex and emphasizes the importance of approaches in which several unlinked regions of the genome are examined.*"

2. ***Explicit species network inference:*** Two approaches:
    a. ***Permutation approach using PhyloNet*** (Wen et al. 2008; Wen et al. 2016). Note that this was not built specifically for polyploids (Blischak et al. 2018) but can accommodate more than two alleles/homoeologues per locus. Here, Oxelman et al. (2017) describe a "minimizing deep coalescence" (MDC) approach to species network inference that was implemented by Oberprieler et al. (2017; notably, Oberprieler is one of the co-authors on this review), first inferring a multi-labeled tree, then reconstructing the species network using the PADRE (Huber et al. 2006; Lott et al. 2009) or Dendroscope (Huson & Scornavacca 2012) programs. "*...the proposed workflow pipeline leads to reliable reconstructions if effective populations sizes are not large and divergences among ancestral taxa are not shallow.*"
    b. ***Simultaneous gene tree and species network inference using AlloppNET*** (Jones et al. 2013; Jones 2017). More on this below (Jones papers)! Briefly, here: this is a model-based approach that take the multi-species coalescent into account (fundamentally important, as incomplete lineage sorting, if undetected, could easily lead to identical gene tree topologies as those resulting from hybridization), sets no limit on the number of samples or loci (though it would get clunky pretty quickly with a "large" phylogenomic dataset; C. Rothfels, pers. commun.), estimates the number

of hybridizations (thereby changing the number of parameters necessitating reversible-jump MCMC), and estimates the homoeologue assignment to subgenomes—"*The identity of the homoeologues (i.e., which sequence came from which diploid species) is estimated along with the other parameters.*". Oxelman et al. (2017) provide a nice overview of the model and priors (Graham Jones is a co-author on this review). Oxelman et al. (2017) describe work done on allotetraploid and -hexaploid *Medicago* by Eriksson and others.

2) **Wen, D., Yu, Y., & Nakhleh, L. (2016). Bayesian inference of reticulate phylogenies under the multispecies network coalescent. *PLoS genetics*, *12*(5), e1006006.**

This paper provides the first (?) Bayesian approach to inferring phylogenetic networks ("a rooted, directed, acyclic graph whose leaves are labeled uniquely by a set of taxa"), allowing for nodes with two parents.

Seems like a reasonable approach, though the input is (somewhat unsatisfyingly) gene trees without branch lengths; Keir would prefer multilocus sequence alignments as input. The output is rooted networks. [Note that an approach using sequence alignments (Wen & Nakhleh, 2018) has been developed, and was employed on haploid, randomly phased sequences from diploid organisms].

Wen et al. (2016) employ a rjMCMC approach to the multispecies network coalescent (MSNC) to simultaneously account for incomplete lineage sorting (ILS) and reticulation (earlier approaches were inferred under maximum parsimony, not modeling ILS).

See Fig. 1B; "*A large divergence time between C and the MRCA of A and B or a small population size of the MRCA of A and B would be unlikely to give rise to the indicated* ***gene genealogy****. However, these same settings coupled with a* ***scenario of hybridization*** *between B and C could very well give rise to the same gene genealogy.*"

The main advantage of this approach seems to be the ability to infer more than a single reticulation in a given tree (i.e., nested reticulations).

3) **Jones, G. R. (2017). Bayesian phylogenetic analysis for diploid and allotetraploid species networks. *BioRxiv*, 129361.**

This builds on the work of Jones et al. (2013), now allowing for multiple hybridization events (earlier version of AlloppNET was restricted to a single hybridization event), and estimates the number of events in a reversible-jump MCMC. However, it still has serious restrictions–shared with all other approaches, as far as I can tell–such as the assumption of zero recombination between subgenomes (or at least between the loci under study). It does allow for missing data, which is great; also, homoeologue assignment (to subgenomes) is estimated,

Implementation seems potentially clunky; there is no BEAUti interface for preparing the required .xml file for input to BEAST. Graham Jones does provide some R scripts for preparing the .xml file, but any node dating or other required changes need to be done by hand http://indriid.com/workingnotes2013.html. Also, it is unclear whether AlloppNET can be implemented in BEAST 2 (the version most of us now use).

"*The main restriction here is that only diploid and allotetraploid species are considered. Thus we assume that the species being analyzed have undergone at most one round of allopolyploidization since the root of the species network. We also assume that within the allotetraploids, there is no recombination between sequences from different parental species. This means that all the sequences can be seen as the result of the evolution of diploid genomes, but after hybridization, the topology, node times, and population sizes are shared.*"

4) **Rothfels, C. J., Pryer, K. M., & Li, F. W. (2017). Next-generation polyploid phylogenetics: Rapid resolution of hybrid polyploid complexes using PacBio single-molecule sequencing. *New Phytologist*, *213*(1), 413-429.**

The paper described a combined wet lab and bioinformatics (PURC) approach to generate nuclear sequence data from polyploid accessions. The authors showed that allele inference by PURC is highly repeatable across three sequencing runs and across six analysis regimes, and the number of inferred alleles was strongly indicative of the

ploidy levels of the accessions. The authors applied this approach to obtain all sequences of four c. 1-kb-long nuclear loci from 28 species in the fern family Cystopteridaceae and inferred a multilabeled species tree using AlloppNet (Jones et al. 2013). The inferred reticulate phylogeny was generally consistent with earlier gene trees of the family. These results provide 1) details to Cystopteridaceae phylogeny, such as inferred extinct diploid progenitors and deep hybridization events, and 2) empirical evidence to general hypotheses about polyploid evolution at large, such as the commonness/rarity of allopolyploidy events in generating polyploids and the 'dead-end' model of polyploid evolution.

Extra readings

Morales-Briones, D. F., Liston, A., & Tank, D. C. (2018). Phylogenomic analyses reveal a deep history of hybridization and polyploidy in the Neotropical genus Lachemilla (Rosaceae). ***New Phytologist*****, 218(4), 1668-1684.**

The authors used PhyloNet (v. 3.6.1) to infer phylogeny within *Lachemilla* (Rosaceae) with sequences of 396 nuclear loci and nearly complete plastome from 27 species. Their analyses revealed extensive discordance among individual gene trees and species trees. Additionally, monophyly of four major groups was consistently recovered, but the relationships between these four clades were unresolved, with the Orbiculate group falling sister to any of the other three groups depending on the dataset or analysis pipeline used. The level of gene tree discordance remained high even after removing known hybrid taxa and the Orbiculate clade altogether. Results from gene genealogy interrogation suggested that the observed conflict between different gene tree and species tree estimation is likely not a product of error in their phylogenetic analyses but the presence of ancestral gene flow. PhyloNet results suggested that ancient and recent reticulation events throughout *Lachemilla* are the most likely explanations for the observed discordance. In particular, the Orbiculate clade is likely of ancient hybrid origin between the Verticillate group and the Pinnate group, and is likely to be the main source of incongruence among species trees.

Jones, G., Sagitov, S., & Oxelman, B. (2013). Statistical inference of allopolyploid species networks in the presence of incomplete lineage sorting. Systematic biology, 62(3), 467-478.

In this paper, Jones and collaborators introduce two models to estimate trees in the presence of a hybridization event: AlloppMUL and AlloppNET. These models are innovative since they allow to infer simultaneously the topology, the nodes, the DNA substitution parameters and the hybridization event but they are restrictive in their assumptions: 1) Diploid parents and one single hybridization event that produces a tetraploid. 2) There is no recombination in the tetraploid between the two parental subgenomes. AlloppMUL infers the polyploidy event as a terminal appearing twice in the phylogeny (resulting in a multi-labeled tree representation) whereas AlloppNET models the hybridization event explicitly as a node in a species network. These models are implemented in BEAST and are based on the multispecies coalescent model with some modifications. The models are tested with simulated data and then used to analyze empirical examples in Brassicaceae and Caryophyllaceae.

Wen, D., & Nakhleh, L. (2017). Coestimating reticulate phylogenies and gene trees from multilocus sequence data. ***Systematic Biology*****,** ***67*****(3), 439-457.**

Uses multilocus sequence data! Apparently a more robust approach than that using gene trees only, as in Wen & Nakhleh (2016).

Blischak, P. D., Mabry, M. E., Conant, G. C., & Pires, J. C. (2018). Integrating networks, phylogenomics, and population genomics for the Study of Polyploidy. ***Annual Review of Ecology, Evolution, and Systematics*****,** ***49*****, 253-278.**

Worthwhile reading for a good background on polyploids; especially read section 4.2 for phylogenetic networks. Much of the paper is beyond the scope of today's "Allopolyploid phylogenetics" topic, but provides a very nice background to the importance of polyploidy and some of the realms of study, questions, hypotheses, etc.

Wen, D., Yu, Y., Zhu, J., & Nakhleh, L. (2018). Inferring phylogenetic networks using PhyloNet. *Systematic Biology*, 67(4), 735-740.
This paper provides a summary of the utility of the software package PhyloNet 3. It outlines some of the commands and approaches, and is a good resource for weeding through the PhyloNet literature and pointing out the maximum likelihood and Bayesian approaches; there are several maximum parsimony approaches (!) as well that have been widely used (e.g., minimizing deep coalescence), and these are identified in this review paper.

Kamneva, O. K., Syring, J., Liston, A., & Rosenberg, N. A. (2017). Evaluating allopolyploid origins in strawberries (*Fragaria*) using haplotypes generated from target capture sequencing. *BMC evolutionary biology*, *17*(1), 180.
The authors proposed an approach to infer reticulate phylogeny using the following step wise approach : 1) consensus sequence assembly and haplotype phasing of sequence alignments; 2) ML gene tree inference for every loci; 3) species tree inference in ASTRAL using the gene trees; 4) cluster-network analysis to generate species cluster with support value above 15%; and 5) putative hybridization inferences using the cluster networks. This strategy "avoids the computational burden of inferring species networks by de-novo likelihood construction". As a case study, they inferred reticulate evolution of 20 *Fragaria* species 257 low-copy nuclear markers. Of the 15 putative hybridizations they inferred, 14 were corroborated using PhyloNet analysis but only 7 by STEMhy. The authors argued that this discrepancy was probably due to the difference in the two methods' response to gene tree errors, and calls for researchers to place confidence on candidate hybridization events supported by more than one inference methods.