Applications in Plant Sciences

REVIEW ARTICLE

INVITED SPECIAL ARTICLE
For the Special Issue: Methods for Exploring the Plant Tree of Life

Practical considerations for plant phylogenomics

Michael R. McKain[1,6,*], Matthew G. Johnson[2,*], Simon Uribe-Convers[3,*], Deren Eaton[4,*], and Ya Yang[5,6,*]

Manuscript received 22 January 2018; revision accepted 13 March 2018.

[1] Department of Biological Sciences, The University of Alabama, Box 870344, Tuscaloosa, Alabama 35487, USA

[2] Department of Biological Sciences, Texas Tech University, 2901 Main Street, Box 43131, Lubbock, Texas 79409, USA

[3] Department of Ecology and Evolutionary Biology, University of Michigan, 830 North University, Ann Arbor, Michigan 48109, USA

[4] Department of Ecology, Evolution, and Environmental Biology, Columbia University, 1200 Amsterdam Avenue, New York, New York 10027, USA

[5] Department of Plant and Microbial Biology, University of Minnesota–Twin Cities, 1445 Gortner Avenue, St. Paul, Minnesota 55108, USA

[6] Authors for correspondence: mrmckain@ua.edu, yangya@umn.edu

*All authors contributed equally.

Citation: McKain, M. R., M. G. Johnson, S. Uribe-Convers, D. Eaton, and Y. Yang. 2018. Practical considerations for plant phylogenomics. *Applications in Plant Sciences* 6(3): e1038.

doi:10.1002/aps3.1038

The past decade has seen a major breakthrough in our ability to easily and inexpensively sequence genome-scale data from diverse lineages. The development of high-throughput sequencing and long-read technologies has ushered in the era of phylogenomics, where hundreds to thousands of nuclear genes and whole organellar genomes are routinely used to reconstruct evolutionary relationships. As a result, understanding which options are best suited for a particular set of questions can be difficult, especially for those just starting in the field. Here, we review the most recent advances in plant phylogenomic methods and make recommendations for project-dependent best practices and considerations. We focus on the costs and benefits of different approaches in regard to the information they provide researchers and the questions they can address. We also highlight unique challenges and opportunities in plant systems, such as polyploidy, reticulate evolution, and the use of herbarium materials, identifying optimal methodologies for each. Finally, we draw attention to lingering challenges in the field of plant phylogenomics, such as reusability of data sets, and look at some up-and-coming technologies that may help propel the field even further.

KEY WORDS genome skimming; microfluidics; phylogenomics; RAD-seq; sequence capture; transcriptomes.

Phylogenomics, or using genome-scale sequence data for phylogenetic analyses, has seen major advancements in recent years. Because of the rapid improvement of high-throughput sequencing (HTS) platforms, reduced representation strategies, and analytical tools, obtaining hundreds to thousands of loci has become routine for many botanical researchers. As of early 2018, Illumina HiSeq and MiSeq short-read technologies (Illumina Inc., San Diego, California, USA) are the workhorses of phylogenomics. Emerging long-read technologies such as Pacific Biosciences (Menlo Park, California, USA; PacBio hereafter) and Oxford Nanopore Technologies (Oxford, United Kingdom; Nanopore hereafter) are facilitating acquisition of long loci as well as improved assembly of whole genomes. A number of analytical approaches have been developed to detect polyploidy and dissect heterogeneity within phylogenetic data sets (Li et al., 2015; Smith et al., 2015; McKain., 2016b; Gompert and Mock, 2017; Gregg et al., 2017), making it easier to address polyploidy and reticulate evolution using genome-scale data. Additionally, the community is pushing for full open access to both data and computer code, making it timely to discuss the tradeoffs each strategy has in terms of resolving complicated evolutionary history and reusability of data. With rapidly evolving new tools and the caveats that they bring, choosing an optimal strategy that takes into consideration cost, available plant tissue, and short- and long-term research goals can be a daunting task, especially for people who are new to the field of phylogenomics.

In this review article, we focus on the costs and benefits of different approaches used in phylogenomics including: microfluidic PCR, restriction enzyme–based methods, genome skimming, target enrichment, and transcriptomics. We highlight unique challenges and opportunities in plant systems—such as polyploidy,

which requires distinguishing homeologous gene copies; reticulate evolution, which requires biparentally inherited loci; and the use of herbarium materials with short and partially degraded DNA—making practical suggestions for each. Finally, we draw attention to challenges such as reusability of data sets and discuss some up-and-coming technologies that may help propel the field even further. Table 1 provides a side-by-side comparison of the methods explored in this manuscript.

TABLE 1. Comparison of cost and utility of phylogenomic methods for plants.[a]

Aspect of method	Sanger-based methods	Microfluidic PCR	Restriction enzyme–based methods	Genome skimming	Target enrichment	Transcriptome
Upfront investment	Design and optimizing PCR primers. Timeframe: weeks; cost: $50–100	Some genomic data (e.g., shotgun libraries). Timeframe: weeks to months; cost: $100–1000	Optional: test alternative restriction digestions for optimal range of fragment sizes. Timeframe: weeks; cost: $100–1000	None	Transcriptome and/or genomes from closely related organisms. Timeframe: months; cost: $100–1000	Freezers and liquid nitrogen containers; logistics for tissue collecting. Timeframe: weeks to months; cost: $1000–10,000
Tissue for sampling: herbarium, silica preserved, flash-frozen, living	All four types but reduced success from low-yield herbarium tissue extractions	All four types, but reduced success from low-yield herbarium tissue extractions	All four types, but reduced success from low-yield herbarium tissue extractions	All four types, but potential reduced success from low-yield herbarium tissue extractions. See Saeidi et al., 2018.	All four types, but reduced success from low-yield herbarium tissue extractions	Flash-frozen or living tissue preserved in RNA*later*
Sequence information type	Coding region, short introns, and short intergenic spacers	Coding region, introns, and short intergenic spacers	Anonymous or reference-mapped short-length loci	Organellar, some nuclear	Coding region and flanking intron	Coding region
Cost per extraction + library prep + sequencing (varies depending on platform)	$1–5 for standard DNA extraction + $0 + $3 for a single read of 800–1000 bp	$1–5 for standard DNA extraction + $0 + $0.40 per microfluidic reaction (~$800 per 48 × 48 plate)	$1–5 for standard DNA extraction + $5–50 + $1200–1800 (HiSeq sequencing 48–384 samples)	$1–5 for standard DNA extraction + $25–150 + $50 (assume 1.5–2 Gb of data per sample)	For 96 samples: $200 for probes, $1–5 for standard DNA extraction + $16 (library in 1/3 volumes) + $1800 for sequencing (MiSeq 2 × 300)	$5–15 for RNA extraction + $50–150 + $200 (assume 25 million reads per transcriptome)
Assembly and cleaning data	Easy	Easy	Relatively easy	Moderately computationally intensive	Moderately computationally intensive	Computationally intensive
Ability to resolve reticulation/ hybridization/ introgression	Yes	Yes, potential to recover single alleles from nuclear loci	Yes, significant power to test for genome-wide or localize admixture	Sometimes, potentially identify hybridization but only if it is biparentally inherited	Yes, can extract alleles if long reads are used	Yes
Ability to infer polyploidy	Yes, but needs time-consuming cloning	Yes, potential to recover single alleles from nuclear loci	Sometimes, can detect polyploidy from read depths, but low potential to separate paralogs	No	Sometimes, can detect abundance of paralogous sequences	Sometimes, if polyploidy event is old enough that homeologs can be separated during transcriptome assembly
Best use	First pass; when maximizing the number of samples is the priority	Closely related species; studies that need specific loci and complete data matrices	Shallow phylogenetic scale with aim to sample many individuals	Deep or shallow phylogenetic scale; detecting parental heritage; potential genome diversity	Deep or shallow phylogenetic scale for up to a few samples per species	When detecting genome duplication and gene family evolution are of interest beyond reconstructing species relationship
Reusability of data	Yes, if using the same loci	Yes, fully reusable within focus group using same loci, limitations with increase in phylogenetic distance	Sometimes, reusable for studies within same study system, not reusable between distant clades	Yes, fully reusable	Sometimes, partially reusable across studies if same loci are targeted	Yes, fully reusable

[a]Costs are given in U.S. dollars (US$) as of 2018.

SANGER SEQUENCING

Primer design from available data → DNA extraction, PCR amplification, and sequencing → read consolidation → phylogenetic inference

In many cases a few loci are sufficient

Low-throughput sequencing-based approaches, typically by PCR amplification of nuclear ribosomal ITS (Baldwin et al., 1995; Baldwin, 1998) and/or a handful of chloroplast regions (Shaw et al., 2005, 2007, 2014) followed by Sanger sequencing, are still commonplace in plant systematics. These methods have the advantages of being low cost (despite relatively high cost per base compared to HTS), requiring only standard molecular lab equipment, and having straightforward data analysis. These methods are often used as the first step in molecular systematics training for students and are useful in many cases for phylogenetic reconstruction. In addition, Sanger-based methods can be used for barcoding vegetative or fragmented samples, or as a quick first pass for selecting samples for collecting genome-scale data.

Due to frequent polyploidy and reticulate evolution in plants, amplified nuclear regions often contain co-amplified paralogs and divergent alleles that must be isolated by laborious procedures like cloning. For this reason, PCR amplification of low-copy genes—a common practice in animal phylogenetics—is not particularly efficient. To avoid cloning, high-throughput single-molecule sequencing approaches, like PacBio, have been effective in sequencing and isolating homeologs from amplified PCR products for a small number of loci in polyploid taxa (Rothfels et al., 2017). For many phylogenetic analyses, however, a lack of phylogenetic signal and potentially high conflict found among few nuclear and chloroplast regions necessitates sequencing a larger number of loci, i.e., a phylogenomics approach. Medium-throughput approaches can be employed by performing PCR amplification of multiple loci per taxon and barcoding them for pooling and sequencing on platforms like the Illumina MiSeq (e.g., Cruaud et al., 2017). However, as overall preparation time and costs for more high-throughput methods continue to decrease, HTS becomes a more cost-effective choice.

MICROFLUIDIC PCR

Primer design from available data → DNA extraction, PCR amplification, and sequencing → read/locus alignment → phylogenetic inference

Among the multiple phylogenomic approaches commonly used today, microfluidic PCR is one that has a familiar feel to it. As the name implies, it is based on PCR amplification of targeted regions but has the advantage of producing much larger amounts of data more efficiently and cost-effectively than standard PCR. Microfluidic PCR is based on the same principles as standard PCR: a DNA template, a set of forward and reverse primers used to target and amplify the region of interest, enzymatic and chemical reagents (i.e., *Taq* polymerase and dNTPs), and a series of heating and cooling steps. The main differences between the two are that (1) two pairs of primers (two forward and two reverse) are used in each microfluidic reaction instead of just one, (2) the volumes of DNA template and other reagents are extremely reduced, (3) all primer pairs have the same annealing temperature, and (4) amplified regions should be of similar lengths. Microfluidic technology has become popular in biomedical fields like cancer research (e.g., Gaedcke et al., 2012; Walter et al., 2012), genotyping of single-nucleotide polymorphisms (SNPs) (e.g., Bhat et al., 2012; Byers et al., 2012; Lu et al., 2012), gene expression (e.g., Dominguez et al., 2013; Moignard et al., 2013; Shalek et al., 2013), and targeted resequencing (e.g., Lohr et al., 2012; Moonsamy et al., 2013). Recently, this technology has been used for generating phylogenetic data sets in microbial systems (Hermann-Bank et al., 2013), haplotyping (identification of individual genome copies) of commercially important plants (Curk et al., 2014), and elucidating recent radiations in plants (Gostel et al., 2015; Uribe-Convers et al., 2016).

Capacity of microfluidic PCR

In phylogenomics, the most commonly used equipment for microfluidics is the Fluidigm Access Array System (Fluidigm Corporation, San Francisco, California, USA). This machine functions as a standard thermocycler except that it amplifies 48 target regions across 48 accessions/taxa (2304 amplicons) per microfluidic array, whereas other versions of this equipment may have higher throughput (e.g., 96 × 96). Each reaction is performed with four primers simultaneously: a pair of primers that amplify the region of interest and a second pair of primers that anneal to the first pair adding barcodes and sequencing adapters to the amplicon. Because the final product contains barcodes, sequencing adapters, and the region of interest, this four-primer PCR approach circumvents the need for other library construction methods. After PCR amplification, all amplicons from each sample are combined in a single pool, and all 48 pools of amplicons (one for each sample) are quantified and sequenced on an HTS platform. The current yield of HTS platforms (e.g., Illumina MiSeq) allows for up to four to six microfluidic plates to be sequenced on one lane and still get the necessary sequencing depth for phylogenomics. Lastly, amplification reactions on the Fluidigm Access Array System can be multiplexed. Fluidigm has shown that one could combine up to 10 primer pairs per well—bringing up the number of amplicons to 23,040 per microfluidic plate—if the primers within a well have no interaction (e.g., primer dimers) and if the target regions are located far enough in the genome so that PCR amplifications do not interfere with each other.

One of the main advantages of microfluidic PCR is that minimal quantities of DNA and PCR reagents are used. The 2304 wells available in the Fluidigm Access Array 48 × 48 plate are only 36 nL in total volume (Cronn et al., 2012), and require just 15 units of *Taq* polymerase to amplify the entire plate. Similarly, only 1 μL of DNA template is necessary to generate all amplicons for each sample. Although Fluidigm recommends that high-quality (50 ng/μL) DNA be used in the reactions, there has been success with much lower concentrations (~10 ng/μL), and importantly, with DNA extracted from herbarium specimens (Uribe-Convers et al., 2016; Latvis et al., 2017).

Microfluidic PCR produces consistent data sets across lineages

Data produced by microfluidic PCR are ideal for phylogenomics. Amplified and sequenced regions have been through a rigorous selection process, often targeting single- or low-copy loci that are highly informative, resulting in data that are useful to resolve relationships in both young and old clades. Moreover, only targeted regions are amplified and sequenced, increasing both the sequencing

depth for each amplicon and the number of loci that are shared among samples. This last point is particularly important as it reduces the amount of missing data in alignment matrices—something that has been shown to be beneficial for phylogenetic inference (Lemmon et al., 2009), although not always a necessity for topology reconstruction (Rubin et al., 2012; Mastretta-Yanes et al., 2015).

Another key advantage of microfluidic PCR is that it can be used with polyploid species. As mentioned above, the highly targeted nature of this approach yields consistently deep sequencing depth among all samples, which facilitates recovery and identification of alleles and homeologs within a locus. Sequence data from microfluidic PCR can be demultiplexed twice: first using the sequencing barcodes and then using the PCR primer sites. This double demultiplexing approach results in groups of sequencing reads that belong to a specific sample and locus, reducing complexity for assembly. If the HTS is done with relatively long (300-bp reads on MiSeq) paired-end reads, no read assembly is needed, except for potentially collapsing reads. By simply aligning the reads back to a reference, many of the computational burdens that an assembly entails are greatly minimized. Variant call information for a locus within a sample can be used to identify different alleles. This approach has been demonstrated using the young genus *Neobartsia* Uribe-Convers & Tank (Orobanchaceae) with great success, and scripts are available to process microfluidic data (Uribe-Convers et al., 2016). Additionally, microfluidic PCR has been used to study the evolutionary histories of *Commiphora* Jacq. (Burseraceae; Gostel et al., 2015) and squashes (*Cucurbita* L., Cucurbitaceae; Kates et al., 2017). Finally, once a set of primers are available, they can be reused within the same group and, in some cases, among allied genera within a family (Latvis et al., 2017) and even an order (Collins et al., 2016).

Upfront investment in primer design

The main disadvantage of microfluidic PCR compared to other HTS methods is the time invested in primer design. Compared to standard PCR primers, all microfluidic primer pairs must have an annealing temperature of 60°C (±1°C). Higher annealing temperatures usually require designing longer primer sequences (~27 bp) in conserved genomic regions, which can be difficult to find across taxa. To compound the challenges of primer design, the regions that are targeted for amplification should all be close in length (usually ~600–900 bp). The latter point is important because sequencing depth could be biased toward shorter regions (~350 bp), or if the regions are too long (>1000 bp), they could interfere with each other during bridge amplification on an Illumina platform. Some prior sequence data (e.g., plastomes, shotgun libraries, transcriptomes) from three to five species within a focal clade are required for designing primers in order to maximize amplification success across a clade. It is recommended to validate primer amplification by simulating the microfluidic PCR conditions in a standard thermocycler. A good primer pair should not create primer dimers nor interfere with barcode and sequencing adapter primers. A good primer pair should also only amplify the region of interest (i.e., no double bands on an agarose gel). Finally, microfluidic amplification is done in a specialized piece of equipment (e.g., Fluidigm Access Array System) that might not be available in standard molecular laboratories or genomic cores. However, microfluidic instruments are used in other amplification-based studies (e.g., genotyping or variant calling), and they are becoming more popular and even commonplace in large genomic cores. Although a researcher new to microfluidics might at first feel discouraged by its limitations, this method provides data matrices with very little missing data, allows complete control over what loci get sequenced, and has no assembling steps and minimal data processing, and the resulting data can outweigh the initial difficulties of primer design.

RESTRICTION ENZYME–BASED METHODS

Restriction enzyme selection→ DNA extraction→ DNA digestion → library preparation and sequencing → read mapping/SNP identification → phylogenetic inference

High-throughput restriction-site-associated DNA sequencing describes a suite of related methods, here referred to collectively as RAD-seq, that utilize restriction enzymes to fragment genomic DNA prior to sequencing (Miller et al., 2007; Baird et al., 2008; Davey et al., 2011; Andrews et al., 2016). Illumina adapters are ligated to digested fragments during library preparation such that all sequenced reads begin at restriction cut sites and extend away for the length of a single sequenced read (typically 75–300 bp). The resulting sequenced reads therefore accumulate at each RAD locus to form high-coverage stacks that can be used to confidently call alleles and SNPs—a convenient outcome relative to many other genomic methods that require tiling partially overlapping sequences, often at lower depths, to construct contigs. Due to their relatively short lengths, however, RAD loci are often not highly informative for constructing gene trees, and instead are typically best suited for SNP-based inference methods (discussed further below). Nevertheless, because RAD-seq methods are affordable and easy to implement, they have been widely adopted in population genetics, phylogeography, and phylogenetics (Hohenlohe et al., 2010; Eaton and Ree, 2013; Nadeau et al., 2013; Wagner et al., 2013; Hipp et al., 2014; Herrera et al., 2015; Leaché et al., 2015; McCluskey and Postlethwait, 2015; DaCosta and Sorenson, 2016).

High flexibility and low cost

One of the greatest strengths of RAD-seq is its flexibility. If you wish to sample more genetic markers, you can simply select a more frequently cutting restriction enzyme or a wider window of digested fragment sizes to include in the sequenced library. If instead you want fewer markers sequenced to higher depth, you can choose a less common cutter. In this way, for most large plant genomes, it is possible to target approximately as many RAD loci for inclusion in your data set as you see fit. This has made RAD-seq particularly promising for the study of recent radiations (Nadeau et al., 2013; Wagner et al., 2013), species delimitation (Leaché et al., 2014), and introgression/admixture (Dasmahapatra et al., 2012; Eaton and Ree, 2013), where a broad sampling of sites from thousands of regions across the genome can be used to characterize heterogeneity in the distribution of gene tree patterns and provide statistical power for tests of reticulate evolution (Durand et al., 2011; Reddy et al., 2017).

The flexibility of RAD-seq can sometimes also lead to confusion as there are now a variety of related methods with similar names but distinct differences. These protocols typically vary in the number and type of enzymes that are used, as well as in the equipment that is required for their preparation, all of which can lead to significant differences in their cost as well as in the quality

and quantity of data generated (Elshire et al., 2011; Peterson et al., 2012; Toonen et al., 2013; Andrews et al., 2016; Glenn et al., 2017; see Andrews et al. [2016] for a review). For example, the original RAD protocol uses only a single restriction enzyme to produce a range of DNA fragment sizes that are subsequently subjected to sonication in order to shear DNA fragments to the appropriate length for short-read sequencing. In contrast, dual-digest methods use two restriction enzymes to digest the genome into variably sized fragments. The resulting fragments are selected based on whether their size falls within an appropriate window for short-read sequencing. The former method can provide more data with less bias caused by mutations to restriction sites, while the latter method can be more flexible in tailoring the number of selected fragments and is cheaper because it requires less specialized equipment and adapters.

The relatively low cost of RAD-seq is one of the primary reasons it has attracted significant use in recent years. The initial cost of materials is low. Subsequent library preparations and sequencing typically range from US$15–75 per sample, and the cost continues to decrease (Andrews et al., 2016). Recent advances to indexed primers (Glenn et al., 2016) and library preparations (e.g., 3RAD; Glenn et al., 2017) have further decreased costs to below US$1/sample for large multiplexed data sets (Hoffberg et al., 2016b). A potentially promising approach for large-scale projects, in which many hundreds or even thousands of individuals are to be sequenced, is to use a hybrid approach like RADcap, which combines restriction digestion with targeted bait capture (Hoffberg et al., 2016a). This combined approach to select and sequence fewer RAD loci provides more even sequencing coverage and allows more samples to be multiplexed. Depending on the number of samples in a study and the number of loci to be targeted, selecting an appropriate protocol can be hugely beneficial and cost saving.

Analysis of RAD-seq data

A major strength of RAD-seq is the enormous quantity of data that can be collected. Because the method does not rely on targeting relatively invariant regions of the genome or coding genes, there are often high rates of variation across RAD loci that make it easy to sample thousands or even millions of SNPs. Typical assembly methods for RAD-seq provide both SNP-based sequence formats and full sequence data for traditional phylogenetic analyses (Catchen et al., 2013; Eaton, 2014), with the addition that reference-mapped data provide the genomic location of markers. In this way, the distribution of sampled RAD loci and SNPs across chromosomes can also be used to study phylogenetic relationships or patterns of admixture as they vary across sliding windows of the genome (e.g., Dasmahapatra et al., 2012). Large SNP data sets may prove particularly useful as SNP-based phylogenetic inference methods continue to develop (see review by Leaché and Oaks, 2017). Such methods that do not require inferring fully resolved gene trees for each locus can expand the utility of methods like RAD-seq while also reducing errors in phylogenetic analyses that arise from assuming gene trees are accurate and that recombination is absent within longer sequences of DNA (Bryant et al., 2012; Chifman and Kubatko, 2014). Still, for relatively deeper-scale phylogenetic analyses, RAD-seq loci are often sufficiently variable to be used in gene-tree-based methods that employ the multi-species coalescent (e.g., Ogilvie et al., 2016; Eaton et al., 2017; Vargas et al., 2017).

Missing data require careful filtering

A particularly relevant concern for RAD-seq is missing data, although the implications of missing data apply similarly to any phylogenomic analysis. Because RAD-seq relies on the conservation of restriction recognition sites across samples in order to target homologous markers, disruption of these sites by mutations (mutation-disruption) leads to missing data. The generation of new restriction recognition sites by mutations can also lead to RAD loci being shared by some samples and not others, but simulations suggest that disruption of ancestrally shared restriction sites is of much greater significance (Eaton et al., 2017). More divergent taxa are thus expected to share fewer conserved restriction sites on average, and thus less pairwise phylogenetic information. This has led to considerable debate as to whether RAD-seq can be accurately applied to deeper phylogenetic scales (Rubin et al., 2012; Cariou et al., 2013). Although it is now clear from empirical applications that RAD-seq can provide significant phylogenetic information over even tens of millions of years divergence (Eaton and Ree, 2013; Escudero et al., 2014; Eaton et al., 2015, 2017; McVay et al., 2017; Tripp et al., 2017; Vargas et al., 2017), the more relevant concern is the scale at which missing data make this method no longer economical compared to alternatives.

Although the problem of missing data is inherent to RAD-seq data sets, in most cases many thousands of loci can be recovered across all or nearly all samples in a study. Bioinformatic methods are employed to filter loci from a data set to select those with some minimum proportion of missing data (Eaton, 2014). Depending on how many restriction enzyme cut sites are targeted, the quality of the library, the sequencing coverage, and the number of samples and their phylogenetic relationships, this may constitute a large proportion of the total loci or a very small proportion (sometimes even none). However, because loci with missing information for some taxa can still provide phylogenetic information for many other taxa, most data sets allow for 30–90% missing data in combined multi-locus data sets (Eaton et al., 2017). In general, missing data tend to have little impact on phylogenetic tree topology (Rubin et al., 2012; Mastretta-Yanes et al., 2015) but can significantly affect other aspects of phylogenetic inference such as branch lengths (Ogilvie et al., 2016).

Although the primary source of missing data in RAD-seq studies is typically assumed to be mutation-disruption, many other aspects of library preparation or sequencing can have an equal or greater effect. Consider that one of the drawbacks of RAD-seq is that loci contain very little variation—only one or a few SNPs per locus. This presents a contradiction to the expectation that mutation-disruption causes most missing data: if few mutations occur in a 100-bp locus, then even fewer mutations should occur in the small restriction recognition site adjacent to the locus. Thus, mutation-disruption would be unlikely to cause 50% of sequences to be missing from a data set. Instead, the amount of missing data in RAD-seq will often depend on many other factors.

In a comparison of 10 phylogenetic-scale RAD-seq data sets, Eaton et al. (2017) showed that selecting an appropriate library preparation method and sequencing depth is of great significance for the amount of phylogenetic information that will be obtained. For example, in a relatively deep-scale phylogenetic analysis of the genus *Viburnum* L., they showed that a 2× increase in sequencing coverage led to >10× increase in the number of phylogenetically informative sites recovered. However, the same return on

sequencing coverage was not observed for all data sets, with the primary difference depending on whether the library was single or dual digest. Single-digest libraries tend to generate many fragments that are often under-sequenced, whereas dual-digest libraries usually select many fewer fragments that are more easily sequenced to sufficient depth. Although the number of reads that must be sequenced to attain a sufficient level of coverage per sample can be estimated based on the expected number of loci in a genome, such estimates are often difficult to make, and it is typically easier to simply base estimates on studies already conducted in related organisms. A discussion of differences among RAD-seq protocols is clearly relevant for designing a project, as well as when comparing RAD-seq to other methods. Although RAD-seq methods are easy to implement, careful attention and troubleshooting of library preparations, including the quality of DNA extractions, restriction digestions, and size selection windows, can have an enormous effect on the results (Graham et al., 2015).

Working with duplications and paralogy

Due to the frequency of gene and genome duplications in plants, anonymous phylogenetic markers have historically received relatively little use, and for many researchers, this reticence applies similarly to RAD-seq. However, anonymity is not necessarily a property of RAD-seq per se, but rather a possible outcome depending on how the data are assembled. If a good reference genome is available, RAD loci can be assembled like many other genomic markers by mapping reads to a reference, in which case paralogy is assessed based on whether reads map to multiple locations in the genome. Due to their short length, however, paralogous loci are typically removed from RAD-seq data sets rather than being further analyzed to try to tease apart paralogous gene tree histories. It is only for cases in which taxa lack a good reference genome that RAD loci are assembled de novo, wherein reads are clustered by sequence similarity to identify homology. Paralogy is more difficult to assess in this case and is usually based on distributions of site frequencies and excesses of heterozygosity or alleles (Eaton, 2014).

Empirically, the effect of paralogs on phylogenetic inference is difficult to assess, as there are many possible ways in which paralogs can be distributed. In their recent phylogenetic study of the plant clade *Viburnum*, Eaton et al. (2017) compared a RAD-seq phylogeny to a tree inferred from Sanger sequence data composed of a nuclear locus and chloroplast sequences (ITS + nine cpDNA regions) and found nearly complete concordance, despite the fact that *Viburnum* has several instances of derived polyploidy. From this, they suggested that any paralogs retained in the RAD data set after filtering likely had relatively little effect on the genome-wide phylogenetic signal. The ratio of phylogenetic signal to noise generated by ancient genome duplications versus more recent species divergences will typically determine the extent to which paralogy is likely to obscure phylogenetic inference. It remains for more detailed studies to investigate the impact of paralogy on various phylogenomic data sets analyzed under different methods. Most applications of RAD-seq for polyploids to date have focused on the detection of ploidy based on read-depth information (Gompert and Mock, 2017); however, there remains a lack of phylogenetic methods for further analysis of polyploids using primarily SNP data.

GENOME SKIMMING

DNA extraction, library preparation, and sequencing → organelle genome assembly → annotation → phylogenetic inference

Genome skimming (also called genome survey sequencing or low-coverage genome shotgun sequencing) is the method of sequencing total genomic DNA without any enrichment (Straub et al., 2012). In plants, the resulting data are a representation of the nuclear, chloroplast, and mitochondrial genomes of the target individual, although contaminants from pathogens, the microbiome, and symbionts (e.g., Nakamura et al., 2013) may also be present. Often, genome skim data contain less than 1× coverage of the nuclear genome, making them inadequate for identification of nuclear genes. However, when sequenced at a higher depth (2–3×), Berger et al. (2017) have demonstrated that it is possible to extract low-copy nuclear genes. Higher coverage is needed not only to ensure complete representation of low-copy loci but also to overcome issues of sequencing error. Other fractions of the data, such as the chloroplast genome (McKain et al., 2016a; Qu et al., 2017), mitochondrial genome (Guo et al., 2016; Petersen et al., 2017), nuclear ribosomal genes (Steele et al., 2012), and repetitive elements (e.g., transposable elements), are in much higher copy numbers, are generally represented in higher relative coverage compared to low-copy loci (>30×), and often allow for assembly of organellar genomes and ribosomal genes or characterization of transposon diversity and quantity (Staton and Burke, 2015b). Genome skimming is a relatively easy first step into phylogenomics because projects require commonly used molecular techniques such as DNA isolation and sequencing library production and do not require data generation prior to initiating a project.

Multiple types of specimens are viable

A benefit of genome skimming is that any source of viable double-stranded DNA can be used. Projects using living (Male et al., 2014; Zhang et al., 2015) and herbarium (Bakker et al., 2016; Teisher et al., 2017) specimens have all successfully generated genome skim data (Staats et al., 2013). Genome skimming is able to use DNA that is otherwise too degraded for PCR-based sequencing approaches (Staats et al., 2011). In recent years, the use of herbaria in genome skim projects has exploded, with molecular workflows modified for isolation and shotgun sequencing of herbarium DNA (Särkinen et al., 2012; Bakker et al., 2016; Saeidi et al., 2018) producing whole chloroplast genomes from samples more than 100 years old (Bakker et al., 2016), endangered species (Welch et al., 2016), and extinct species known only from herbarium collections (Zedane et al., 2016). Because these workflows produce viable libraries for HTS, they are also amenable to sequence capture, as discussed below.

Chloroplast genomes are readily and inexpensively obtained

Organellar genomes make up a large component of total genomic DNA, with cpDNA ranging from <0.3% in *Picea abies* (L.) H. Karst. needles to 37% in *Asclepias syriaca* L. leaves (Twyford and Ness, 2016) and mitochondrial DNA abundance 5–10% that of cpDNA (Bock et al., 2014). Although genome size in flowering plants varies from 63.6 Mbp in *Genlisea aurea* A. St.-Hil. (Leushkin et al., 2013) to almost 152.23 Gbp in *Paris japonica* (Franch. & Sav.) Franch. (Pellicer et al., 2010), there does not appear to be a direct negative

correlation between genome size and percent total organellar DNA, suggesting that genome skimming is a viable option for obtaining organellar genomes across many taxa (Bakker et al., 2016; Twyford and Ness, 2016). Because of their relative abundance, (usual) structural simplicity, and historical significance in systematics, chloroplast genomes have become a primary target of genome skimming projects.

When designing projects for chloroplast genome sequencing, two factors should be considered to maximize data over cost. First and foremost, it must be decided if full chloroplast genomes are necessary or if protein-coding gene space will be adequate. Acquiring full chloroplast genomes for each sample can add extra time to data generation and analysis; however, multiple assembly pipelines (e.g., ACRE [Wysocki et al., 2014], IOGA [Bakker et al., 2016], NOVOPlasty [Dierckxsens et al., 2017], Fast-Plast [https://github.com/mrmckain/Fast-Plast], and a *k*-mer-based approach [Izan et al., 2017]) have been developed that are capable of assembling complete chloroplast genomes from short-read data. Complete chloroplast genomes can potentially provide more phylogenetic signal from intergenic regions for reconstructing relationships among closely related species (Carbonell-Caballero et al., 2015). When sequencing of complete chloroplast genomes is not feasible or necessary, read mapping–based approaches can provide adequate assemblies for chloroplast gene space for population-level studies (Vallejo-Marín et al., 2016). De novo assemblies of complete chloroplast genomes often require higher coverage than mapping-based approaches, so this must be taken into account in project design. The second factor to consider is the relative percentage of total genomic DNA that is chloroplast for the taxa being sequenced. An underestimate can result in chloroplast genomes not sequenced deeply enough for adequate data acquisition, and an overestimate can result in wasted potential taxon sampling. In practice, whole chloroplast genomes can be assembled from an estimated 50–100× average coverage, although coverage is not always consistent across the chloroplast genome. Regions rich in either AT or GC repeats often see decreases in coverage (Benjamini and Speed, 2012) meaning some lineages may need a higher average coverage for complete chloroplast genome assembly. Estimating the percentage of total DNA can be done by mapping existing reads from the taxon group to a representative chloroplast genome (e.g., Twyford and Ness, 2016). If such data are not available, quantitative PCR (qPCR) can be used to estimate the relative percentage of chloroplast DNA in a sample (Lutz et al., 2011). Through qPCR, cpDNA percentage for each taxon can be estimated, allowing for optimal multiplexing of samples to relatively equal chloroplast read representation across taxa. Ultimately, the number of taxa in a sequencing run is related to the sequencing potential of the run (i.e., total reads and read length), the average size of the chloroplast genome for the group, the desired average coverage of the chloroplast genome, and the relative percentage of total genomic DNA that comes from the chloroplast. An Illumina run of 100 million 150-bp paired-end reads is capable of sequencing 50–60 samples for complete chloroplast genomes (Teisher et al., 2017), although this will vary among taxa and DNA source (fresh vs. herbarium).

More than just chloroplast genomes

In addition to chloroplast genomes, genome skimming provides a first look at genome composition. Genome skimming provides data for both development of target enrichment probes (Schmickl et al., 2016) and the isolation of low-copy nuclear genes given sufficient coverage (Berger et al., 2017). Transposable element diversity and composition can also be discerned through genome skim data. Although multiple studies have focused on the use of longer read technology such as 454 pyrosequencing (e.g., Harkess et al., 2016), the development of RepeatExplorer (Novák et al., 2013) and Transposome (Staton and Burke, 2015b) enable short reads to be used for transposon identification (e.g., Staton and Burke, 2015a). Using these approaches, genome skim data are able to provide novel insights into the evolution of nuclear genomes, not just organellar genomes. When used in conjunction with chloroplast-based phylogenies derived from the same data, understanding of transposon and genome evolution is greatly extended. These types of analyses are specifically suited to genome skimming, which provides unbiased genome sampling compared to enrichment-based approaches. Genome skimming also allows for the assembly of other highly repetitive nuclear regions, such as nuclear ribosomal DNA (Kim et al., 2015), potentially providing phylogenetically informative nuclear loci.

Limitations caused by non-biparental inheritance in reconstructed phylogenies

As in most sequencing-based projects, some taxa can be difficult to work with and not all samples will result in useful data due to DNA quality. Genome skimming can provide complete chloroplast genome sequences, but these can be limiting in the resolution they offer to phylogenomics projects if hybridization and polyploidy are abundant. Chloroplast genomes are usually uniparentally inherited, which becomes problematic in the identification of hybridization and polyploid events. Phylogenetic signal from whole chloroplast genomes can, at times, suggest incomplete lineage sorting or introgression (i.e., chloroplast capture), especially at the inter- and intraspecific levels (Wambugu et al., 2015; Zhou et al., 2017). It can be difficult to tease apart incomplete lineage sorting and hybridization because they display similar phylogenetic patterns, especially if only chloroplast genomes are considered. In cases in which bidirectional hybridization occurs and both parental species donate chloroplast genomes, it may be possible to detect a hybridization event using chloroplast-based phylogenies. However, a combined approach of nuclear and chloroplast genomes will be much more powerful than either alone. As such, both the biology of the species (i.e., hybridization frequency, recent radiations) and the research questions must be amenable to organellar-based phylogenies for genome skimming to be successful. It should be noted, however, that chloroplast phylogenies often recover comparable phylogenetic relationships to those seen in nuclear-based studies (e.g., Gitzendanner et al., 2018), suggesting that these limitations are situation specific. Although the use of genome skimming for transposon diversity studies is promising, the methodology has primarily been tested in lineages with well-documented transposons (e.g., Asteraceae and Poaceae) and may be less accurate in understudied lineages due to fewer or no reference genomes being available.

TARGET ENRICHMENT

Probe design from available data → DNA extraction, library preparation, hybridization, and sequencing → locus assembly → homology and orthology inference → phylogenetic inference

Although genome-skimming methods are useful for extracting phylogenetic data from organellar and other high-copy regions in plants (Stull et al., 2013), they are not typically feasible for recovering nuclear genes. Several methods have emerged for enrichment of shotgun (i.e., Illumina) sequencing libraries for genes of interest, including ultraconserved elements (Faircloth et al., 2012), anchored phylogenomics (Lemmon et al., 2012), exon capture (Mandel et al., 2014), and Hyb-Seq (Weitemier et al., 2014). Each of these methods, which we will collectively refer to as "target enrichment" (Mamanova et al., 2010), work via the use of short (60–120 bp) RNA probes that hybridize to sequence library fragments. The hybridized fragments are typically bound to magnetic beads while the remainder of the library is discarded. Commonly used target enrichment methods differ in the types of DNA that are targeted. Ultraconserved elements (UCEs) and anchored phylogenomics target slow-evolving genomic regions (that may or may not be associated with protein-coding genes) using universal probes that can be used across a wide phylogenetic diversity of organisms. These methods rely on sequence variation in regions flanking the conserved genomic elements. In contrast, protein-coding genes are used to design probes for exon capture and Hyb-Seq approaches. Depending on the phylogenetic breadth of the study, the exon data may be used directly. Alternatively, flanking intron regions are also captured and can be useful for informing more recent relationships. In the Hyb-Seq approach, exon capture is combined with analysis of off-target organellar reads to retrieve both nuclear and organellar data in the same sequencing run.

Consistently recovering large regions from multiple DNA sources

As with PCR-based methods, target enrichment requires some prior genomic knowledge about the target organisms. Although several genomes spanning the breadth of target organisms are required for probe design for the UCE and anchored phylogenomics methods, they are not required for Hyb-Seq. Furthermore, it has proven difficult to identify ultraconserved elements in plants, likely due to the high amounts of genome duplication. For target enrichment in plants, many groups have instead chosen to focus on low-copy protein-coding genes, using exon capture or Hyb-Seq designs (Johnson et al., 2016; Crowl et al., 2017; Landis et al., 2017; Villaverde et al., unpublished manuscript). Target enrichment has also been used to capture chloroplast exons directly (Medina et al., 2018; Heyduk et al., 2016a, 2016b).

The availability of public transcriptome data, including over 1400 green plants as part of the One Thousand Plants project (OneKP; Matasci et al., 2014), has simplified the process of probe design for many plant groups. Transcriptome sequence data are now available for most angiosperm plant families and can be used for Hyb-Seq probe design. Predicted protein sequences from several species can be sorted into low-copy orthogroups based on sequence similarity using tools such as OrthoFinder (Emms and Kelly, 2015). One disadvantage of the transcriptome-only approach to probe design is that probes spanning intron boundaries will not be effective during sequence capture. Identification of intron-exon boundaries is possible using MarkerMiner (Chamala et al., 2015), which aligns transcriptome data to reference genome sequences and returns intron-masked multiple-sequence alignments. If no reference genome is available, a low-coverage genome sequence (10–15× coverage) can also be used to design probes around intron boundaries (Gardner et al., 2016). Finally, the pipeline Sondovač (Schmickl et al., 2016) uses a combination of transcriptome and genome skimming data to identify possible nuclear exons, and their introns, to be captured. There are many filtering steps in the pipeline to assure that the target loci are orthologous and putatively single copy.

It is generally advisable to design target enrichment probes using orthologous sequences from multiple species. In addition to ensuring the loci are truly single-copy in the target taxa, probes designed from orthologous sequences will extend the breadth of phylogenetic utility of the probe set (Johnson et al., 2016; Villaverde et al., unpublished manuscript). Further discussion of bait design considerations can be found at: https://github.com/mossmatters/KewHybSeqWorkshop. A number of broad-scale target enrichment projects are under development in plants, including by the Plant and Fungal Tree of Life (PAFTOL; Royal Botanical Gardens Kew, Richmond, Surrey, United Kingdom), Genealogy of Flagellate Plants (GoFlag; University of Florida, Gainesville, Florida, USA), and Anchored Phylogenomics (Léveillé-Bourret et al., 2018) groups. The possibility of a universal set of genes that can be used for any plant species is an exciting future direction for targeted sequencing.

Being able to use herbarium specimens in phylogenomics analyses is one of the major advantages of the target enrichment approach. DNA from herbarium specimens is often degraded into very small fragments, meaning PCR-based approaches are unsuccessful at amplifying loci. In contrast, target enrichment has proven successful for antique DNA collections in many organisms, including 100-year-old herbarium specimens (Villaverde et al., unpublished manuscript). Target enrichment also has an advantage over phylotranscriptomic methods in groups in which live tissue is difficult to collect but herbarium collections exist.

Workflow easily accomplished in a modern molecular lab

A typical molecular lab workflow would involve six steps: DNA extraction, genomic DNA fragmentation via sonication, HTS library preparation, sequence capture, PCR, and sequencing. The sonication step may be omitted for many herbarium specimens, which typically have highly fragmented DNA. Depending on the methods of DNA extraction and library preparation, a 96-well plate of samples may be prepared for sequencing in as little as two or three weeks. One difference between genome skimming and target enrichment is that it may not be economical to send DNA extracts to a third-party for library preparation. The libraries would need to be returned to researchers for hybrid enrichment and then sent back for sequencing. Additional cost-cutting measures may be used, including the preparation of streptavidin beads. For an example of one possible low-cost workflow, see: https://github.com/mossmatters/KewHybSeqWorkshop.

If probe design is conducted using existing transcriptome and genomic resources, there is little up-front cost for target enrichment studies. The typical cost for probe sequences and target enrichment reagents is US$200 per reaction with myBaits kits (Arbor Biosciences, Ann Arbor, Michigan, USA), and a single reaction can be used to enrich up to 96 Illumina libraries for hundreds of loci. The cost of library preparation is similar to other methods and can utilize typical DNA library preparation kits such as TruSeq Nano (Illumina Inc.) or more economical kits such as those provided by KAPA (Roche Sequencing, Pleasanton, California, USA) and NEBNext (New England BioLabs, Ipswich, Massachusetts, USA). Studies may employ different strategies for sequencing: for example,

if capture of flanking intron regions is important, MiSeq 2 × 300 PE reads would be ideal, whereas if many herbarium specimens are used, a shorter read length may be a more appropriate choice.

Data analysis is amendable to type of enrichment

After sequencing, data analysis involves reconstructing the loci from sequencing reads before proceeding to sequence alignment and phylogenetic reconstruction. Several pipelines have been developed to assist: for example, HybPiper (Johnson et al., 2016) was designed for Hyb-Seq and exon-based approaches, whereas Phyluce (Faircloth, 2015) was designed for the UCE approach. Users should pay special attention to the detection of paralogous sequences recovered by sequencing. In some cases, paralogous sequences may not affect further analysis (if they are recent enough to be monophyletic for each sample). In other cases, paralogs may prove useful to identify further loci for phylogenetics; when the relative age of a genome duplication is known, reads from the two paralogs can be sorted and assembled into separate, orthologous alignments (e.g., see Johnson et al., 2016).

Sequence alignments generated by target capture can be concatenated into a supermatrix or used for gene-tree-based methods of phylogenetic reconstruction, including ASTRAL-III (Mirarab et al., 2014; Zhang et al., 2017), ASTRID (Vachaspati and Warnow, 2015), and BUCKy (Larget et al., 2010). This is especially useful with exon capture and Hyb-Seq approaches, because loci are likely to be long enough to contain many variable sites and produce gene trees with high confidence (Folk et al., 2015; Johnson et al., 2016; Villaverde et al., unpublished manuscript). Filtering potential target genes to ensure the presence of long coding regions (>1000 bp) and, when intron position is known, long exon regions (>500 bp) will increase the probability that gene trees may be resolved. Recovery of organellar DNA as a byproduct of nuclear target enrichment can depend on many factors, including how much of the extraction contains organellar DNA and the efficiency of target enrichment. One method for increasing the off-target organellar coverage is to add a dilution of the unenriched library to the post-hybridization library (K. Weitemier, pers. comm.).

Large segment enrichment

Recent advances in sequence capture have introduced the capacity to enrich for not only exons but also intergenic regions. One such method, known as region-specific extraction (Dapprich et al., 2016), utilizes primers, designed in similar fashion as target enrichment probes, and a second-strand synthesis using biotinylated nucleotides that facilitates the enrichment of long pieces of DNA using a standard magnet approach. This method allows for sampling of long and phylogenetically informative regions outside of exonic regions and has the potential to allow for easy assembly and identification of paralogs, acquisition of conserved regulatory regions, and identification of structural variation that would otherwise not be obtainable in taxa without fully sequenced genomes. Another approach is the use of capture probes on gene-seized DNA fragments, followed by sequencing with long-read technology such as Nanopore and PacBio (Giolai et al., 2016, 2017). This approach has the benefit of being very similar to general target enrichment, making adoption an easier transition. Another strength of these approaches is the possibility of targeted sequencing of large genomic regions in non-model species, a powerful tool for both phylogenomics and evolutionary genomics.

TRANSCRIPTOMES

Live tissue → RNA extraction, library preparation, and sequencing → transcriptome assembly → homology and orthology inference → phylogenetic inference

Using transcriptome sequencing to generate protein coding sequences for phylogenomics, or "phylotranscriptomics," has the versatility to inform relationships from closely related species (Pease et al., 2016) to ancient relationships with relatively slow-evolving coding sequences (Wickett et al., 2014). Transcriptomes contain rich information on both gene sequences and gene expression. No prior knowledge of sequences is required, and the transcriptome data generated for one project are reusable for a different project. Although phylotranscriptomics has been gaining popularity during the past several years, its use is still restricted to a relatively small number of research groups due to the relatively high cost per sample (US$260–350 plus labor) and logistics in obtaining and handling tissue due to unstable RNA molecules. In addition, data analysis of transcriptome data requires command line tools and overcoming computational challenges such as handling isoforms, incomplete/missing gene sequences, and misassembly. However, most of these hurdles have been lowered by recently developed equipment, commercially available kits, and analytical tools.

Proper planning for collection of living tissue

Living collections, seed banks, and reputable commercial seed and live plant providers are the primary sources for tissue used for phylotranscriptomics. To supplement existing collections, there are two approaches that can be used to collect tissue samples suitable for transcriptome analysis from wild populations. RNA*later* (Thermo Fisher Scientific, Waltham, Massachusetts, USA) stabilizes both DNA and RNA and allows for collecting samples in ambient temperature. The key step for using RNA*later* is to slice thick tissue thin so that the solution penetrates tissue quickly. An alternative strategy, using liquid nitrogen, is logistically more challenging but enables a wider range of analyses in addition to DNA and RNA, such as protein and secondary metabolites (Sedio et al., 2018). Having collaborators based in a local institution near the field site allows for support in shipping and storing equipment and samples, especially if collecting internationally. Permits for collecting tissue for RNA isolation may be more difficult to obtain compared to those needed for silica-preserved samples. Advanced planning and communication are essential for planning these trips. Once tissue samples are obtained, long-term storage of either RNA*later*-preserved or flash-frozen materials can be expensive as they occupy freezer space and do not tolerate thawing. See Yang et al. (2017) for an example of a field collection and tissue storage workflow.

Standardization of tissue may not be critical

Due to the dynamic nature of gene expression, one of the most frequently asked questions in phylotranscriptomic project design is what tissue to use. Traditionally, most transcriptome studies have focused on differentially expressed genes. For phylotranscriptomic purposes, we typically aim to recover as many genes as possible, especially housekeeping genes. Ideally a mixture of plant tissues should be used. However, logistic constraints of field collection often limit collection to vegetative tissues that vary in growth stages. Conditions such as temperature, day length, and time of day for

collecting can be difficult to standardize but are less important when the goal is to recover housekeeping genes. A useful rule of thumb is to collect young leaves, flower buds, and meristems, which have a relatively high RNA concentration and are low in secondary metabolites compared to mature leaves, making them easier to work with (Johnson et al., 2012).

Avoiding contamination in RNA extraction

A number of phylotranscriptomic protocols have been developed in various plant lineages (Johnson et al., 2012; Yockteng et al., 2013; Jordon-Thaden et al., 2015; Yang et al., 2017). Suitable extraction protocols can be lineage specific, and we recommend starting from existing protocols that have proven effective in closely related plant groups. Despite having to keep tissue frozen until extraction, the RNA extraction procedure is quite similar to DNA extraction. RNA extraction is best done in small batches of 12 or less because it is necessary to move relatively quickly to avoid RNA degradation.

Extreme care should be taken to avoid contamination, especially from closely related plants. This is because sequencing coverage varies by several orders of magnitude among genes in any given transcriptome. Unlike DNA analysis, in which contamination can be filtered out by low sequence coverage, highly expressed genes from contaminants can be difficult to remove analytically, especially if they are from a closely related species. Unfortunately, publicly available transcriptome data retrieved from the National Center for Biotechnology Information (NCBI) Short Read Archive (SRA) often contain contaminated reads, or even hybrid or mixed samples. As *rbcL* and *matK* genes can often be recovered from transcriptomes, they can be used to compare against sequences in GenBank to detect potential contamination. The recently developed tool CroCo (Simion et al., 2018) can be used to detect potential cross-contamination from transcriptome data sets generated by the same research group.

Data analysis requires powerful computers and command line tools

Due to memory requirements for de novo transcriptome assembly, a high-end desktop computer with at least 64 Gb of memory or high-performance computing clusters are needed for data processing. Although point-and-click software platforms such as Galaxy (Afgan et al., 2016), CLC Genomics Workbench (QIAGEN, Valencia, California, USA), and DNA Subway (https://dnasubway.cyverse.org/) are available for de novo assembly, downstream analysis steps such as data filtering, homology and orthology inference, matrix construction, and gene tree analyses are still active areas of research and development. No existing point-and-click tools can properly handle the entire phylotranscriptomic workflow, and command line skills are required to properly analyze transcriptome data. We highly recommend familiarizing yourself with a scripting language (such as Python) and Unix command line tools through bioinformatics courses, workshops, and online classes (such as those offered through Coursera [https://www.coursera.org/]). Recently, Carey and Papin (2018) published a guide for biologists learning to program, which provides a practical resource for those just getting started. The "simple fool's guide" (De Wit et al., 2012) and the Eel Pond mRNAseq Protocol (http://khmer-protocols.readthedocs.io/en/v0.8.4/mrnaseq/index.html) are good examples to start with for data analysis, although updated protocols should be considered when available.

Isoforms, incomplete sequences, and gene and genome duplication

With proper orthology inference and filtering, phylotranscriptomic data sets using housekeeping genes can achieve matrix occupancy similar to Sanger-based methods (Yang and Smith, 2014; Yang et al., 2015, 2018). Methods developed without explicit consideration for gene and genome duplication events, such as HaMStR (Ebersberger et al., 2009) and OrthoMCL (Li et al., 2003), are potentially problematic in plants, especially when non-single-copy gene families are of interest. Recently developed tools such as OrthoFinder (Emms and Kelly, 2015), in our experience, perform better than OrthoMCL in retaining gene family structure instead of breaking gene families apart. Approaches such as constructing phylomes (the collection of gene phylogenies for a taxon; Huerta-Cepas et al., 2011) followed by tree-based orthology pruning (Yang et al., 2018) or all-by-all BLAST followed by Markov clustering and tree-based orthology pruning (Yang and Smith, 2014) are more appropriate for the challenges of plant orthology inference, especially with complex gene and genome duplication scenarios. Optimal homology and orthology inference in plants is still an active research area, with novel tools being developed by multiple research groups including the Joint Genome Institute (https://phytozome.jgi.doe.gov). Multiple challenges remain in orthology inference: all-by-all homology search is computationally intensive (less of a concern with DIAMOND; Buchfink et al., 2015), inflation values have an unknown impact in Markov clustering (van Dongen, 2000), and *E*-value saturation effects from BLAST are unknown. On the other hand, baited methods, such as building phylomes or sorting transcriptomes using a core set of orthogroups (e.g., McKain et al., 2016b), rely on a high-quality core gene set or focal proteome, as errors and incompleteness in these sets can propagate into subsequent analyses.

Detecting gene and genome duplication events

Due to the complexity of de novo transcriptome assembly, it is difficult to distinguish isoforms and alleles from recently duplicated paralogs. Our experience is that detecting polyploids formed during the past few million years is often difficult using transcriptomes due to paralog divergence, taxon sampling, and incomplete lineage sorting. Transcriptome data are suitable for detecting more ancient polyploidy events given proper homology inference (Li et al., 2015; McKain et al., 2016b; Yang et al., 2018). Large data sets, such as OneKP, have demonstrated the utility of such an approach through the identification of hundreds of whole genome duplication events (M. Barker, pers. comm.). Moving forward, some exciting aspects of phylotranscriptomic analysis include gene loss/silencing, relative expression levels, and substitution rates between paralogous pairs.

In summary, while sample handling is delicate, with transcriptome data, housekeeping genes can be used for species tree reconstruction, while gene duplication and loss/silencing can be used for functional inference. The learning curve for data analysis is steep, but the return allows for novel biological insights in non-model systems.

DISCUSSION

With so many phylogenomic methods available for plants, many limitations that previously plagued systematics projects can be

TABLE 2. Recommendations for data sharing and archiving.

Archive format/ platform	Best practice	Information to include	Special considerations
Vouchers	Deposit specimen with appropriate characteristics to identify to species level in a herbarium	GPS point, locality data, collector, collection number	Permits often required. Special permission for living collections (e.g., botanical gardens, arboreta)
NCBI Short Read Archive	Submit all raw read data	Taxon (as specific as possible); voucher information; tissue type; methods for collection, extraction, and library preparation; read type (paired or single end)	Submit biological replicates separately if sampling for RNA-Seq experiment. Link populations and accessions together in a BioProject
Dryad	Provide details to reproduce results including commands, scripts, program versions, and log files. Major steps in data analysis should be included. Provide final data sets from which major conclusions are drawn	Cleaned and assembled reads; intermediate and final analysis files; parameters for analyses; scripts as used in associated analyses; details not presented in manuscript but necessary to replicate results	Links to Github, Bitbucket, or other online repositories for updated versions of scripts; simply stating "custom scripts" is not acceptable. Provide documentation of code used and parameters, such as a Jupyter Notebook (http://jupyter.org), to promote repeatability

alleviated through a multifaceted approach. Good planning with insight into the biology of one's system as well as into the potential limitations of the methods used can improve the likelihood of success. When planning a phylogenomic project, one should first consider data already available from public databases such as NCBI (GenBank, SRA, and the Transcriptome Shotgun Assembly Sequence Database [TSA]) and Phytozome (https://phytozome.jgi.doe.gov). It is increasingly common to start by summarizing and/or re-analyzing existing data when designing phylogenomic projects. Existing data can give a researcher a head start, circumventing the need to generate preliminary data when they are necessary for certain approaches (e.g., microfluidic PCR and target enrichment). Although the phylogenomic community has been moving toward increased transparency and data/code sharing, more often than not it is frustrating to re-use published data sets. As such, researchers should be sure to contribute responsibly to the community by making data and analyses openly available and well-annotated with metadata. Additionally, vouchers should be made for samples to link a physical plant specimen to generated data (Funk et al., 2017). Here, we make recommendations on metadata and data sharing to optimize re-usability of data and transparency of research (Table 2).

Given your short- and long-term research goals, consider the longevity of the data generated and ask the question: would my data become obsolete in five years? Tissue collection should take into consideration the improvement of phylogenomic approaches, as well as other future approaches. For example, is collecting seeds and frozen or RNA*later*-preserved tissue in addition to silica-preserved samples feasible for at least some species? These will provide resources for future transcriptome and whole genome sequencing, even if it is not the current goal of a project. With the advance of whole genome sequencing, a well-curated living collection will become increasingly important not only to you but to the community.

Multiple approaches may be combined for developing phylogenomic projects. Transcriptomes and genomes can be used to design Hyb-Seq and PCR primers. Whole genomes can be used for reference-based mapping of transcriptome, RAD-seq, and target enrichment data and help inform lineages with recent divergence or phase homologous sequences. Combinations of approaches can provide novel insight into relationships by using the strengths of these approaches to inform each other. For example, identification of a hybridization event is possible through most of the approaches depicted here. The uniparental inheritance of organellar genomes recovered from genome skimming can elucidate the parental history of a hybrid, identifying the maternal genome donor and determining whether the event is unidirectional or bidirectional.

Finally, it is becoming increasingly attractive to develop "model clades" with a combination of whole genomes, transcriptomes, Hyb-Seq/genome skimming at species level, and RAD-seq at population level. With a suite of tools, we can not only reconstruct the evolutionary history of these clades but also start asking questions about genetic mechanisms underlying trait evolution and adaptation. Phylogenomic methods provide much more than just evolutionary history, they provide insight into different aspects of a plant's genomes, which can lead to novel discoveries in previously intractable lineages.

ACKNOWLEDGMENTS

We would like to thank the editors of this special issue for the invitation to submit this review, D. Morales-Briones for comments on the manuscript, and two anonymous reviewers for help improving the clarity and expanding the scope of the manuscript.

LITERATURE CITED

Afgan, E., D. Baker, M. van den Beek, D. Blankenberg, D. Bouvier, M. Čech, J. Chilton, et al. 2016. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Research* 44(W1): W3–W10.

Andrews, K. R., J. M. Good, M. R. Miller, G. Luikart, and P. A. Hohenlohe. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics* 17: 81.

Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, Z. A. Lewis, E. U. Selker, et al. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* 3: e3376.

Bakker, F. T., D. Lei, J. Yu, S. Mohammadin, Z. Wei, S. van de Kerke, B. Gravendeel, et al. 2016. Herbarium genomics: Plastome sequence assembly from a range of herbarium specimens using an Iterative Organelle Genome Assembly pipeline. *Biological Journal of the Linnean Society* 117: 33–43.

Baldwin, B. G. 1998. Phylogenetic utility of the external transcribed spacer (ETS) of 18S-26S rDNA: Congruence of ETS and ITS trees of *Calycadenia* (Compositae). *Molecular Phylogenetics and Evolution* 10: 449–463.

Baldwin, B. G., M. J. Sanderson, J. M. Porter, M. F. Wojciechowski, C. S. Campbell, and M. J. Donoghue. 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.

Benjamini, Y., and T. P. Speed. 2012. Summarizing and correcting GC content bias in high-throughput sequencing. *Nucleic Acids Research* 40: e72.

Berger, B. A., J. Han, E. B. Sessa, A. G. Gardner, K. A. Shepherd, V. A. Ricigliano, R. S. Jabaily, and D. G. Howarth. 2017. The unexpected depths of genome-skimming data: A case study examining Goodeniaceae floral symmetry genes. *Applications in Plant Sciences* 5(10): 1700042.

Bhat, S., A. M. Polanowski, M. C. Double, S. N. Jarman, and K. R. Emslie. 2012. The effect of input DNA copy number on genotype call and characterising SNP markers in the humpback whale genome using a nanofluidic array. *PLoS One* 7: e39181.

Bock, D. G., N. C. Kane, D. P. Ebert, and L. H. Rieseberg. 2014. Genome skimming reveals the origin of the Jerusalem Artichoke tuber crop species: Neither from Jerusalem nor an artichoke. *New Phytologist* 201: 1021–1030.

Bryant, D., R. Bouckaert, J. Felsenstein, N. A. Rosenberg, and A. RoyChoudhury. 2012. Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution* 29: 1917–1932.

Buchfink, B., C. Xie, and D. H. Hudson. 2015. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* 12: 59–60.

Byers, R. L., D. B. Harker, S. M. Yourstone, P. J. Maughan, and J. A. Udall. 2012. Development and mapping of SNP assays in allotetraploid cotton. *Theoretical and Applied Genetics* 124: 1201–1214.

Carbonell-Caballero, J., R. Alonso, V. Ibañez, J. Terol, M. Talon, and J. Dopazo. 2015. A phylogenetic analysis of 34 chloroplast genomes elucidates the relationships between wild and domestic species within the genus *Citrus*. *Molecular Biology and Evolution* 32: 2015–2035.

Carey, M. A., and J. A. Papin. 2018. Ten simple rules for biologists learning to program. *PloS Computational Biology* 14: e1005871.

Cariou, M., L. Duret, and S. Charlat. 2013. Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization. *Ecology and Evolution* 3: 846–852.

Catchen, J., P. A. Hohenlohe, S. Bassham, A. Amores, and W. A. Cresko. 2013. Stacks: An analysis tool set for population genomics. *Molecular Ecology* 22: 3124–3140.

Chamala, S., N. García, G. T. Godden, V. Krishnakumar, I. E. Jordon-Thaden, R. De Smet, W. B. Barbazuk, et al. 2015. MarkerMiner 1.0: A new application for phylogenetic marker development using angiosperm transcriptomes. *Applications in Plant Sciences* 3(4): 1400115.

Chifman, J., and L. Kubatko. 2014. Quartet inference from SNP data under the coalescent model. *Bioinformatics* 30: 3317–3324.

Collins, E. S., M. R. Gostel, and A. Weeks. 2016. An expanded phylogenomic PCR toolkit for Sapindales. *Applications in Plant Sciences* 4(12): 1600078.

Cronn, R., B. J. Knaus, A. Liston, P. J. Maughan, M. Parks, J. V. Syring, and J. Udall. 2012. Targeted enrichment strategies for next-generation plant biology. *American Journal of Botany* 99: 291–311.

Crowl, A. A., C. Myers, and N. Cellinese. 2017. Embracing discordance: Phylogenomic analyses provide evidence for allopolyploidy leading to cryptic diversity in a Mediterranean *Campanula* (Campanulaceae) clade. *Evolution* 71: 913–922.

Cruaud, P., J.-Y. Rasplus, L. J. Rodriguez, and A. Cruaud. 2017. High-throughput sequencing of multiple amplicons for barcoding and integrative taxonomy. *Scientific Reports* 7: 41948.

Curk, F., G. Ancillo, A. Garcia-Lor, F. Luro, X. Perrier, J.-P. Jacquemoud-Collet, L. Navarro, and P. Ollitrault. 2014. Next generation haplotyping to decipher nuclear genomic interspecific admixture in *Citrus* species: Analysis of chromosome 2. *BMC Genetics* 15: 152.

DaCosta, J. M., and M. D. Sorenson. 2016. ddRAD-seq phylogenetics based on nucleotide, indel, and presence–absence polymorphisms: Analyses of two avian genera with contrasting histories. *Molecular Phylogenetics and Evolution* 94: 122–135.

Dapprich, J., D. Ferriola, K. Mackiewicz, P. M. Clark, E. Rappaport, M. D'Arcy, A. Sasson, et al. 2016. The next generation of target capture technologies: Large DNA fragment enrichment and sequencing determines regional genomic variation of high complexity. *BMC Genomics* 17: 486.

Dasmahapatra, K. K., J. R. Walters, A. D. Briscoe, J. W. Davey, A. Whibley, N. J. Nadeau, A. V. Zimin, et al. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* 487: 94.

Davey, J. W., P. A. Hohenlohe, P. D. Etter, J. Q. Boone, J. M. Catchen, and M. L. Blaxter. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* 12: 499.

De Wit, P., M. H. Pespeni, J. T. Ladner, D. J. Barshis, and S. R. Palumbi. 2012. The simple fool's guide to population genomics via RNA-Seq: An introduction to high-throughput sequencing data analysis. *Molecular Ecology Resources* 12: 1058–1067.

Dierckxsens, N., P. Mardulyn, and G. Smits. 2017. NOVOPlasty: De novo assembly of organelle genomes from whole genome data. *Nucleic Acids Research* 45: e18.

Dominguez, M. H., P. K. Chattopadhyay, S. Ma, L. Lamoreaux, A. McDavid, G. Finak, R. Gottardo, et al. 2013. Highly multiplexed quantitation of gene expression on single cells. *Journal of Immunological Methods* 391: 133–145.

Durand, E. Y., N. Patterson, D. Reich, and M. Slatkin. 2011. Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution* 28: 2239–2252.

Eaton, D. A. R. 2014. PyRAD: Assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics* 30: 1844–1849.

Eaton, D. A. R., and R. H. Ree. 2013. Inferring phylogeny and introgression using RADseq data: An example from flowering plants (*Pedicularis*: Orobanchaceae). *Systematic Biology* 62: 689–706.

Eaton, D. A. R., A. L. Hipp, A. González-Rodríguez, and J. Cavender-Bares. 2015. Historical introgression among the American live oaks and the comparative nature of tests for introgression. *Evolution* 69: 2587–2601.

Eaton, D. A. R., E. L. Spriggs, B. Park, and M. J. Donoghue. 2017. Misconceptions on missing data in RAD-seq phylogenetics with a deep-scale example from flowering plants. *Systematic Biology* 66: 399–412.

Ebersberger, I., S. Strauss, and A. von Haeseler. 2009. HaMStR: Profile hidden markov model based search for orthologs in ESTs. *BMC Evolutionary Biology* 9: 157.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler, and S. E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6: e19379.

Emms, D. M., and S. Kelly. 2015. OrthoFinder: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* 16: 157.

Escudero, M., D. A. R. Eaton, M. Hahn, and A. L. Hipp. 2014. Genotyping-by-sequencing as a tool to infer phylogeny and ancestral hybridization: A case study in *Carex* (Cyperaceae). *Molecular Phylogenetics and Evolution* 79: 359–367.

Faircloth, B. C. 2015. PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics* 32: 786–788.

Faircloth, B. C., J. E. McCormack, N. G. Crawford, M. G. Harvey, R. T. Brumfield, and T. C. Glenn. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology* 61: 717–726.

Folk, R. A., J. R. Mandel, and J. V. Freudenstein. 2015. A protocol for targeted enrichment of intron-containing sequence markers for recent radiations: A phylogenomic example from *Heuchera* (Saxifragaceae). *Applications in Plant Sciences* 3(8): 1500039.

Funk, V. A., M. Gostel, A. Devine, C. L. Kelloff, K. Wurdack, C. Tuccinardi, A. Radosavljevic, et al. 2017. Guidelines for collecting vouchers and tissues intended for genomic work (Smithsonian Institution): Botany Best Practices. *Biodiversity Data Journal* 5: e11625.

Gaedcke, J., M. Grade, J. Camps, R. Sokilde, B. Kaczkowski, A. J. Schetter, M. J. Difilippantonio, et al. 2012. The rectal cancer microRNAome: microRNA expression in rectal cancer and matched normal mucosa. *Clinical Cancer Research* 18: 4919–4930.

Gardner, E. M., M. G. Johnson, D. Ragone, N. J. Wickett, and N. J. C. Zerega. 2016. Low-coverage, whole-genome sequencing of *Artocarpus camansi* (Moraceae) for phylogenetic marker development and gene discovery. *Applications in Plant Sciences* 4(7): 1600017.

Giolai, M., P. Paajanen, W. Verweij, L. Percival-Alwyn, D. Baker, K. Witek, F. Jupe, et al. 2016. Targeted capture and sequencing of gene-sized DNA molecules. *Biotechniques* 61: 315–322.

Giolai, M., P. Paajanen, W. Verweij, K. Witek, J. D. G. Jones, and M. D. Clark. 2017. Comparative analysis of targeted long read sequencing approaches for characterization of a plant's immune receptor repertoire. *BMC Genomics* 18: 564.

Gitzendanner, M. A., P. S. Solits, G. K-S. Wong, B. R. Ruhfel, and D. E. Solits. 2018. Plastid phylogenomic analysis of green plants: A billion years of evolutionary history. *American Journal of Botany* 105: in press.

Glenn, T. C., R. Nilsen, T. J. Kieran, J. W. Finger, T. W. Pierson, K. E. Bentley, S. Hoffberg, et al. 2016. Adapterama I: Universal stubs and primers for thousands of dual-indexed Illumina libraries (iTru & iNext). *bioRxiv* 49114.

Glenn, T. C., N. J. Bayona-Vasquez, T. J. Kieran, T. W. Pierson, S. L. Hoffberg, P. A. Scott, K. E. Bentley, et al. 2017. Adapterama III: Quadruple-indexed, triple-enzyme RADseq libraries for about $1USD per sample (3RAD). *bioRxiv* 205799.

Gompert, Z., and K. E. Mock. 2017. Detection of individual ploidy levels with genotyping-by-sequencing (GBS) analysis. *Molecular Ecology Resources* 17: 1156–1167.

Gostel, M. R., K. A. Coy, and A. Weeks. 2015. Microfluidic PCR-based target enrichment: A case study in two rapid radiations of *Commiphora* (Burseraceae) from Madagascar. *Journal of Systematics and Evolution* 53: 411–431.

Graham, C. F., T. C. Glenn, A. G. McArthur, D. R. Boreham, T. Kieran, S. Lance, R. G. Manzon, et al. 2015. Impacts of degraded DNA on restriction enzyme associated DNA sequencing (RADSeq). *Molecular Ecology Resources* 15: 1304–1315.

Gregg, W. C. T., S. H. Ather, and M. W. Hahn. 2017. Gene-tree reconciliation with MUL-Trees to resolve polyploidy events. *Systematic Biology* 66: 1007–1018.

Guo, W., F. Grewe, W. Fan, G. J. Young, V. Knoop, J. D. Palmer, and J. P. Mower. 2016. *Ginkgo* and *Welwitschia* mitogenomes reveal extreme contrasts in gymnosperm mitochondrial evolution. *Molecular Biology and Evolution* 33: 1448–1460.

Harkess, A., F. Mercati, L. Abbate, M. McKain, J. C. Pires, T. Sala, F. Sunseri, et al. 2016. Retrotransposon proliferation coincident with the evolution of dioecy in *Asparagus*. *G3— Genes|Genomes|Genetics* 6: 2679–2685.

Hermann-Bank, M. L., K. Skovgaard, A. Stockmarr, N. Larsen, and L. Mølbak. 2013. The Gut Microbiotassay: A high-throughput qPCR approach combinable with next generation sequencing to study gut microbial diversity. *BMC Genomics* 14: 788.

Herrera, S., H. Watanabe, and T. M. Shank. 2015. Evolutionary and biogeographical patterns of barnacles from deep-sea hydrothermal vents. *Molecular Ecology* 24: 673–689.

Heyduk, K., M. R. McKain, F. Lalani, and J. Leebens-Mack. 2016a. Evolution of a CAM anatomy predates the origins of Crassulacean acid metabolism in the Agavoideae (Asparagaceae). *Molecular Phylogenetics and Evolution* 105: 102–113.

Heyduk, K., D. W. Trapnell, C. F. Barrett, and J. Leebens-Mack. 2016b. Phylogenomic analyses of species relationships in the genus *Sabal* (Arecaceae) using targeted sequence capture. *Biological Journal of the Linnean Society* 117: 106–120.

Hipp, A. L., D. A. R. Eaton, J. Cavender-Bares, E. Fitzek, R. Nipper, and P. S. Manos. 2014. A framework phylogeny of the American oak clade based on sequenced RAD data. *PLoS One* 9: e93975.

Hoffberg, S., T. J. Kieran, J. M. Catchen, A. Devault, B. C. Faircloth, R. Mauricio, and T. C. Glenn. 2016a. Adapterama IV: Sequence capture of dual-digest RADseq libraries with identifiable duplicates (RADcap). *bioRxiv* 44651.

Hoffberg, S. L., T. J. Kieran, J. M. Catchen, A. Devault, B. C. Faircloth, R. Mauricio, and T. C. Glenn. 2016b. RADcap: Sequence capture of dual-digest RADseq libraries with identifiable duplicates and reduced missing data. *Molecular Ecology Resources* 16: 1264–1278.

Hohenlohe, P. A., S. Bassham, P. D. Etter, N. Stiffler, E. A. Johnson, and W. A. Cresko. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics* 6: e1000862.

Huerta-Cepas, J., S. Capella-Gutierrez, L. P. Pryszcz, I. Denisov, D. Kormes, M. Marcet-Houben, and T. Gabaldón. 2011. PhylomeDB v3.0: An expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic Acids Research* 39: D556–D560.

Izan, S., D. Esselink, R. G. F. Visser, M. J. M. Smulders, and T. Borm. 2017. De novo assembly of complete chloroplast genomes from non-model species based on a *K*-mer frequency-based selection of chloroplast reads from total DNA sequences. *Frontiers in Plant Science* 8: 1271.

Johnson, M. G., E. M. Gardner, Y. Liu, R. Medina, B. Goffinet, A. J. Shaw, N. J. C. Zerega, and N. J. Wickett. 2016. HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in Plant Sciences* 4(7): 1600016.

Johnson, M. T. J., E. J. Carpenter, Z. Tian, R. Bruskiewich, J. N. Burris, C. T. Carrigan, M. W. Chase, et al. 2012. Evaluating methods for isolating total RNA and predicting the success of sequencing phylogenetically diverse plant transcriptomes. *PLoS One* 7: e50226.

Jordon-Thaden, I. E., A. S. Chanderbali, M. A. Gitzendanner, and D. E. Soltis. 2015. Modified CTAB and TRIzol protocols improve RNA extraction from chemically complex Embryophyta. *Applications in Plant Sciences* 3(5): 1400105.

Kates, H. R., P. S. Soltis, and D. E. Soltis. 2017. Evolutionary and domestication history of *Cucurbita* (pumpkin and squash) species inferred from 44 nuclear loci. *Molecular Phylogenetics and Evolution* 111: 98–109.

Kim, K., S.-C. Lee, J. Lee, Y. Yu, K. Yang, B.-S. Choi, H.-J. Koh, et al. 2015. Complete chloroplast and ribosomal sequences for 30 accessions elucidate evolution of *Oryza* AA genome species. *Scientific Reports* 5: 15655.

Landis, J. B., D. E. Soltis, and P. S. Soltis. 2017. Comparative transcriptomic analysis of the evolution and development of flower size in *Saltugilia* (Polemoniaceae). *BMC Genomics* 18: 475.

Larget, B. R., S. K. Kotha, C. N. Dewey, and C. Ané. 2010. BUCKy: Gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics* 26: 2910–2911.

Latvis, M., S. M. E. Mortimer, D. F. Morales-Briones, S. Torpey, S. Uribe-Convers, S. J. Jacobs, S. Mathews, and D. C. Tank. 2017. Primers for *Castilleja* and their utility across Orobanchaceae: I. Chloroplast primers. *Applications in Plant Sciences* 5(9): 1700038.

Leaché, A. D., M. K. Fujita, V. N. Minin, and R. R. Bouckaert. 2014. Species delimitation using genome-wide SNP data. *Systematic Biology* 63: 534–542.

Leaché, A. D., A. S. Chavez, L. N. Jones, J. A. Grummer, A. D. Gottscho, and C. W. Linkem. 2015. Phylogenomics of phrynosomatid lizards: Conflicting signals from sequence capture versus restriction site associated DNA sequencing. *Genome Biology and Evolution* 7: 706–719.

Leaché, A. D., and J. R. Oaks. 2017. The utility of single nucleotide polymorphis (SNP) data in phylogenetics. *Annual Review of Ecology, Evolution, and Systematics* 48: 69–84.

Lemmon, A. R., J. M. Brown, K. Stanger-Hall, and E. M. Lemmon. 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. *Systematic Biology* 58: 130–145.

Lemmon, A. R., S. A. Emme, and E. M. Lemmon. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology* 61: 727–744.

Leushkin, E. V., R. A. Sutormin, E. R. Nabieva, A. A. Penin, A. S. Kondrashov, and M. D. Logacheva. 2013. The miniature genome of a carnivorous plant *Genlisea aurea* contains a low number of genes and short non-coding sequences. *BMC Genomics* 14: 476.

Léveillé-Bourret, É., J. R. Starr, B. A. Ford, E. Moriarty Lemmon, and A. R. Lemmon. 2018. Resolving rapid radiations within angiosperm families using anchored phylogenomics. *Systematic Biology* 67: 94–112.

Li, L., C. J. Stoeckert, and D. S. Roos. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research* 13: 2178–2189.

Li, Z., A. E. Baniaga, E. B. Sessa, M. Scascitelli, S. W. Graham, L. H. Rieseberg, and M. S. Barker. 2015. Early genome duplications in conifers and other seed plants. *Science Advances* 1: e1501084.

Lohr, J. G., P. Stojanov, M. S. Lawrence, D. Auclair, B. Chapuy, C. Sougnez, P. Cruz-Gordillo, et al. 2012. Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. *Proceedings of the National Academy of Sciences USA* 109: 3879–3884.

Lu, X., L. Wang, S. Chen, L. He, X. Yang, Y. Shi, J. Cheng, et al. 2012. Genome-wide association study in Han Chinese identifies four new susceptibility loci for coronary artery disease. *Nature Genetics* 44: 890–894.

Lutz, K. A., W. Wang, A. Zdepski, and T. P. Michael. 2011. Isolation and analysis of high quality nuclear DNA with reduced organellar DNA for plant genome sequencing and resequencing. *BMC Biotechnology* 11: 54.

Male, P. J., L. Bardon, G. Besnard, E. Coissac, F. Delsuc, J. Engel, E. Lhuillier, et al. 2014. Genome skimming by shotgun sequencing helps resolve the phylogeny of a pantropical tree family. *Molecular Ecology Resources* 14: 966–975.

Mamanova, L., A. J. Coffey, C. E. Scott, I. Kozarewa, E. H. Turner, A. Kumar, E. Howard, et al. 2010. Target-enrichment strategies for next-generation sequencing. *Nature Methods* 7: 111–118.

Mandel, J. R., R. B. Dikow, V. A. Funk, R. R. Masalia, S. E. Staton, A. Kozik, L. Rieseberg, and J. M. Burke. 2014. A target enrichment method for gathering phylogenetic information from hundreds of loci: An example from the Compositae. *Applications in Plant Sciences* 2(2): 1300085.

Mastretta-Yanes, A., N. Arrigo, N. Alvarez, T. H. Jorgensen, D. Piñero, and B. C. Emerson. 2015. Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources* 15: 28–41.

Matasci, N., L.-H. Hung, Z. Yan, E. J. Carpenter, N. J. Wickett, S. Mirarab, N. Nguyen, et al. 2014. Data access for the 1,000 Plants (1KP) project. *GigaScience* 3: 10–17.

McCluskey, B. M., and J. H. Postlethwait. 2015. Phylogeny of zebrafish, a "model species", within Danio, a "model genus." *Molecular Biology and Evolution* 32: 635–652.

McKain, M. R., J. R. McNeal, P. R. Kellar, L. E. Eguiarte, J. C. Pires, and J. Leebens-Mack. 2016a. Timing of rapid diversification and convergent origins of active pollination within Agavoideae (Asparagaceae). *American Journal of Botany* 103: 1717–1729.

McKain, M. R., H. Tang, J. R. McNeal, S. Ayyampalayam, J. I. Davis, C. W. dePamphilis, T. J. Givnish, et al. 2016b. A phylogenomic assessment of ancient polyploidy and genome evolution across the Poales. *Genome Biology and Evolution* 8: evw060.

McVay, J. D., A. L. Hipp, and P. S. Manos. 2017. A genetic legacy of introgression confounds phylogeny and biogeography in oaks. *Proceedings of the Royal Society B, Biological Sciences* 284: 20170300.

Medina, R., M. G. Johnson, N. Wilding, T. Hedderson, N. J. Wickett, and B. Goffinet. 2018. Evolutionary dynamism in bryophytes: Phylogenomic inferences confirm rapid radiation in the moss family Funariaceae. *Molecular Phylogenetics and Evolution* 120: 240–247.

Miller, M. R., J. P. Dunham, A. Amores, W. A. Cresko, and E. A. Johnson. 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research* 17: 240–248.

Mirarab, S., R. Reaz, M. S. Bayzid, T. Zimmermann, M. S. Swenson, and T. Warnow. 2014. ASTRAL: Genome-scale coalescent-based species tree estimation. *Bioinformatics* 30: i541–i548.

Moignard, V., I. C. Macaulay, G. Swiers, F. Buettner, J. Schütte, F. J. Calero-Nieto, S. Kinston, et al. 2013. Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nature Cell Biology* 15: 363–372.

Moonsamy, P. V., T. Williams, P. Bonella, C. L. Holcomb, B. N. Höglund, G. Hillman, D. Goodridge, et al. 2013. High throughput HLA genotyping using 454 sequencing and the Fluidigm Access Array™ system for simplified amplicon library preparation. *Tissue Antigens* 81: 141–149.

Nadeau, N. J., S. H. Martin, K. M. Kozak, C. Salazar, K. K. Dasmahapatra, J. W. Davey, S. W. Baxter, et al. 2013. Genome-wide patterns of divergence and gene flow across a butterfly radiation. *Molecular Ecology* 22: 814–826.

Nakamura, Y., N. Sasaki, M. Kobayashi, N. Ojima, M. Yasuike, Y. Shigenobu, M. Satomi, et al. 2013. The first symbiont-free genome sequence of marine red alga, Susabi-nori (*Pyropia yezoensis*). *PLoS One* 8: e57122.

Novák, P., P. Neumann, J. Pech, J. Steinhaisl, and J. Macas. 2013. RepeatExplorer: A Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* 29: 792–793.

Ogilvie, H. A., J. Heled, D. Xie, and A. J. Drummond. 2016. Computational performance and statistical accuracy of *BEAST and comparisons with other methods. *Systematic Biology* 65: 381–396.

Pease, J. B., D. C. Haak, M. W. Hahn, and L. C. Moyle. 2016. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. *PLoS Biology* 14(2): e1002379.

Pellicer, J., M. F. Fay, and I. J. Leitch. 2010. The largest eukaryotic genome of them all? *Botanical Journal of the Linnean Society* 164: 10–15.

Petersen, G., A. Cuenca, A. Zervas, G. T. Ross, S. W. Graham, C. F. Barrett, J. I. Davis, and O. Seberg. 2017. Mitochondrial genome evolution in Alismatales: Size reduction and extensive loss of ribosomal protein genes. *PLoS One* 12: 1–21.

Peterson, B. K., J. N. Weber, E. H. Kay, H. S. Fisher, and H. E. Hoekstra. 2012. Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One* 7: e37135.

Qu, X. J., J. J. Jin, S. M. Chaw, D. Z. Li, and T. S. Yi. 2017. Multiple measures could alleviate long-branch attraction in phylogenomic reconstruction of Cupressoideae (Cupressaceae). *Scientific Reports* 7: 1–11.

Reddy, C. B., M. J. Hickerson, L. A. F. Frantz, and K. Lohse. 2017. Blockwise site frequency spectra for inferring complex population histories and recombination. *bioRxiv* 77958.

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

Rubin, B. E. R., R. H. Ree, and C. S. Moreau. 2012. Inferring phylogenies from RAD sequence data. *PLoS One* 7: e33394.

Saeidi, S., M. R. McKain, and E. A. Kellogg. 2018. Robust DNA isolation and high-throughput sequencing library construction for herbarium specimens. *Journal of Visualized Experiments (JoVE)* 133: e56837.

Särkinen, T., M. Staats, J. E. Richardson, R. S. Cowan, and F. T. Bakker. 2012. How to open the treasure chest? Optimising DNA extraction from herbarium specimens. *PLoS One* 7: e43808.

Schmickl, R., A. Liston, V. Zeisek, K. Oberlander, K. Weitemier, S. C. K. Straub, R. C. Cronn, et al. 2016. Phylogenetic marker development for target enrichment from transcriptome and genome skim data: The pipeline and its application in southern African *Oxalis* (Oxalidaceae). *Molecular Ecology Resources* 16: 1124–1135.

Sedio, B. E., C. A. Boya P., and J. C. Rojas Echeverri. 2018. A protocol for high-throughput, untargeted forest community metabolomics using mass spectrometry molecular networks. *Applications in Plant Sciences* 6(3): e1033.

Shalek, A. K., R. Satija, X. Adiconis, R. S. Gertner, J. T. Gaublomme, R. Raychowdhury, S. Schwartz, et al. 2013. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* 498: 1–5.

Shaw, J., E. B. Lickey, J. T. Beck, S. B. Farmer, W. S. Liu, J. Miller, K. C. Siripun, et al. 2005. The tortoise and the hare II: Relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *American Journal of Botany* 92: 142–166.

Shaw, J., E. B. Lickey, E. E. Schilling, and R. L. Small. 2007. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: The tortoise and the hare III. *American Journal of Botany* 94: 275–288.

Shaw, J., H. L. Shafer, O. R. Leonard, M. J. Kovach, M. Schorr, and A. B. Morris. 2014. Chloroplast DNA sequence utility for the lowest phylogenetic and phylogeographic inferences in angiosperms: The tortoise and the hare IV. *American Journal of Botany* 101: 1987–2004.

Simion, P., K. Belkhir, C. Francois, J. Veyssier, J. C. Rink, M. Manuel, H. Philippe, and M. J. Telford. 2018. A software tool 'CroCo' detects pervasive cross-species contamination in next generation sequencing data. *BMC Biology* 16: 28.

Smith, S. A., M. J. Moore, J. W. Brown, and Y. Yang. 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evolutionary Biology* 15: 150.

Staats, M., A. Cuenca, J. E. Richardson, R. Vrielink-van Ginkel, G. Petersen, O. Seberg, and F. T. Bakker. 2011. DNA damage in plant herbarium tissue. *PLoS One* 6: e28448.

Staats, M., R. H. J. Erkens, B. van de Vossenberg, J. J. Wieringa, K. Kraaijeveld, B. Stielow, J. Geml, et al. 2013. Genomic treasure troves: Complete genome sequencing of herbarium and insect museum specimens. *PLoS One* 8: e69189.

Staton, S. E., and J. M. Burke. 2015a. Evolutionary transitions in the Asteraceae coincide with marked shifts in transposable element abundance. *BMC Genomics* 16: 623.

Staton, S. E., and J. M. Burke. 2015b. Transposome: A toolkit for annotation of transposable element families from unassembled sequence reads. *Bioinformatics* 31: 1827–1829.

Steele, P. R., K. L. Hertweck, D. Mayfield, M. R. McKain, J. Leebens-Mack, and J. C. Pires. 2012. Quality and quantity of data recovered from massively parallel sequencing: Examples in Asparagales and Poaceae. *American Journal of Botany* 99: 330–348.

Straub, S. C. K., M. Parks, K. Weitemier, M. Fishbein, R. C. Cronn, and A. Liston. 2012. Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. *American Journal of Botany* 99: 349–364.

Stull, G. W., M. J. Moore, V. S. Mandala, N. A. Douglas, H.-R. Kates, X. Qi, S. F. Brockington, et al. 2013. A targeted enrichment strategy for massively parallel sequencing of angiosperm plastid genomes. *Applications in Plant Sciences* 1(2): 1200497.

Teisher, J. K., M. R. McKain, B. A. Schaal, and E. A. Kellogg. 2017. Polyphyly of Arundinoideae (Poaceae) and evolution of the twisted geniculate lemma awn. *Annals of Botany* 102: 1493–1505.

Toonen, R. J., J. B. Puritz, Z. H. Forsman, J. L. Whitney, I. Fernandez-Silva, K. R. Andrews, and C. E. Bird. 2013. ezRAD: A simplified method for genomic genotyping in non-model organisms. *PeerJ* 1: e203.

Tripp, E. A., Y.-H. E. Tsai, Y. Zhuang, and K. G. Dexter. 2017. RADseq dataset with 90% missing data fully resolves recent radiation of *Petalidium* (Acanthaceae) in the ultra-arid deserts of Namibia. *Ecology and Evolution* 7: 7920–7936.

Twyford, A. D., and R. W. Ness. 2016. Strategies for complete plastid genome sequencing. *Molecular Ecology Resources* 17(5): 858–868.

Uribe-Convers, S., M. L. Settles, and D. C. Tank. 2016. A phylogenomic approach based on PCR target enrichment and high throughput sequencing: Resolving the diversity within the South American species of *Bartsia* L. (Orobanchaceae). *PLoS One* 11(2): e0148203.

Vachaspati, P., and T. Warnow. 2015. ASTRID: Accurate Species TRees from Internode Distances. *BMC Genomics* 16(Suppl 1): S3.

Vallejo-Marín, M., A. M. Cooley, M. Y. Lee, M. Folmer, M. R. McKain, and J. R. Puzey. 2016. Strongly asymmetric hybridization barriers shape the origin of a new polyploid species and its hybrid ancestor. *American Journal of Botany* 103: 1272–1288.

van Dongen, S. 2000. Graph clustering by flow simulation. PhD thesis, University of Utrecht, Utrecht, The Netherlands.

Vargas, O. M., E. M. Ortiz, and B. B. Simpson. 2017. Conflicting phylogenomic signals reveal a pattern of reticulate evolution in a recent high-Andean diversification (Asteraceae: Astereae: *Diplostephium*). *New Phytologist* 214: 1736–1750.

Wagner, C. E., I. Keller, S. Wittwer, O. M. Selz, S. Mwaiko, L. Greuter, A. Sivasundar, and O. Seehausen. 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Molecular Ecology* 22: 787–798.

Walter, K., T. Holcomb, T. Januario, P. Du, M. Evangelista, N. Kartha, L. Iniguez, et al. 2012. DNA methylation profiling defines clinically relevant biological subsets of non-small cell lung cancer. *Clinical Cancer Research* 18: 2360–2373.

Wambugu, P. W., M. Brozynska, A. Furtado, D. L. Waters, and R. J. Henry. 2015. Relationships of wild and domesticated rices (*Oryza* AA genome species) based upon whole chloroplast genome sequences. *Scientific Reports* 5: 13957.

Weitemier, K., S. C. K. Straub, R. C. Cronn, and M. Fishbein. 2014. Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences* 2(9): 1400042.

Welch, A. J., K. Collins, A. Ratan, D. I. Drautz-Moses, S. C. Schuster, and C. Lindqvist. 2016. The quest to resolve recent radiations: Plastid phylogenomics of extinct and endangered Hawaiian endemic mints (Lamiaceae). *Molecular Phylogenetics and Evolution* 99: 16–33.

Wickett, N. J., S. Mirarab, N. Nguyen, T. Warnow, E. Carpenter, N. Matasci, S. Ayyampalayam, et al. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences USA* 111: E4859–E4868.

Wysocki, W. P., L. G. Clark, S. A. Kelchner, S. V. Burke, J. C. Pires, P. P. Edger, D. R. Mayfield, et al. 2014. A multi-step comparison of short-read full plastome sequence assembly methods in grasses. *Taxon* 63: 899–910.

Yang, Y., and S. A. Smith. 2014. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: Improving accuracy and matrix occupancy for phylogenomics. *Molecular Biology and Evolution* 31: 3081–3092.

Yang, Y., M. J. Moore, S. F. Brockington, D. E. Soltis, and G. K. Wong. 2015. Dissecting molecular evolution in the highly diverse plant clade Caryophyllales using transcriptome sequencing. *Molecular Biology and Evolution* 32: 2001–2014.

Yang, Y., M. J. Moore, S. F. Brockington, A. Timoneda, T. Feng, H. E. Marx, J. F. Walker, and S. A. Smith. 2017. An efficient field and laboratory workflow for plant phylotranscriptomic projects. *Applications in Plant Sciences* 5(3): 1600128.

Yang, Y., M. J. Moore, S. F. Brockington, J. Mikenas, J. Olivieri, J. F. Walker, and S. A. Smith. 2018. Improved transcriptome sampling pinpoints 26 ancient and more recent polyploidy events in Caryophyllales, including two allopolyploidy events. *New Phytologist* 217: 855–870.

Yockteng, R., A. M. R. Almeida, S. Yee, T. Andre, C. Hill, and C. D. Specht. 2013. A method for extracting high-quality RNA from diverse plants for next-generation sequencing and gene expression analyses. *Applications in Plant Sciences* 1(12): 1300070.

Zedane, L., C. Hong-Wa, J. Murienne, C. Jeziorski, B. G. Baldwin, and G. Besnard. 2016. Museomics illuminate the history of an extinct, paleoendemic plant lineage (*Hesperelaea*, Oleaceae) known from an 1875 collection from Guadalupe Island, Mexico. *Biological Journal of the Linnean Society* 117: 44–57.

Zhang, C., E. Sayyari, and S. Mirarab. 2017. ASTRAL-III: Increased scalability and impacts of contracting low support branches. *In* J. Meidanis and L. Nakhleh [eds.], Comparative genomics. RECOMB-CG 2017. Lecture Notes in Computer Science, vol. 10562, 53–75. Springer International Publishing, Cham, Switzerland.

Zhang, N., J. Wen, and E. A. Zimmer. 2015. Congruent deep relationships in the grape family (Vitaceae) based on sequences of chloroplast genomes and mitochondrial genes via genome skimming. *PLoS One* 10: e0144701.

Zhou, Y., L. Duvaux, G. Ren, L. Zhang, O. Savolainen, and J. Liu. 2017. Importance of incomplete lineage sorting and introgression in the origin of shared genetic variation between two closely related pines with overlapping distributions. *Heredity* 118: 211–220.

