Commentary

Annals of Botany **123:** iv–v, 2019
Available online at www.academic.oup.com/aob
doi: 10.1093/aob/mcz055

Whole genomes: the holy grail. A commentary on: 'Molecular phylogenomics of the tribe Shoreeae (Dipterocarpaceae) using whole plastidgenomes'

Richard G. Olmstead and Ana M. Bedoya
Department of Biology, University of Washington, Seattle, WA 98195, USA

Corresponding author details: Richard G. Olmstead, olmstead@uw.edu

Heckenhauer *et al.* (2019) address a problem that is as old as phylogenetic systematics: why is it that different datasets sometimes infer different relationships? What's new these days is that we have harnessed the data from whole genomes. However, the question still remains.

In the case of tribe Shoreeae (Dipterocarpaceae), four genera together form a clade, but the relationships among those genera are in conflict. Increasing amounts of evidence in a series of studies (reviewed in Heckenhauer *et al.*, 2018) confirm that two of these genera (*Shorea* and *Rubroshorea*) together are monophyletic, but the relationship of this clade to the two others (*Parashorea* and *Richetia*) conflicts in trees based on plastid and nuclear genome data.

For as long as we have been constructing trees of evolutionary relationships using DNA sequences in plants, authors have been able to pass off responsibility for resolving conflict to future generations, or at least to future studies, by suggesting that with more taxa and more sequence data, the relationships will become clear. What happens when we meet the end of the road? In this study, excellent representative sampling is assessed for complete plastid genome sequences AND reduced representation data of the nuclear genome derived from RADseq to the tune of nearly 20 000 loci and over 100 000 SNPs. Despite the vast amount of data, the conflict between chloroplast and nuclear phylogenetic trees remains. The authors are left to speculate why.

This is a good case study of a problem that is common enough that anyone in this field encounters it sooner or later, and that occurs in many groups of organisms. Specifically, the situation in which extant members of a group form three clades resulting from divergence early in the evolutionary history of the group, and the crown of each clade is subtended by a relatively long branch going back to the early divergence events (Fig. 1).

The authors offer two commonly attributed causes for this conflict, (1) hybridization and subsequent introgression leading to conflict among loci, or (2) incomplete lineage sorting in which individual loci may have evolutionary histories with different branching patterns as a result of retained polymorphisms that sort out differentially sometime after the branching event. The short internode and relatively long stem branches leading to each clade suggest that recent hybridization is not involved. This is further supported by their SplitsTree analysis. However, hybridization is generally thought of as a phenomenon restricted primarily to recently diverged species, and, other things being equal (e.g., population size, outcrossing rate) incomplete lineage sorting is more likely to occur when branches in a phylogeny occur in close temporal proximity. This makes distinguishing the two difficult, even with large amounts of data. In a similar study, Lee-Yaw *et al.* (2019) use whole plastid genomes and RADseq data to test introgression vs. incomplete lineage sorting among annual species of *Helianthus*, an extraordinarily well-studied group, where divergence times are estimated to be no more than 50 000 to 250 000 years. In what may be an ideal situation, they were able to tease apart evidence for multiple introgression events from lineage sorting. In the Dipterocarpaceae, with the estimated age of 28My (+/- 9My) for the critical nodes (Heckenhauer *et al.*, 2017), it is unlikely that even the most sophisticated analyses currently available will be able to determine what process may have been at play.

All of the above assumes that the trees are correctly reconstructed by the large amounts of data obtained. A third possible explanation for the conflict is also possible. With the increased sophistication of probabilistic, model-based approaches to tree construction, we often forget that the results we obtain are only as good as the computer programs we use are at modeling the data. We learned long ago (Felsenstein, 1973) that even subtle biases in model specification can introduce systematic error leading to incorrect results that increase in confidence with increasing amounts of data. Differences in base composition, substitution bias, and patterns of inheritance between nuclear and plastid genomes are difficult to model accurately and could conceivably be amplified by the large amounts of data obtained by next generation sequencing to make conflict seem more significant than it is. This could result in two large datasets with exactly the same branching history producing two different results.

Switching gears, reading this paper gave the first author pause to reflect on the arc of molecular systematics research over the course of the 30+ years he has been participating in it. The earliest

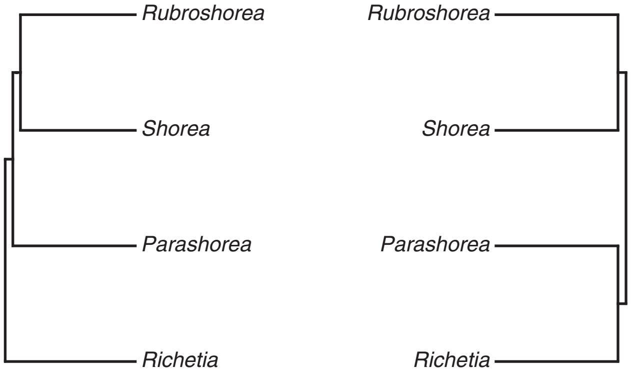

FIG. 1. Conflict in the phylogenetic reconstruction of whole chloroplast genome (left) and RADSeq data (right) reported by Heckenhauer *et al.* (2019). Branch lengths are exaggerated to illustrate the scenario where short internodes may result in incomplete lineage sorting or hybridization causing tree discordance among data sets.

efforts to use DNA for phylogenetic studies in plants involved the use of restriction enzymes to cut DNA into fragments that could be assessed by gel electrophoresis. This proved to be problematic for the nuclear genome, because it was so large, existed in two complements, and had a lot of noncoding regions that were highly variable in size, making fragment analysis difficult. Also, mutations that affected the restriction enzyme recognition sites altered fragment sizes. But for the smaller chloroplast genome, this was very effective (Palmer and Zamir, 1982), and when fragments were mapped, even mutated restriction sites could be interpreted correctly. Largely for this reason, use of the nuclear genome lagged far behind the plastid for systematic purposes.

With the advent of cloning, DNA sequencing, and ultimately PCR and direct sequencing of PCR products, comparison of sequences themselves took over. Emphasis remained with the plastid genome, for similar reasons—scale and accessibility. Early studies used only *rbcL* (Chase *et al.*, 1993), but eventually, individual loci from the nuclear genome were developed for DNA sequencing, starting with the coding and spacer (ITS) regions of the ribosomal repeat (Baldwin *et al.*, 1995), which behave much like plastid DNA, because of their high copy number and rapid copy correction. Soon, multilocus plastid datasets were common and, with effort, multilocus nuclear datasets were developed to complement them (Yuan *et al.*, 2010). An exponential increase in the knowledge of plant genomes has led to an increasing number of gene regions for systematic use.

The advent of next generation sequencing offers the possibility of accessing huge portions of the plant genomes for phylogenetic reconstruction and we have now come full circle. Obtaining whole plastid genomes comes as a byproduct of whole genome sequencing (their presence in high copy makes them easy to capture). No longer is advanced knowledge of the genome necessary (although targeted sequence capture methods make effective use of prior knowledge), and restriction enzyme technology once again is ascendant; now used effectively for sequencing the nuclear genome. RADseq generates thousands of restriction fragment sequences randomly positioned in the genome.

Where do we stand today? Whole genome sequencing, the ‘holy grail’ of molecular phylogenetics, provides us with more and powerful insights into plant phylogeny, but still leaves us pondering some of the same thorny questions we’ve struggled with for decades.

Literature cited

Baldwin BG, Sanderson MJ, Porter JM, *et al.* 1995. The ITS region of nuclear ribosomal DNA: a valuable source of evidence on angiosperm phylogeny. *Annals of the Missouri Botanical Garden* **82**: 247–277.

Chase MW, Soltis DE, Olmstead RG, *et al.* 1993. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbcL*. *Annals of the Missouri Botanical Garden* **80**: 528–580.

Felsenstein J. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Zoology* **22**: 240–249.

Heckenhauer J, Samuel R, Ashton PS, *et al.* 2017. Phylogenetic analyses of plastid DNA suggest a different interpretation of morphological evolution than those used as the basis for previous classifications of Dipterocarpaceae (Malvales). *Botanical Journal of the Linnean Society* **185**: 1–26.

Heckenhauer J, Samuel R, Ashton PS, *et al.* 2018. Phylogenomics resolves evolutionary relationships and provides insights into floral evolution in the tribe Shoreeae (Dipterocarpaceae). *Molecular Phylogenetics and Evolution* **127**: 1–13.

Heckenhauer J, Paun O, Chase MW, *et al.* 2019. Molecular phylogenomics of the tribe Shoreeae (Dipterocarpaceae) using whole plastid genomes. *Annals of Botany* **123**: 857–865.

Lee-Yaw JA, Grassa CJ, Joly S, *et al.* 2019. An evaluation of alternative explanations for widespread cytonuclear discordance in annual sunflowers *(Helianthus)*. *New Phytologist* **221**: 515–526.

Palmer JD, Zamir D. 1982. Chloroplast DNA evolution and phylogenetic relationships in Lycopersicon. *Proceedings of the National Academy USA* **79**: 5006–5010.

Yuan Y-W, Liu C, Marx HE, *et al.* 2010. An empirical demonstration of using PPR (pentatricopeptide repeat) genes as phylogenetic tools: phylogeny of Verbenaceae and the *Verbena* complex. *Molecular Phylogenetics and Evolution* **54**: 23–35.