The LINNEAN SOCIETY of London BOTANICAL Journal of the Linnean Society

Botanical Journal of the Linnean Society, 2016, **182**, 234–255. With 2 figures

An introduction to plant phylogenomics with a focus on palms

CRAIG F. BARRETT[1,*], CHRISTINE D. BACON[2], ALEXANDRE ANTONELLI[2,3], ÁNGELA CANO[4] and TOBIAS HOFMANN[2]

[1]*Division of Plant and Soil Sciences, West Virginia University, 100 Research Way G153 South Agricultural Sciences Building, Morgantown, 26506, WV, USA*
[2]*Department of Biological and Environmental Sciences, University of Gothenburg, Box 461, SE-405 30 Göteborg, Sweden*
[3]*Gothenburg Botanical Garden, Carl Skottsbergs gata 22A, SE-41319 Göteborg, Sweden*
[4]*Plant Systematics and Biodiversity Laboratory, University of Geneva and Conservatoire et Jardin botaniques de la ville de Genève, C.P. 60, 1292 Chambésy, Switzerland*

Received 13 October 2015; revised 26 January 2016; accepted for publication 2 February 2016

Phylogenomics refers to the use of phylogenetic trees to interpret gene function and genome evolution and to the use of genome-scale data to build phylogenetic trees. The field of phylogenomics has advanced rapidly in the past decade due to the now widespread availability of next generation sequencing technologies, which themselves continue to change at a rapid pace and drive down the cost of sequencing per base pair. In this review, we discuss genomic resources available to palm biologists in the form of complete genomes (plastid, mitochondrial, nuclear) and sequenced transcriptomes, all of which can be leveraged to study non-model palm taxa. We also discuss various approaches to generating phylogenomic data in palms, such as next-generation sequencing technologies and methodological approaches that allow acquisition of large volumes of biologically and phylogenetically meaningful data without the need to sequence entire genomes (e.g. genome skimming, RAD-seq, targeted sequence capture). This review was designed for those unfamiliar with phylogenomics and associated methods, but who are interested in engaging in phylogenomics research. We discuss several considerations required for designing phylogenetic projects using genomic data, such as available computing capabilities and level of bioinformatics expertise. We then review some recent, empirical examples of palm phylogenomic studies and how they are shaping the future of palm systematics and evolutionary biology. © 2016 The Linnean Society of London, *Botanical Journal of the Linnean Society*, 2016, **182**, 234–255

ADDITIONAL KEYWORDS: arecaceae – arecales – bioinformatics – genomics – high throughput sequencing – next generation sequencing – NGS – palmae.

INTRODUCTION: WHAT IS ‘PHYLOGENOMICS?’

The development and subsequent widespread availability of DNA sequencing technologies have twice revolutionized biology: with the development of Sanger sequencing (Sanger, Nicklen & Coulson, 1977) and next generation sequencing (NGS). This is especially true in the field of genetics, allowing for a new discipline called genomics. The term ‘phylogenomics’ (Eisen, 1998) emerged later, initially describing the use of phylogenetic trees to predict gene/genome function. However, genome-scale data can also be used to build highly resolved and supported phylogenetic trees, which in turn can be used to study genome function and evolution in a phylo-comparative framework. The field of systematics, which was traditionally based upon information from morphology, anatomy, behaviour, physiology and geography, likewise experienced a revolution with the introduction of Sanger sequencing (e.g. Hillis & Moritz, 1990; Soltis, Soltis & Doyle, 1992). The application of the polymerase chain reaction (PCR; Mullis *et al.*, 1986), cloning (Sambrook, Fritsch & Maniatis, 1989) and Sanger sequencing (see below) produced data that allowed hypotheses previously based on morphological

*Corresponding author: E-mail: cfb0001@mail.wvu.edu and other data sources to be tested, substantially increasing phylogenetic resolution and support. Molecular data have repeatedly provided evidence for major taxonomic rearrangements, thus greatly advancing our understanding of the Tree of Life.

More recently, as NGS technologies came into broad use (see below; Table 1), a more general definition of phylogenomics now dominates: the use of genome-scale data to build phylogenetic trees. Here we distinguish between two commonly used meanings: the first refers to using trees to study e.g. genomic function, gene family evolution, comparative genome evolution, and horizontal gene transfer; the second simply refers to the use of genome-scale data to build phylogenetic trees. Thus, the second is not mutually exclusive of the first; on the contrary, building phylogenetic trees is essentially a single but necessary component of phylogenomics (Eisen, 1998; Sjölander, 2004).

Recent phylogenomic examples include the studies of Szöllősi *et al.* (2015), who used phylogenetic trees to interpret genome-scale data on the frequency of horizontal gene transfer among groups of fungi, and Davies *et al.* (2015), who used transcriptome- and genome-based phylogenetic trees to study adaptive evolution in African mole rats as a result of a subterranean lifestyle (see below for a definition and overview on transcriptomes and their use in phylogenetics). An example in plants is that of Jiao *et al.* (2014), who used publicly available nuclear genomes to build a phylogenetic tree among representative clades of monocots for the inference of ancestral genome duplication events. This study identified several such events, including one uniting the commelinid monocots, a clade of immense ecological/economic importance and high diversity (including palms, gingers, grasses and their relatives).

In recent years, systematists have become increasingly interested in building phylogenetic trees using genomic data. This is certainly the case in plant systematics, as evidenced by growing numbers of references to phylogenomic studies on the Angiosperm Phylogeny Website (http://www.mobot.org/MOBOT/research/APweb; Stevens, 2001; onwards) and in published papers (Stevens & Davis, 2005; APG III, 2009). One recent example of such a study is the analysis of 360 complete plastid genomes (protein-coding regions) from public databases across the green plants (Ruhfel *et al.*, 2014), which at that time represented the largest plastid dataset yet analysed. The authors sampled comprehensively across the green plants, based on all publicly available complete plastid genome data, providing resolution and support for a great number of relationships, but also identifying areas of uncertainty. Another study (Wickett *et al.*, 2014) used transcriptome sequencing to generate a dataset of >1000 low/single copy nuclear loci across a sample of 92 representative green plant taxa and was able to improve resolution of some of the deepest but most recalcitrant nodes in the green plant tree of life with strong branch support. Furthermore, use of numerous, non-recombining loci of the nuclear genome avoids relying on the plastid genome, which represents a single, albeit powerful and informative history, and allows the use of additional phylogenetic approaches such as the multispecies coalescent (e.g. Ané *et al.*, 2007; Degnan & Rosenberg, 2009; Liu *et al.*, 2009a; Heled & Drummond, 2010).

This review will primarily focus on the more current use of phylogenomics (using genome-scale data to build trees). This is not to diminish or trivialize the ‘original’ meaning; indeed, as we will argue below, the interpretation of genomic data based on phylogenetic trees can be particularly powerful in systematics and evolutionary biology (e.g. comparative genomics and differential gene expression).

Palms (order Arecales, family Arecaceae) are a diverse group of ecologically and economically important monocot angiosperms comprising 2600 species in 181 genera (Baker & Dransfield, 2016; this volume), with a rich history of systematic studies dating back several centuries (see references in Uhl & Dransfield, 1987; Dransfield *et al.*, 2008; Baker & Dransfield, 2016; this volume). Earlier systematic work based on morphology was greatly advanced by the application of Sanger sequencing (e.g. Baker *et al.*, 1999, 2009, 2011; Lewis & Doyle, 2001; Roncal *et al.*, 2005; Asmussen *et al.*, 2006), resulting in phylogenetically informed tribal and subfamilial classification systems (Dransfield *et al.*, 2005; Asmussen *et al.*, 2006; Dransfield *et al.*, 2008; for a more detailed description of systematic advances in palms, refer to Baker & Dransfield, 2016). However, many areas of uncertainty remain in palm relationships, particularly at the genus and species level (e.g. resolution among genera in subtribes of Trachycarpeae; Bacon, Baker & Simmons, 2012). Most uncertainties are due to the need for greater numbers of informative phylogenetic markers. Furthermore, palms have been shown to have extremely slow substitution rates compared to other monocot clades (e.g. plastid RuBisCO large subunit, nuclear alcohol dehydrogenase; Bousquet *et al.*, 1992; Gaut *et al.*, 1992; Barrett *et al.*, 2015). Thus, the use of genome-scale data has the potential to resolve difficult issues in palm systematics (Baker & Dransfield, 2016) and will furthermore lay a phylogenetically robust foundation for systematically relevant studies of genomics and macroevolution and other fields.

An exhaustive treatment of all topics relevant to phylogenomics is not possible in a single review and

Table 1. Definitions of frequently encountered terminology associated with phylogeonomics. Note that these techniques are not mutually exclusive, are often loosely defined and may be used differently in different fields

Sanger sequencing (here, first-generation sequencing): A method of DNA sequencing that involves chain termination using dideoxynucleotides, usually resulting in sequence reads of 500–1000 bp. This method requires cloning or amplification via polymerase chain reaction (PCR) to provide sufficient DNA template for sequencing. This was the first widely available sequencing technology that drove the 'molecular revolution' and resulted in numerous single or multigene phylogenies and the first complete genomes (e.g. the first human genome).

Next generation sequencing (NGS): A broad class of technologies that became available in the 2000s, post-Sanger sequencing, allowing massively parallel sequencing of DNA or RNA. Examples include pyrosequencing (454), sequencing by ligation (SOLiD), single molecule fluorescence sequencing (Helicos), sequencing by synthesis (Illumina), single molecule real-time sequencing (Pacific Biosciences), nanopore sequencing (Oxford Nanopore) etc.

Second generation sequencing: A class of sequencing technologies that includes pyrosequencing, sequencing by ligation, single molecule fluorescence sequencing, sequencing by synthesis etc. This classification is somewhat arbitrary, but usually implies the class of sequencing technologies that were (or are still) widely available from the mid 2000s to the present day and that usually result in reads < 1,000 bp in length, e.g. Illumina, 454.

Third generation sequencing: A class of technologies including single molecule real time sequencing (SMRT: Pacific Biosciences, or 'PacBio') and nanopore sequencing (Oxford Nanopore). These are distinct from second generation technologies, and produce read lengths >1,000 bp (and potentially much longer). High sequencing errors remain a drawback for some technologies.

Whole genome shotgun sequencing: A method of randomly sequencing the genome that involves either cloning or fragmenting genomic DNA, sequencing using one of a variety of technologies (e.g. Sanger, Illumina, PacBio) and assembling the reads to cover the genome at some depth.

Genome skimming (genome survey sequencing, shallow sequencing): Whole genome shotgun sequencing at levels typically far too low to recover the single or low-copy elements of the nuclear genome, but enabling recovery of the 'high-copy fraction' of genomic DNA. In plants, this includes plastid genomes, mitochondrial genes or genomes, rDNA cistrons, transposable elements and other high-copy elements. Several samples may be tagged with unique barcodes, pooled, sequenced in multiplex and sorted bioinformatically, greatly increasing cost-effectiveness.

Library preparation: Laboratory procedures necessary to prepare samples for NGS. This usually includes fragmentation of genomic DNA, followed by size-selection of fragments and ligation of sequencing adapters, primers and barcode indexes.

Paired-end sequencing: In NGS, sequencing both ends of a genomic DNA fragment, as opposed to only one end (single end sequencing); e.g. 100 bp paired-end sequencing of ~500 bp fragments yields 100 bp of sequence data on each end of the fragment with 300 bp of unknown sequence between them.

Coverage depth: The number of bases covering a particular position of the genome; e.g. 100 × coverage depth means that there is an average of 100 bases contributing to the consensus sequence of each position across a genome.

Genome coverage: Percentage or proportion of the genome that is covered by sequence data, based on some coverage depth criterion; e.g. 95% of the genome is covered at a depth of 100 bp or more.

Transcriptome: All of the expressed messenger RNA (mRNA) transcripts from a given tissue or tissues at a given point in development or during some particular stage of a physiological or developmental process.

RNA sequencing (RNA-seq): Shotgun sequencing of total RNA from a transcriptome.

Target capture, sequence capture, hybrid sequence capture, hyb-seq, seq-cap etc. A class of methods by which specific, predetermined regions of the genome are captured via DNA or RNA probe hybridization and sequenced using various technologies (Illumina, Sanger, 454 etc.).

Reduced representation: A broad category of techniques by which a particular subset of loci are selected from across the genome (either randomly or non-randomly) that are particularly informative for the question at hand, greatly reducing the complexity of whole-genome analyses. Examples include restriction site-associated sequencing (RAD-seq), genotyping-by-sequencing (GBS), targeted sequence capture etc.

Metagenomics: Sequencing of all environmental or clinical DNA or RNA at a given location or from a particular specimen, allowing both identification and functional characterization of (usually microbial) communities; e.g. plant root-associated microbial communities, human gut communities, water samples). This is not exactly equivalent to **meta-barcoding** (a form of DNA barcoding), which instead involves amplicon sequencing of a single gene for molecular identification, e.g. through ribosomal DNA amplification and sequencing.

thus some topics are necessarily beyond the scope of this paper. Here we briefly summarize the history of commonly used sequencing technologies and discuss methodological approaches currently available for generating genome-scale phylogenetic trees. Lastly, we review some recent phylogenomic analyses of palms at multiple taxonomic levels, including different methodological approaches, provide suggestions for palm biologists interested in using phylogenomic tools and propose ideas for the immediate future of palm phylogenomic research.

A BRIEF TIMELINE OF COMMONLY USED SEQUENCING TECHNOLOGIES

The earliest forms of DNA sequencing became available to a general audience of researchers in the late 1970s (Maxam & Gilbert, 1977; Sanger *et al.*, 1977). Sanger sequencing, which uses dideoxy chain termination (Table 1), became the dominant mode of sequencing in the 1980s until after the turn of the millennium. With the addition of PCR, specific regions of the genome could be amplified to produce sufficient material for generating sequences on the order of a few hundred bases to a few kilobases in length. The earliest complete genomes of viruses, prokaryotes, single-celled eukaryotes, fungi, plants and animals (including humans) were sequenced with the Sanger method via whole genome shotgun sequencing (Table 1); many sequencing projects employed sophisticated robotic machinery, took several years to complete, required massive consortia of researchers and cost enormous amounts of money based on the technology available (e.g. International Human Genome Sequencing Consortium, 2001).

In the mid-2000s, various technologies became available that allowed massively parallel sequencing of genomic DNA and RNA, the so-called ‘second generation sequencing technologies’ (Table 1; reviewed in Glenn, 2011; Mardis, 2013; van Dijk *et al.*, 2014). Some prominent examples include pyrosequencing (Roche 454 Life Sciences, Branford, CT, USA), sequencing by synthesis (Solexa/Illumina Inc., San Diego, CA, USA), sequencing by ligation (i.e. SOLiD, Applied Biosystems/Life Technologies, Waltham, MA, USA) and single molecule fluorescent sequencing (Helicos BioSciences, Cambridge, MA, USA) (Fig. 1). The classification here of second generation sequencing technologies is somewhat arbitrary, but despite differing greatly in chemistry, they generally produce reads < 1,000 bp in length. These technologies decreased the cost of sequencing per base pair enormously compared to Sanger sequencing (for a comparison of technologies and cost per base pair, see Glenn, 2011, and subsequent updates, e.g. http://www.molecularecologist.com/next-gen-fieldguide-2014; van Dijk *et al.*, 2014), thus precipitating a second

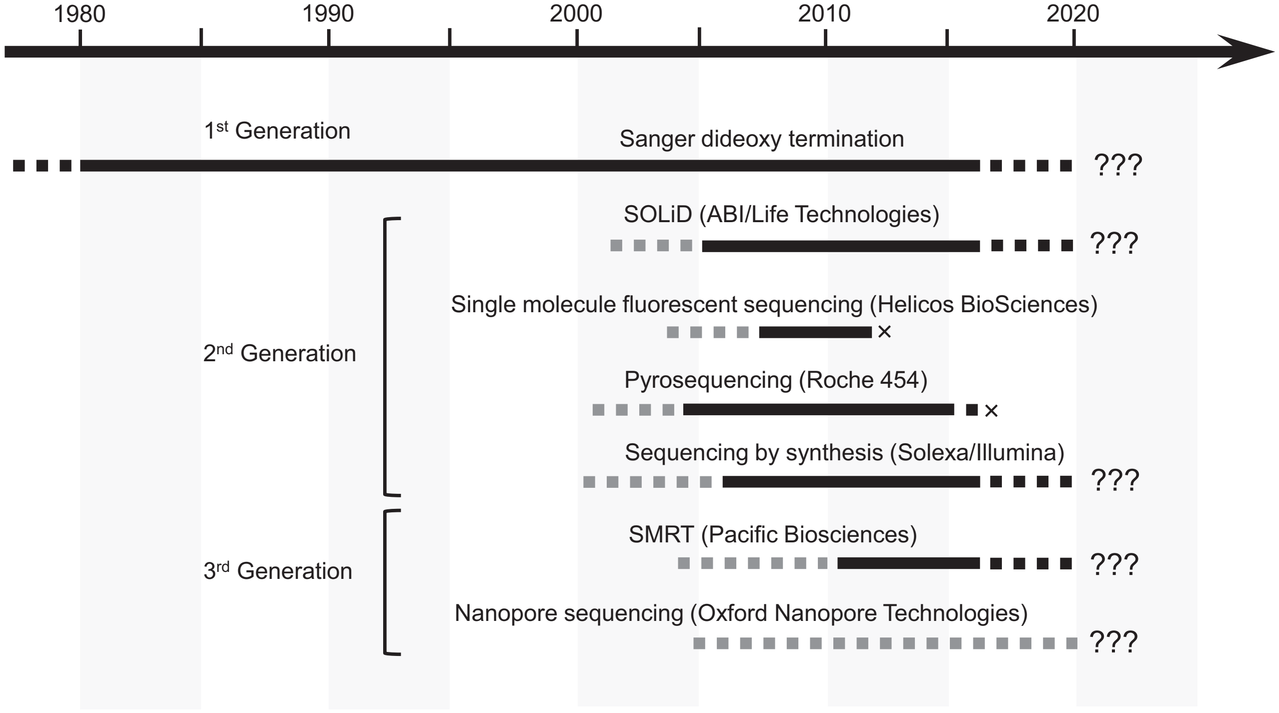

Figure 1. A brief timeline of commonly used DNA/RNA sequencing technologies. Arrow above = year; dotted grey line = approximate time in development; black lines = dates when openly available to researchers up to the present; black dotted lines and question marks = uncertainty of availability in the future; ‘×’ = discontinuation or lack of support (either in the past or planned for the future); ‘SOLiD’ = Sequencing by Oligo/Ligation and Detection; ‘ABI’ = Applied Biosystems; ‘SMRT’ = Single-Molecule Real Time.

revolution in genomics. Genomes that took years to sequence with Sanger technology could now be completed in days, for a miniscule fraction of the cost. Figure 1 details a timeline of some of the more commonly employed sequencing technologies. For a review of the technical details and advantages/disadvantages of each second generation sequencing technology, refer to Mardis (2013) and van Dijk *et al.* (2014). Currently, Illumina technologies dominate the global sequencing market, due to lower cost and higher throughput relative to other second generation technologies (van Dijk *et al.*, 2014).

The shorter read lengths of second generation technologies make assembly into complete genomes rather difficult, requiring advanced skills in programming and bioinformatics. Other technologies have been or are being developed (Fig. 1) that provide longer read lengths and potentially better genome assemblies. Single molecule real time sequencing (SMRT, Pacific Biosciences of California, Menlo Park, CA, USA) is one platform currently available that produces reads up to 60 kb and it is particularly useful in whole genome sequencing for creating scaffolds to cross regions that cannot be resolved with the shorter read lengths of second generation technologies (e.g. long repeats or low complexity regions that are common among genomes). These ‘third generation technologies’ (Table 1) tend to have high error rates (but see below regarding SMRT sequencing), despite their desirable long read lengths (Schadt, Turner & Kasarskis, 2010). Thus, a common strategy has been to combine the higher output and sequence accuracy/depth of second generation sequencing with lower output but longer read lengths of third generation sequencing to achieve both deep coverage and longer assembled contigs (‘hybrid’ assembly of Illumina and PacBio data; e.g. Bashir *et al.*, 2012).

Improvements to SMRT sequencing have led to increased read lengths and lower overall error rates; it should be noted that the error distribution along a read is random for SMRT sequencing, as opposed to that of Illumina, for which quality tends to decrease toward the 3′ portions of reads. The use of ‘circular consensus sequencing’ allows multiple interrogations of base calls at a given position and accurate allelic phasing within fragments (Eid *et al.*, 2009; Travers *et al.*, 2010). Thus with deeper coverage, the error profile of SMRT sequencing becomes minimal; numerous recent studies have employed this technology alone to assemble high-quality genomes and full-length transcripts (Chin *et al.*, 2013; Koren & Phillippy, 2015; Pendleton *et al.*, 2015; Westbrook *et al.*, 2015).

An exciting third generation technology involves nanopore sequencing (Oxford Nanopore), with the potential to produce reads > 100 kb, but this technology is not yet widely available. Such long read lengths would make genome assemblies much easier and of higher finished quality, but initial tests of this technology suggest high sequencing error rates (Laver *et al.*, 2015); hopefully improvements in the technology can be made that will allow for high throughput, real-time analyses of long reads with low error rates.

GENOMIC RESOURCES FOR THE PALMS

The widespread availability of NGS technology means that plant biologists no longer need to rely exclusively on model systems [e.g. *Arabidopsis* Heynh. (Brassicaceae), *Oryza* L.*, Zea* L. (Poaceae)], but have the ability to build their own models (e.g. the milkweed *Asclepias syriaca* L.; Straub *et al.*, 2011). Palms are no exception and genomic resources for the family have been rapidly accumulating (Fig. 2). In 2010, the first complete plastid genome (= plastome) of a palm was published (date palm, *Phoenix dactylifera* L.; Yang *et al.*, 2010), followed by plastomes of African oil palm (*Elaeis guineensis* Jacq.; Uthaipaisanwong *et al.*, 2012) and coconut (*Cocos nucifera* L.; Huang, Matzke & Matzke, 2013). The first complete palm mitochondrial genome was published in 2012 (*Phoenix dactylifera*; Fang *et al.*, 2012). More recently, a number of additional plastid genomes have been sequenced: Barrett *et al.* (2015) and Comer *et al.* (2015) published a combined 68 plastomes representing all five subfamilies and nearly all tribes across the palms (Fig. 2). Assemblies for additional mitochondrial genomes are currently underway (C.F. Barrett, unpubl. data).

The publication of two annotated nuclear genomes in 2013 represented a milestone in palm biology (*Phoenix dactylifera*, Al-Mssallem *et al.*, 2013; *Elaeis guineensis*, Singh *et al.*, 2013). These are economically important species for human nutrition and have been sequenced mainly for the purpose of understanding the genetic underpinnings of, for example, fruit development and improvement. They also represent important annotated references for the sequencing of a great diversity of other palm genomes in the future and have been/will continue to be especially crucial in comparative palm phylogenomics. Current nuclear genome sequencing efforts are targeting additional palms [e.g. *Chamaedorea tepejilote* Liebm. (subfamily Arecoideae), J. Tregear, unpubl. data; *Geonoma undata* Klotzsch (Arecoideae), C. Lexer, unpubl. data; *Mauritia flexuosa* L.f. (Calamoideae), Tregear *et al.*, unpubl. data]. In addition to organellar and nuclear genomic sequencing, several transcriptome datasets generated via RNA-seq (defined in Table 1) are available, from representative species across three of the five palm subfamilies (Bourgis *et al.*, 2011; Matasci *et al.*,

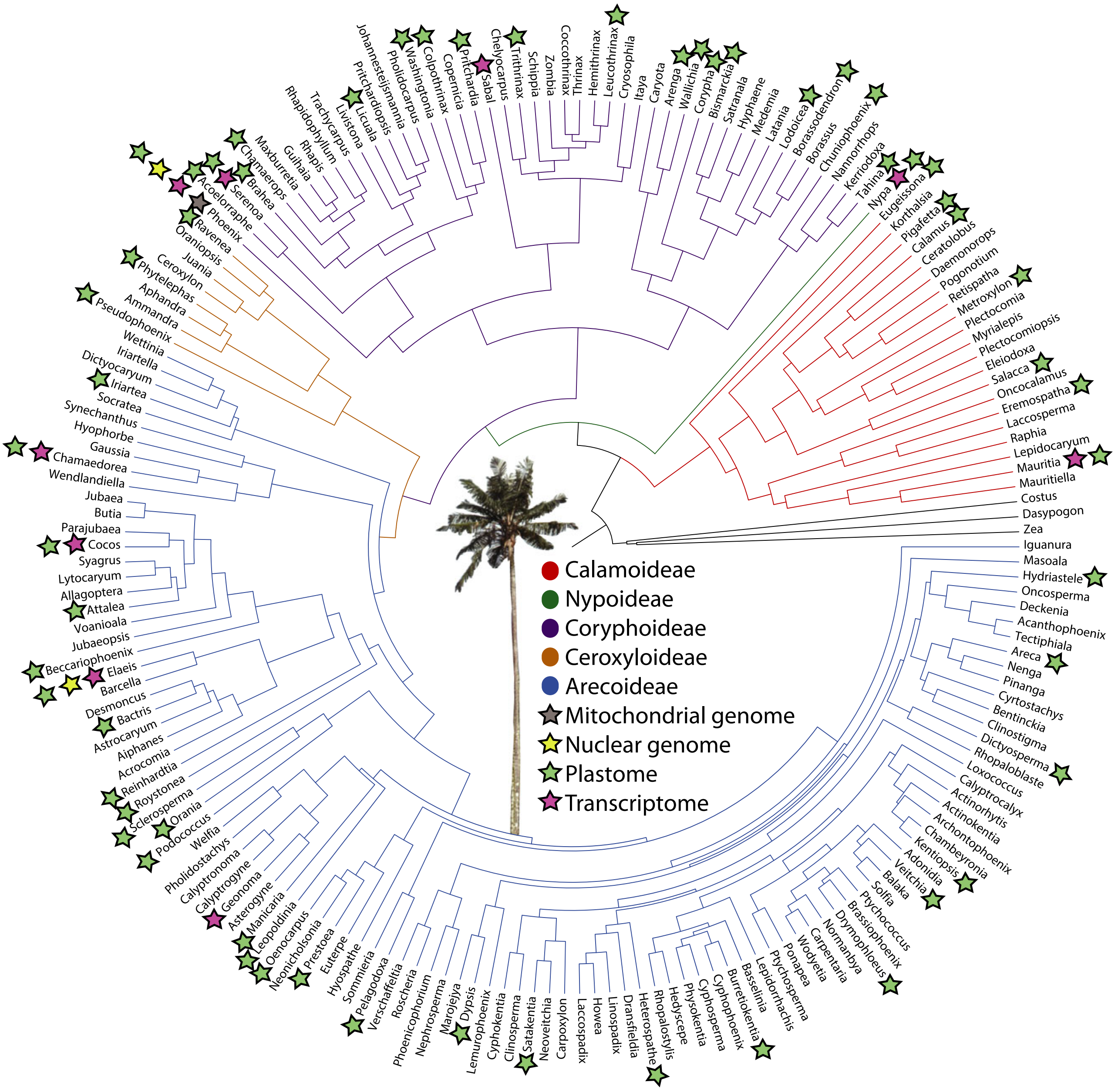

Figure 2. Phylogenetic tree of the palms (Arecaceae) with generic-level sampling based on the analysis of Couvreur, Forest & Baker (2011) showing each of the five palm subfamilies indicated with colours. Palm genera with genomic resources (complete plastomes, mitochondrial genomes, nuclear genomes and/or transcriptomes) are also shown with coloured star symbols indicating available data.

2014). These include: oil palm, coconut (both Arecoideae), and date palm; *Nypa fruticans* Wurmb (Nypoideae); *Serenoa repens* (W. Bartram) Small and *Sabal bermudana* L.H. Bailey (subfamily Coryphoideae); available at https://sites.google.com/a/ualberta.ca/onekp. Several more are to be released by the US National Science Foundation-funded 'Assembling the Monocot Tree of Life' project (http://www.botany.wisc.edu/givnish/monocotatol.htm).

METHODOLOGICAL APPROACHES TO GENERATING PHYLOGENOMIC DATA

Whole genome sequencing/re-sequencing (Wgs)

This describes methods and approaches for generating complete or nearly complete, annotated nuclear genomes that can be used for myriad purposes, including comparative phylogenomics (e.g. Nater *et al.*, 2015). Next generation sequencing technologies have made genome sequencing an attainable objective for individual labs or small collaborations, whereas before they required expensive, large-scale efforts of massive consortia (e.g. *Arabidopsis* Genome Initiative, 2000; Schnable *et al.*, 2009). Although generating genome-scale data is becoming much more affordable and efficient, assembling those data into finished, fully annotated genomes is complicated, with the need for high-performance, parallel computing and a high level of bioinformatics expertise (e.g. Schatz, Witkowski & McCombie, 2012). It is the latter consideration that perpetuates the slow rate at which new genomes are assembled, annotated, published and analysed. Although the analysis of complete nuclear genomes to produce highly resolved phylogenetic trees is a major goal of phylogenomics, it is also the most financially, computationally and technically demanding approach.

Assembling whole nuclear genomes usually requires paired end NGS data (see Table 1), and it is beneficial to use different fragment sizes. This is in order to achieve deep coverage with libraries of smaller fragment size while also using libraries with longer fragment sizes at a lower sequencing depth to serve as bridges between contigs, i.e. ‘scaffolds’. Paired end libraries with larger fragment sizes (usually > 1000 bp) are particularly crucial in genome assembly owing to their ability to cross problematic repetitive regions and to create contig scaffolds when the ends of a read are anchored in two assembled contigs. For an in-depth review of whole genome sequencing, assembly and annotation, refer to Ekblom & Jochen (2014). However, in many cases researchers will not need whole nuclear genomes to address significant questions in phylogenomics. Various techniques are discussed below (see Table 2) that reduce the vast volume of genomic data through which biologists must sort, but that still allow recovery of rich genomic information.

Genome skimming

A basic, relatively straightforward and rapid way to generate genome-scale data is via ‘genome skimming’ (e.g. Straub *et al.*, 2011, 2012; Papadopoulou, Taberlet & Zinger, 2015) or ‘genome survey sequencing’ (e.g. Steele *et al.*, 2012). This method takes advantage of the high-copy fraction of total genomic DNA, including organellar DNA (plastomes, mitochondrial genomes) and other multi-copy elements such as the nuclear ribosomal DNA (rDNA) cistron, transposable elements, some multigene families etc. This approach is possible because there are multiple organelles per cell (plastids furthermore contain multiple copies of the plastid genome) and of other high-copy elements in the nuclear genome. Thus, as opposed to sequencing one individual organism at deep levels, sequencing many samples at levels that result in low-coverage depth (defined in Table 1) of the majority of the nuclear genome (often <1 × coverage depth) will still yield relatively high coverage of organellar genomes and other high-copy elements of the nuclear genome.

Several samples (e.g. species, individuals from a population) can be multiplexed by adding unique sequence indexes, pooling and sequencing simultaneously, resulting in a cost-effective approach to generating phylogenomic data. Nuclear ribosomal DNA regions often have extremely high coverage (due to their repetitive nature in eukaryotic genomes), followed by plastid genomes and then mitochondrial genomes, respectively (Straub *et al.*, 2011, 2012; Barrett *et al.*, unpubl. data). It is often possible to

Table 2. A relative comparison of common methods used to generate phylogenomic data. For ease of comparison, a sample across 40 species (or individuals) is assumed. ‘Technical requirements’ refers to the level of sophistication required in terms of laboratory equipment and expertise; ‘Bioinformatics’ refers to the level of computational expertise and computational capacity (e.g. RAM) required; ‘Wasted data’ provides an approximate idea of what proportion of the final data are generally unusable for phylogenomic analysis

	Cost	Technical requirements	Bioinformatics	Wasted data
Whole genome sequencing	extremely high*	extremely high	extremely high	little
Genome skimming	low	low	low	vast majority
Transcriptomes (phylogenetics)	high	moderate	high	majority
Transcriptomes (gene expression)	very high	moderate	high	little
Sequence capture	moderate–high**	high	high	little

*However, cost per base pair continues to decrease at a rapid rate as technological improvements are made (e.g. Illumina).

**Some protocols for sequence capture require equipment such as hybridization ovens, quantitative PCR machines, sonication machines etc., which may require a large initial investment.

assemble complete rDNA cistrons, complete plastid genomes and partial to complete mitochondrial genomes or gene sets, although complete mitochondrial genomes are often difficult to assemble due to complex rearrangements, repeats and alternative structural configurations (e.g. Alverson *et al.*, 2011). Genome skimming is an excellent way for researchers unfamiliar with NGS technologies and analyses to gain experience in genomics and bioinformatics, while producing a highly resolved phylogenetic hypothesis for their clade of interest based on data from all three genomic compartments. The major disadvantage is that genome skimming does not give adequate coverage of the vast majority of phylogenomically relevant data housed in the nuclear genome, the low/single copy fraction.

Metagenomics

Plant-associated microbes are often difficult or impossible to culture, introducing biases into comparisons of microbial communities among plant species, habitats, experimental treatments etc. NGS technologies allow a work-around: metagenomics. This is the sequencing of environmental DNA or RNA samples (both intra- and extra-cellular) which allows inherent biases and roadblocks to culturing to be overcome. Most current studies using these techniques are focused on water or soil samples (Venter *et al.*, 2004; Daniel, 2005; Ramirez *et al.*, 2014; Delmont *et al.*, 2015). Universal PCR primers have been typically used to amplify ribosomal DNA from a broad group of interest (e.g. bacteria, fungi) followed by extensive cloning and Sanger sequencing to characterize microbial communities. More recently, similar amplicon-sequencing approaches have been undertaken using NGS of amplified rDNA to achieve much deeper and more cost-effective sampling of microbial communities (e.g. 454 or Illumina sequencing; e.g. Kembel *et al.*, 2014), which eliminates the requirement of costly and laborious cloning. Although reliance on single markers such as rDNA is useful and convenient in microbial identification, it does not provide detailed information on functional aspects of, for example, endophytic or rhizosphere microbial communities and whether or not these display phylogenetic structure with respect to their hosts. Particular care needs to be given, however, to taxonomic biases introduced by amplification steps (e.g. markers capturing only certain taxa and skewing the true environmental diversity) and contamination risks. The sequencing of environmental DNA or RNA can be extremely useful to those specifically interested in how plant-associated microbial communities differ, both taxonomically and functionally (Gilbert & Hughes, 2011; Carvalhais *et al.*, 2012). In other words, researchers can not only identify which microbes are present but also identify what genes are present and/or expressed under certain conditions and across associated plant species/communities.

Transcriptomics

Using transcriptomes gives a snapshot of gene expression, spatially across tissue types and temporally across developmental stages. Researchers interested in generating phylogenetic data from across the genome have benefitted immensely from using RNA-seq (Table 1) to acquire hundreds to thousands of markers for comparative evolutionary studies (e.g. Bazinet *et al.*, 2013; Wickett *et al.*, 2014; González *et al.*, 2015). Here we describe two types of transcriptomic analyses, in accordance to the two aforementioned definitions of phylogenomics: generating phylogenomic markers and studying comparative genome evolution or gene expression.

Using transcriptomes to identify markers useful in phylogenetic analysis has advantages and challenges. It allows the acquisition of massive amounts of biologically and phylogenetically informative data for systematics at multiple taxonomic levels (protein coding genes: expressed exons only) and excludes a major portion of the genome that may not be particularly useful in some studies (introns, intergenic spacers, repetitive DNA), thus simplifying the computational burden on researchers. Compared to genome skimming, one can generate enormous amounts of data from across the vast nuclear genome, with potentially higher information content, allowing for powerful approaches to phylogenetic inference, including multispecies coalescent analyses (e.g. Liu *et al.*, 2009a,b; Heled & Drummond, 2010). Transcriptomes can be indexed, pooled and sequenced in multiplex, but it is often necessary to sequence fewer samples per run to obtain sufficient coverage, relative to genome skimming. In other words, read data must cover only ~150 kb to obtain complete plastomes, whereas to obtain all expressed coding regions of a transcriptome, one needs a larger number of total reads per sample to obtain adequate coverage. Challenges include having to spend more money to obtain data relative to genome skimming (given the same number of taxa to be analysed) and difficulties with handling RNA, which is highly unstable and degrades rapidly, especially in remote, tropical field conditions as is often the case for palm research. Use of preservatives like RNAlater (ThermoFisher Scientific, Waltham, MA, USA) removes the need for taking liquid nitrogen into the field (which was previously prohibitive or at least extremely difficult and costly). However, in our experience such preservatives can yield a considerably lower quality and quantity of final RNA in some taxa as compared to utilizing fresh material. Thick, succulent, waxy or lignified tissues must be cut into small pieces to ensure adequate penetration of the preservative and it is advantageous to keep samples as cool as possible until they can be frozen at −80 °C or extracted.

Functional phylotranscriptomics requires the use of RNA-seq data to study differences in gene expression among taxa, natural conditions, tissues or experimental manipulations. Aside from whole-genome comparisons, this is perhaps the most powerful approach to phylogenomics, as it can give information about functional and/or adaptive variation, returning to the primary definition of phylogenomics. Not only does it potentially generate thousands of informative markers, but it also allows one to explore regions of the genome that may be directly involved with reproductive or ecological isolation among species or populations or that show evidence of adaptive, divergent evolution. It is also the most expensive of the approaches aside from whole genome sequencing, in that it requires extensive replication, which comes in two forms: biological replication, which is often focused on including multiple individuals across species or populations of a species; and technical replication, which involves using several libraries from the same individual to reduce experimental error (Dunn, Luo & Wu, 2013). Comparison of single-replicate samples may be informative as to the presence/absence of specific genes or gene families (all else being equally controlled), but this basic approach lacks any statistical power in that it provides no measure of internal or stochastic variation in differential gene expression. Replication allows the application of powerful statistical tests for comparing gene expression profiles across species, samples, experimental variables etc.

Reduced representation library sequencing

This represents a broad class of cost-effective methods to obtain genome-scale data from non-model organisms across the tree of life (e.g. Lemmon & Lemmon, 2013). Examples of these methods include reduced-representation shotgun sequencing (RRSS; Altshuler *et al.*, 2000), genotyping-by-sequencing (GBS; Elshire *et al.*, 2011) and restriction-site-associated DNA sequencing (RAD-seq; Miller *et al.*, 2007; Baird *et al.*, 2008). To reduce the volume of genomic data recovered, taking an example from RAD-seq, DNA samples are digested with one or more restriction enzymes. The resulting fragments are then size-selected to recover a subset of DNA fragments that are subsequently used for library preparation and sequencing. Although originally implemented for single nucleotide polymorphism (SNP) discovery and genotyping for genetic mapping and population genetics (e.g. Hohenlohe *et al.*, 2010, 2011, 2013), these methods are now also being used for shallow-level phylogenetics (e.g. Eaton & Ree, 2013; Cruaud *et al.*, 2014; Pante *et al.*, 2015) and phylogeography (e.g. Emerson *et al.*, 2010; Reitzel *et al.*, 2013; Leaché *et al.*, 2015b). The result can be tens of thousands of variable genetic markers from across the genome (including all three genomes in plants). A fundamental limitation of these methods concerns base substitutions within restriction sites across taxa. Mutations at these sites will cause the number of sequenced loci to drop at deeper phylogenetic scales and may reduce the usefulness of the resulting data in phylogenetic analyses, since there may be little overlap in orthologous sequences among different studies or across divergent taxa (e.g. Rubin, Ree & Moreau, 2012; Cariou, Duret & Charlat, 2013). Despite this, advances have been made towards accounting for and explicitly addressing these problems (Peterson *et al.*, 2012; Leaché *et al.*, 2015a). RAD-seq and related methods are particularly useful in conservation genomics, in that they allow genome-wide assessments of genetic variation within and among populations for many individuals and, more importantly, can lead researchers to regions of the genome that may display signatures of locally adapted variation (e.g. Hohenlohe *et al.*, 2010).

Targeted sequence capture

This has become a preferred method in many taxonomic groups because of the power to identify hundreds of variable loci (e.g. Faircloth *et al.*, 2012; Lemmon, Emme & Lemmon, 2012; McCormack *et al.*, 2012; Prum *et al.*, 2015). Instead of sequencing the whole genome, efforts are focused on those loci useful for a particular taxonomic scope, reducing the volume of unusable data recovered, increasing cost-effectiveness and, in many cases, providing high phylogenetic resolution and strong branch support for species trees.

Sequence capture specifically targets loci of interest and is combined with high-throughput sequencing methods, allowing the acquisition of hundreds to thousands of unlinked loci, distributed across the nuclear genome, at sufficient coverage depth. Because only specific regions of the genome are captured and sequenced, researchers can sample many more taxa or individuals relative to WGS, genome skimming and transcriptome sequencing, thereby wasting fewer data (e.g. Cronn *et al.*, 2012; Grover, Salmon & Wendel, 2012; Stull *et al.*, 2013; Weitemier *et al.*, 2014; Table 2).

To generate probes for targeted capture, sequences corresponding to orthologous loci across the taxon sample can be derived from genomic sequences such as annotated genomes and transcriptomes using a BLAST-based approach. Putative single-copy, orthologous loci can be parsed from these data and aligned, then filtered further to omit sequences that have low or extremely high pairwise distances (i.e. if they are invariant, have been converted to pseudogenes, contain low-complexity regions or repeats etc.), allowing for capture at various taxonomic scopes. Although there is inherent variation in the approach to generating probe sets, most methods are generally similar and there are automated pipelines available (e.g. Chamala *et al.*, 2015) and companies that synthesize nucleic acid probe sets at reasonable prices.

Probe design can be tailored to particular questions, whether the objective is to resolve relationships across deep or shallow evolutionary time scales. For example, one can choose to target conserved coding regions [e.g. ultra-conserved elements (UCEs); Faircloth *et al.*, 2012; McCormack *et al.*, 2012; Lemmon & Lemmon, 2013] at the same time as capturing high-variation introns. It may be difficult to assess homology among intron sequences at deeper phylogenetic scales and it may therefore be necessary to trim the intron sequences and thus only analyse coding regions. At intermediate taxonomic levels, researchers can usually include both exons and introns, to maximize the number of variable characters. One way to ensure a maximal coverage of flanking intron regions is to use paired-end sequencing, thereby extending coverage of the regions neighbouring the targeted exon. It is difficult, however, to completely recover longer introns (e.g. > 1000 bp), as capture success typically decreases as a function of physical distance from the exon (e.g. Bi *et al.*, 2012; Peñalba *et al.*, 2014). This approach does not absolutely require a nuclear genome, although this information helps in specifying introns of optimal length (often 100–600 bp; e.g. de Sousa *et al.,* 2014). For questions at or below the species level (e.g. population or conservation genetics), intron sequences may be used specifically to design probes; this often requires either a complete genome or several previously sequenced intron regions, but recent studies have been highly successful (e.g. Folk, Mandel & Freudenstein, 2015).

Once probes have been designed, they are then synthesized typically as RNA ‘baits’ (sequences to which the genomic samples will be hybridized). Sonication is then used to shear physically genomic DNA from all samples into small fragments (usually 200–800 bp, depending on the specific application), indexed as in other NGS approaches (i.e. a short, unique ‘barcode’ sequence is ligated to the fragment) and pooled at equimolar ratios across samples. The DNA libraries are then enriched for the selected loci by hybridizing genomic DNA libraries to the pre-synthesized probes/baits. Non-target fragments are removed and captured target fragments are eluted and typically enriched using PCR. Various protocols are available, including solution-based (e.g. Blumenstiel *et al.*, 2010) or array-based capture (e.g. Hodges *et al.*, 2009). The enriched library is then sequenced using, for example, an Illumina machine (e.g. Lemmon & Lemmon, 2013). Similar to the development of probe sequences, wet laboratory and bioinformatics protocols for targeted sequence capture are generally similar and available online for consultation (e.g. https://github.com/AntonelliLab/palm_pipeline).

BIOINFORMATICS AND COMPUTATIONAL CONSIDERATIONS FOR PHYLOGENOMICS

Although one can analyse NGS data on any operating system, many programs are written for a UNIX/Linux environment. Therefore, a basic familiarity with the command line interface is required to process read data and for subsequent analyses (e.g. UNIX is excellent for basic text file manipulation, scripting and programming), although some proprietary graphical user interfaces such as Sequencher (GeneCodes Corporation, AnnArbor, MI, USA), Geneious (Biomatters, Ltd., Auckland, New Zealand) and CLC Genomics (CLC bio, a QIAGEN Company, Venlo, Netherlands) are increasingly useful for NGS applications.

NGS data are typically returned from the sequencer to most researchers in FASTQ format (Cock *et al.*, 2010), in which each read consists of four lines; the most important lines are the sequence itself and a quality score for each base. Base calls towards the 3′ ends of reads tend to decrease in quality. Quality is typically described using the PHRED scale (Ewing & Green, 1998), which is a logarithmic expression of sequencing error probability. For example, a PHRED score of $10 = 10^{-1}$ or a probability of a sequencing error = 0.1; whereas PHRED $20 = 10^{-2}$ or 0.01 probability of an error; and so on, with a limit at PHRED = 40. There are numerous freely available programs and scripts [e.g. NGS QC Toolkit, (Patel & Jain, 2012); Trimmomatic (Bolger, Lohse & Usadel, 2014)] that allow filtering and/or trimming of poor quality reads or bases, according to user-defined PHRED thresholds, removal of any remaining adaptor contaminants, merging of overlapping paired-end reads, removal of non-unique reads etc. These are important steps that contribute to the quality and fidelity of assembled genes and genomes.

A variety of assembly programs exist for genome and transcriptome assembly (e.g. Zerbino & Birney, 2008; Simpson *et al.*, 2009; Li *et al.*, 2010; see Bradnam *et al.*, 2013). For in-depth reviews of genome assembly, see Haridas *et al.* (2011), Nagarajan & Pop (2013) and Ekblom & Jochen (2014). For target sequence capture data, the assembled contigs can be blasted against the reference sequences that were used for the bait design, in order to match contigs to known and identifiable genetic loci. When working with complete genome sequence data, it is necessary to annotate the retained contigs in order to identify their position in the genome. Annotating genomes can be laborious and time consuming, but represents one of the most important aspects of genomic research, in that it provides resources necessary for future bioinformatics analyses by other researchers globally. Methods of annotation can be complicated (in line with the complexity of eukaryotic genomes), although many tools exist to help researchers: e.g. DOGMA (Wyman, Jansen & Boore, 2004), ACRE (Wysocki *et al.*, 2014), and Verdant (McKain M & Hartsock R, unpubl. data) for plastomes; Mitofy for mitochondrial genomes (Alverson *et al.*, 2010), and a large suite of software available for nuclear genomes [e.g. The NCBI Eukaryotic Genome Annotation Pipeline, http://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/; Galaxy (Giardine *et al.*, 2005), available via web interface at usegalaxy.org; Genome Tools (Gremme, Steinbiss & Kurtz, 2013)]. Particular challenges emerge when dealing with polyploidy, which is common in many plant taxa, due to the difficulties in phasing allelic variation.

Building phylogenetic trees from genomic data also requires substantial RAM (random access memory); especially for multi-partition, model-based analyses. However, this is more closely related to the number of taxa than the number of characters per taxon (e.g. Day, 1983). As a rule, more complex models and genome-scale data require high performance computing (e.g. model-based coalescent analyses, divergence time estimation, phylo-comparative trait analyses). Workarounds have been developed that can greatly improve performance of phylogenomic data analysis (e.g. BEAGLE; Ayres *et al.*, 2012) and task parallelization in tree-building software such as RAxML (Stamatakis, 2006; Stamatakis & Aberer, 2013; Stamatakis, 2014), MrBayes (Huelsenbeck & Ronquist, 2001; Altekar *et al.*, 2004) and BEAST (Drummond & Rambaut, 2007; Drummond *et al.*, 2012), which help to divide the workload and speed up analyses. Some freely available online clusters can often handle modest to medium-sized analyses [e.g. CIPRES; Miller, Pfeiffer & Schwartz (2010), https://www.phylo.org] and many research-oriented institutions offer their own computational resources.

Despite these computational advances in parallelizing phylogenetic analyses, many challenges remain when facing large NGS datasets, which as of late routinely exceed 100 loci. For example, Bayesian methods (e.g. BEAST), which are a popular choice for phylogenetic inference, are pushed to their boundaries when analysing large, multi-locus datasets, as the underlying Markov chain Monte Carlo algorithms may take tens of millions of iterations (and weeks to months) to converge among replicate runs, if they converge at all. Methods that estimate gene trees and species trees jointly can provide a dynamic exploration of parameter space (e.g. Liu & Pearl, 2007; Liu, 2008). When adding a large number of loci to the analysis, the corresponding, introduced operators may lead to an overwhelming number of variables, for which it may not be possible to explore parameter space efficiently and thus may cause issues with parameter identifiability (see Ponciano *et al.*, 2012). Commonly chosen alternatives to Bayesian tree estimation are fast maximum likelihood (ML) methods such as RAxML, which are computationally more tractable when facing massive phylogenomic datasets. When using ML methods, gene trees and species trees have to be estimated sequentially. The common practice is to estimate a separate gene tree for each individual locus and use all gene trees to estimate a most likely species tree (e.g. BuCKy, MP-EST, ASTRAL; Ané *et al.*, 2007; Larget *et al.*, 2010; Liu, Yu & Edwards, 2010; Mirarab *et al.*, 2014). However, this approach may be problematic if many of the individual loci do not contain sufficient phylogenetic signal in order to estimate gene trees (discussed in Gatesy & Springer, 2014; Roch & Warnow, 2015). This becomes a challenge when analysing shallow phylogenies, where divergence times are relatively recent and not enough mutations have accumulated in order to infer robust gene trees, and also in deep phylogenies containing short internal branches.

In addition to practical limitations when dealing with vast genomic data for phylogenetic inference, one must take into account methodological aspects such as concatenation of data across loci versus the use of coalescent methods. This is currently a topic of intense debate (e.g. Song *et al.*, 2012; Wu *et al.*, 2013; Gatesy & Springer, 2014; Xi *et al.*, 2014; Simmons & Gatesy, 2015; Springer & Gatesy, 2016), which remains to be settled. Furthermore, despite the obvious benefits of obtaining data from hundreds to thousands of variable loci from across the genome, it may not yet be possible to use the full power of these data effectively due to the immense computational burden required to model robustly the complex evolutionary processes that have shaped these loci (e.g. Roch & Warnow, 2015). This is especially true for the inference of divergence times and species trees (among other data-intensive analyses), due to the sheer dimensions of these datasets in terms of taxa, loci, partitions and corresponding model parameter space. Another important aspect regarding tree inference based on genomic data is how to account for missing data (e.g. Streicher, Schulte & Wiens, 2016), not just with RAD-seq approaches as described above, but also with other approaches. Researchers must keep these considerations in mind when interpreting the results of phylogenomic analyses of relationships; improvements to phylogenetic reconstruction methods for genome-scale data represent an area of intense effort in computational biology.

SOME RECENT PHYLOGENOMIC STUDIES IN PALMS

Phylogenomics applied to palm systematics and evolution is a relatively new endeavour and has benefitted immensely from ongoing genome and transcriptome sequencing projects in date and oil palms and from widely available NGS technologies. These advances are having a profound effect on palm systematics. Up to this point, palm molecular systematics has relied heavily on a few loci from the plastid and nuclear genomes and the focus has been on increasing taxon sampling. Asmussen *et al.* (2006), Baker *et al.* (2009, 2011), Roncal *et al.* (2010, 2012) and Bacon *et al.* (2012) have provided recent examples of dense taxon sampling using data generated by Sanger sequencing across the palms at various levels (among and within subfamilies, tribes etc.). Although these and many other molecular studies based on Sanger sequencing have greatly improved our knowledge of palm taxonomy, biogeography and morphology, among other fields, there are still regions of the palm phylogenetic tree that remain unresolved and/or have low branch support, thus necessitating the vast character information contained in genome-scale data.

A recently published study by Barrett *et al.* (2015) used genome skimming to resolve ‘deep’ relationships among subfamilies and tribes of the non-arecoid palms (i.e. sampling was focused on the non-arecoid subfamilies Calamoideae, Coryphoideae, Nypoideae and Ceroxyloideae), and on the placement of the palm order (Arecales) among the commelinid monocots. Nearly all protein-coding regions of the plastome (75 genes), and whole aligned plastomes including intergenic spacers and introns, provided high resolution and strong support for nearly all nodes on the final tree. That study also recovered the same pattern of ‘deep’ relationships as seen in earlier studies: (Calamoideae, (Nypoideae, (Coryphoideae, (Ceroxyloideae, Arecoideae)))), all with 100% bootstrap support. Furthermore, tribal relationships outside Arecoideae were resolved with strong support, with the exception of Eugeissoneae in Calamoideae, the position of which remains unsupported. More broadly, the palms are placed with moderate to strong support as sister to the commelinid family Dasypogonaceae (boostrap support = 81–91, depending on the analysis), which consists of four genera native to Australia, unplaced to order. Thus, although most relationships were strongly supported, this and other studies have shown that genome-scale data from complete plastomes, or even in some cases from across nuclear genomes, do not *guarantee* complete resolution and support of all clades (e.g. Barrett *et al.*, 2013, 2014; Ruhfel *et al.*, 2014; Wickett *et al.*, 2014). Genomic data should be considered in the broader context of anatomy/morphology, fossils, development etc.

A striking finding from Barrett *et al.* (2015) was the extensive heterogeneity in plastome-wide substitution rates among palms and other commelinid orders. This represents the most comprehensive analysis of plastid substitution rates among the commelinid monocots, in terms of taxon and character sampling, and corroborates previous findings of slow evolutionary rates in the palms relative to other orders based on one or a few genes (e.g. Gaut *et al.*, 1992; but also see Scarcelli *et al.*, 2011). Based on nearly complete coding regions of the plastome, some lineages of Poales and Zingiberales display rates $> 5\times$ greater than those observed across a broad sample of palms (Barrett *et al.*, 2015). Although the causal factor(s) for this discrepancy in rates is not known, it is notable that palms contain most of the tallest species among all monocots, which may have contributed, at least in part, to their notoriously slow substitution rates (see discussion of plant height and substitution rates in Lanfear *et al.*, 2013). Future research efforts could include an expanded sampling of taxa for plastomes and many loci across the nuclear genome and available trait and environmental data could be used in phylo-comparative analyses of rates across monocots.

Another example of the use of genome skimming in palms is a preliminary study of the genus *Brahea* Mart. ex Endl., which is native to Mexico and Central America. Using data from the plastome, mitochondrial genome and nearly complete ribosomal DNA cistron, J.R. Medina, S.C. Lahmeyer, & C.F. Barrett, unpubl. data analysed 11 of 13 *Brahea* spp. to test subgeneric delimitation based on morphology (subgenera *Brahea* and *Erythea*, *sensu* Quero & Yáñez, 2000) and assessed the evolution of acaulescent growth forms across the genus. Plastomes provided resolution and support for relationships, but these differed slightly among the three genomes, suggesting that incomplete lineage sorting and possibly interspecific gene flow may be at work. Future comparisons may focus on sampling of multiple individuals of each described species across the geographical range of the genus and employing numerous nuclear loci to test species delimitation, detect gene flow and build a resolved species tree for the genus.

Comer *et al.* (2015) used a combination of genome skimming, long-range PCR, sequence capture and 454 + Illumina sequencing of plastid genomes to help resolve relationships among 31 representative species across the most species-rich palm subfamily, Arecoideae. Using this approach, they were able to resolve many of the deep relationships among tribes of Arecoideae with strong branch support, although some relationships among the 'core arecoid' clade remain unresolved based on protein-coding regions of the plastome. This study has important implications for the biogeographic history of Arecoideae, which contains over half of all palm species and has a pantropical distribution, and further demonstrates the effectiveness of capture-based approaches at higher taxonomic levels. Current efforts are focused on using sequence capture to generate a dataset of several hundred single copy nuclear loci (Comer *et al.*, in review) to help resolve deep-level relationships of this subfamily.

Sequence capture is being used across populations of the African genera *Podococcus* G. Mann & H. Wendl. (two species) and *Sclerosperma* Mann & H.Wendl. (three species) to target complete plastid genomes for > 100 individuals in a cost-effective manner (A. Faye, unpubl. data). Plastid probes were designed following the protocol published in Mariac *et al.* (2014) and Scarcelli *et al.* (2016). These data are now being analysed in combination with ecological climate models to test the presence of past tropical refugia and infer range dynamics of tropical forests along the Atlantic coast of Africa. This provides an example of how palms and NGS data are being used as a model to infer the evolutionary dynamics of tropical rainforests (e.g. Couvreur & Baker, 2013).

Heyduk *et al.* (2015) generated a probe kit for sequence capture for the study of taxonomic relationships and diversification in the American genus *Sabal* Adans. (Coryphoideae). One hundred and seventy-six loci were derived from targeted sequence capture, of which 133 were suitable for phylogenetic analysis, and well-supported relationships were resolved that largely reflect the geographical distributions of members of this genus. These results contrast in some areas with those from the plastome, which did not fully resolve species relationships, demonstrating the resolving power of low/single-copy loci across the nuclear genome. This paper also provides for the first time in palm systematics a glimpse of the high degree of topological conflict across loci of the nuclear genome at the species level and is informative in terms of systematically relevant processes such as incomplete lineage sorting and gene flow across species boundaries. Because of the sampling of genomic information used to design the probe kit, their approach is useful across multiple taxonomic levels in palms (e.g. across Arecoideae: J.R. Comer, unpubl. data) and therefore represents an immensely important genomic resource for palm systematists. Indeed, ongoing studies in Chamaedoreeae (A. Cano, unpubl. data), in the species complex *Geonoma macrostachys* Mart. (C.D. Bacon, unpubl. data), and phylogeographic study of *Mauritia flexuosa* (Bacon unpubl. data) show that the probe kit designed by Heyduk *et al.* (2015) is feasible and informative across taxonomic and temporal scales (Table 3).

GENERAL RECOMMENDATIONS FOR PALM BIOLOGISTS INTERESTED IN PHYLOGENOMICS

Sequencing technology and experimental approach

The most important piece of advice to palm researchers would be to let the questions decide the technology and not the other way around. In other words,

Table 3. Summary statistics related to the utility of targeted sequence capture across taxonomic levels in Arecaceae. Average number of raw reads, average number of recovered contigs and average number of genes represented in each dataset are shown

	Taxonomic level	Raw reads	Contigs	Genes
Chamaedoreeae*	Tribe	620 000	84 855	233
Sabal†	Inter-specific	166 208	11 588	157
Geonoma macrostachys‡	Intra-specific	1 737 895	43 026	227

*Bacon *et al.*, unpubl. data.
†Probe kit developed by Heyduk *et al.* (2015).
‡Cano *et al.*, unpubl. data.

technology is a secondary consideration after biologically meaningful questions are asked. Much of the laboratory work can be outsourced, removing the technical burden on the researcher, but the trade-off is that outsourcing can be more expensive. It may be cheaper to do everything in-house, but outsourcing may remove costly technical errors. If one chooses to carry out protocols in-house, it is advisable to collaborate with those who have technical expertise, equipment etc. or to visit a laboratory in which protocols are routinely practiced.

Learn a scripting language

A variety of 'scripting' languages exist that are freely available to help with the daunting task of processing and analysing overwhelming volumes of NGS data: Perl, Python, R, UNIX/Linux etc. These platforms are powerful and adaptable, especially in dealing quickly with massive text files, processing input/output data and integrating other software (i.e. forming pipelines). Various bioinformatics-based scripting initiatives exist with code customized for routine analyses of NGS data and for phylogenomics [Bioperl.org, Biopython.org, Biocunductor.org (R) etc.]. Collaborations with bioinformaticians or computer scientists who are interested in biological applications are also beneficial. However, collaborations of this nature should be *quid pro quo*, and collaborators should be considered for co-authorship on papers or included as co-principal investigators on grant proposals. An alternative is to hire a computer savvy student as a collaborator (e.g. an undergraduate majoring in computer science, bioinformatics or biological sciences with some coding experience). A number of tutorials are available on the web (e.g. YouTube.com, Lynda.com, code.org), as are online courses, books etc., to aid in coding and bioinformatics skills and these are especially useful for beginning and intermediate levels. To become proficient at coding, one needs to practice often, even just 15 min per day a few days a week; coding is analogous to learning a new language or musical instrument. Most researchers have extremely busy schedules and coding practice requires self-motivation. A way to ensure good practice is to organize a formal seminar group (or even informal meetings) or to partake in an intensive workshop to learn a scripting language. Coding should be viewed not as a skill to be learned overnight, but as a career-long commitment.

Data storage and computation

NGS data files are massive, often in the order of gigabytes per file. Data from just a few projects can easily take up terabytes of storage space and, when coupled with the need for backing up, this means a simple internal hard drive will not suffice. Data storage on servers, the cloud and external hard drives (as backups) is strongly recommended and should follow well-developed protocols and guidelines (e.g. Osborne *et al.*, 2014). Assembling reads into contigs and eventually genomes often requires a massive amount of RAM. Some analyses can be run on laptops, desktops or workstations, such as *de novo* assembly of genome skim data into plastomes (e.g. Barrett *et al.*, 2014, 2015). Larger analyses such as whole nuclear genome or transcriptome *de novo* assemblies may require much more memory (often > 500GB RAM) and highly parallelized computing.

Researchers should take advantage of any available high-performance computation at his/her home institution or should take the initiative to build his/her own system if funds are available to do so. New faculty or researchers, if given startup allowance, should allocate resources specifically for high-performance computing and storage, as should those applying for grant funding. For example, a workstation-style desktop computer with a RAM upgrade is currently sufficient for genome skim and many aspects of sequence capture bioinformatics. Many contemporary desktop computers can handle up to 32GB RAM and workstations can handle up to 64GB. Servers, clusters and cloud services can provide much greater storage capacity and RAM capability, with the added feature of allowing parallel computing, which is crucial for tasks such as whole genome assembly. Servers and clusters can be expensive for individual laboratories or researchers; it is often beneficial to pool resources with collaborators or colleagues at one's institution.

Organization

In order to carry out phylogenomic research, one needs to be highly organized not only in the laboratory, but in terms of data management (Noble, 2009). NGS data from just a few projects will quickly fill hard drives and the numerous files resulting from the various steps of bioinformatics analysis will quickly become overwhelming. It is strongly recommended to keep a hierarchically structured, clean file (directory) system, with directory and file names as detailed as possible. A systematic naming system of files is also recommended, to allow one to find specific files several months or even years after they are created. When not in use, files should be zipped (compressed) to save storage space and all data/results should be backed up regularly and systematically, ideally in more than one location.

THE FUTURE OF PALM PHYLOGENOMICS

We now have the tools to feasibly generate a species-level phylogenetic tree of all palms (or close to it) based on > 100 single-copy nuclear loci designed, for example, from probes used in Heyduk *et al.* (2015). The importance of a densely sampled, genome-scale, species-level phylogenetic hypothesis of all palms cannot be understated and is currently underway thanks to the collaborative efforts of many palm biologists, including ourselves and collaborators. Such a phylogenetic framework will allow for improved divergence time estimates, interpretation of morphology, development, ecology, macroevolution and genome evolution. A species-level phylogenetic tree for the palms will also help us explore other potential crop species as not to rely exclusively on the few that are in large-scale cultivation (i.e. date, oil palms). Lastly, palms have been recognized as a model for understanding tropical forest palaeoecology (Bacon, 2013; Couvreur & Baker, 2013) and a fully resolved, strongly-supported phylogenetic tree will allow for more accurate interpretation of the fossil record and its implications for the evolution of the Earth's most productive biome (tropical rainforests) through space and time.

These tools can also be used to study gene family evolution across the palm family and among more exclusive taxonomic levels. Improvements in sequencing technology, experimental protocols, analytical models and bioinformatics pipelines for data processing may allow the acquisition and use of genomic data that are typically discarded or not targeted (e.g. multi-gene families with numerous copies), but that probably contain a wealth of information relevant to comparative phylogenomics. For example, improvements to 3rd generation sequencing technologies may allow single-molecule sequencing and subsequent assembly of individual paralogous members of gene families without the need for cloning, whereas commonly used second generation shotgun approaches do not currently allow this. These methods are not only useful for phylogenomics at higher taxonomic levels, but also for species and population levels. Using sequence capture to target highly variable introns has major potential in population and conservation genetics, acquiring markers for genotype/phenotype association studies, finding loci under selection or that are locally adapted across species ranges and identifying markers important in plant breeding for desirable traits, among others.

Improvements on the one currently available probe kit (Heyduk *et al.*, 2015) for targeted sequence capture in palms can possibly be made by including new and unpublished palm genomes to increase the number of single-copy loci to 500 or more. An important but unexploited advantage to targeted capture in palm phylogeonomics is the ability to recover genome-scale data from herbarium specimens, taking advantage of the already degraded ('pre-sheared') DNA in these samples, as has been done in recent studies in 'museomics' (e.g. Staats *et al.*, 2013; Besnard *et al.*, 2014). This holds particular promise for species and populations that have become rare or endangered due to habitat degradation or even for extinct taxa (Miller *et al.*, 2009; Bi *et al.*, 2013; Zedane *et al.*, 2016) and will also allow a phylogenomic perspective on the effects of climate change on palm genetic diversity.

It is an exciting time to be a palm biologist, as sequencing technologies and analytical capabilities have made genomic approaches a reality. There is an inevitable and increasing role for phylogenomics in palm research in the coming years and we will continue to see the development of genomic resources for these economically and ecologically important, emblematic components of global tropical ecosystems.

ACKNOWLEDGEMENTS

C.F.B. was supported by National Science Foundation Award DEB-0830020 (Monocot Tree of Life Project), the California State University Program for Education and Research in Biotechnology, and the Alexander von Humboldt Institute. C.D.B., T.H. and A.A. were supported by the Swedish Research Council (B0569601), the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013, ERC Grant Agreement n. 331024), and a Wallenberg Academy Fellowship. We thank Karolina Heyduk, Jason Comer and Jim Leebens-Mack for expert technical and bioinformatics advice/assistance. We especially thank Julissa Roncal and Wolf Eiserhardt for their valuable insights into the manuscript and Rodrigo Bernal, Gloria Galeano, Jean-Cristophe Pintaud, James Tregear and Henrik Balslev for organizing the World Palm Symposium 2015 in Quindío, Colombia. *The authors fully disclose that there are no conflicts of interest with any of the for-profit institutions or companies mentioned in this review and the authors do not preferentially advocate the use of any one product over any other.

REFERENCES

Al-Mssallem IS, Hu S, Zhang X, Lin Q, Liu W, Tan J, Yu X, Liu J, Pan L, Zhang T, Yin Y, Xin C, Wu H, Zhang G, Ba Abdullah MM, Huang D, Fang Y, Alnakhli YO,

Jia S, Yin A, Alhuzimi EM, Alsaihati BA, Al-Owayyed SA, Zhao D, Zhang S, Al-Otaibi NA, Sun G, Majrashi MA, Li F, Tala Wang J, Yun Q, Alnassar NA, Wang L, Yang M, Al-Jelaify RF, Liu K, Gao S, Chen K, Alkhaldi SR, Liu G, Zhang M, Guo H, Yu J. 2013. Genome sequence of the date palm *Phoenix dactylifera* L. *Nature Communications* **4:** doi:10.1038/ncomms3274.

Altekar G, Dwarkadas S, Huelsenbeck JP, Ronquist F. 2004. Parallel metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics* **20:** 407–415.

Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, Lander ES. 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407:** 513–516.

Alverson AJ, Rice DW, Dickinson S, Barry K, Palmer JD. 2011. Origins and recombination of the bacterial-sized multichromosomal mitochondrial genome of cucumber. *Plant Cell* **23:** 2499–2513.

Alverson AJ, Wei X, Rice DW, Stern DB, Barry K, Palmer JD. 2010. Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). *Molecular Biology and Evolution* **27:** 1436–1448.

Ané C, Larget B, Baum DA, Smith SD, Rokas A. 2007. Bayesian estimation of concordance among gene trees. *Molecular Biology and Evolution* **24:** 412–426.

APG III. 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Botanical Journal of the Linnean Society* **161:** 105–121.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796–815.

Asmussen CB, Dransfield J, Deickmann V, Barfod AS, Pintaud JC, Baker WJ. 2006. A new subfamily classification of the palm family (Arecaceae): evidence from plastid DNA phylogeny. *Botanical Journal of the Linnean Society* **151:** 15–38.

Ayres DL, Darling A, Zwickl DJ, Beerli P, Holder MT, Lewis PO, Huelsenbeck JP, Ronquist F, Swofford DL, Cummings MP, Rambaut A, Suchard MA. 2012. BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. *Systematic Biology* **61:** 170–173.

Bacon CD. 2013. Biome evolution and biogeographical change through time. *Frontiers of Biogeography* **5:** 227–231.

Bacon CD, Baker WJ, Simmons MP. 2012. Miocene dispersal drives island radiations in the palm tribe Trachycarpeae (Arecaceae). *Systematic Biology* **61:** 426–442.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* **3:** e3376.

Baker WJ, Asmussen CB, Barrow SC, Dransfield J, Hedderson TA. 1999. A phylogenetic study of the palm family (Palmae) based on chloroplast DNA sequences from the *trnL–trnF* region. *Plant Systematics and Evolution* **219:** 111–126.

Baker WJ, Dransfield J. 2016. Beyond *Genera Palmarum*: progress and prospects in palm systematics. *Botanical Journal of the Linnean Society* doi:10.1111/boj.12401.

Baker WJ, Norup MV, Clarkson JJ, Couvreur TLP, Dowe JL, Lewis CE, Pintaud JC, Savolainen V, Wilmot T, Chase MW. 2011. Phylogenetic relationships among arecoid palms (Arecaceae: arecoideae). *Annals of Botany* **108:** 1417–1432.

Baker WJ, Savolainen V, Asmussen-Lange CB, Chase MW, Dransfield J, Forest F, Harley MM, Uhl NW, Wilkinson M. 2009. Complete generic-level phylogenetic analyses of palms (Arecaceae) with comparisons of supertree and supermatrix approaches. *Systematic Biology* **58:** 240–256.

Barrett CF, Comer JR, Leebens-Mack J, Li J, Mayfield-Jones DR, Medina J-R, Perez L, Pires JC, Santos C, Stevenson DW, Zomlefer WB, Davis JI. 2015. Plastid genomes reveal support for deep phylogenetic relationships and extensive rate variation among palms and other commelinid monocots. *New Phytologist* **209:** 855–870.

Barrett CF, Davis JI, Leebens-Mack J, Conran JG, Stevenson DW. 2013. Plastid genomes and deep relationships among the commelinid monocot angiosperms. *Cladistics* **29:** 65–87.

Barrett CF, Specht CD, Leebens-Mack J, Stevenson DW, Zomlefer WB, Davis JI. 2014. Resolving ancient radiations: can complete plastid gene sets elucidate deep relationships among the tropical gingers (Zingiberales)? *Annals of Botany* **113:** 119–133.

Bashir A, Klammer AA, Robins WP, Chin CS, Webster D, Paxinos E, Hsu D, Ashby M, Wang S, Peluso P, Sebra R, Sorenson J, Bullard J, Yen J, Valdovino M, Mollova E, Luong K, Lin S, Lamay B, Joshi A, Rowe L, Frace M, Tarr CL, Turnsek M, Davis BM, Kasarskis A, Mekalanos JJ, Waldor MK, Schadt EE. 2012. A hybrid approach for the automated finishing of bacterial genomes. *Nature Biotechnology* **30:** 701–707.

Bazinet AL, Cummings MP, Mitter KT, Mitter CW. 2013. Can RNA-Seq resolve the rapid radiation of advanced moths and butterflies (Hexapoda: Lepidoptera: *Apoditrysia*)? *An exploratory study. PLoS One* **8:** e82615.

Besnard G, Christin P-A, Malé P-JG, Lhuillier E, Lauzeral C, Coissac E, Vorontsova MS. 2014. From museums to genomics: old herbarium specimens shed light on a C3 to C4 transition. *Journal of Experimental Botany* **65:** 6711–6721.

Bi K, Linderoth T, Vanderpool D, Good JM, Nielsen R, Moritz C. 2013. Unlocking the vault: next-generation museum population genomics. *Molecular Ecology* **22:** 6018–6032.

Bi K, Vanderpool D, Singhal S, Linderoth T, Moritz C, Good JM. 2012. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics* **13:** 403.

Blumenstiel B, Cibulskis K, Fisher S, DeFelice M, Barry A, Fennell T, Abreu J, Minie B, Costello M,

Young G, Maquire J, Kernytsky A, Melnikov A, Rogov P, Gnirke A, Gabriel S. 2010. Targeted exon sequencing by in-solution hybrid selection. *Current Protocols in Human Genetics*. **18:** 18.4. doi: 10.1002/0471142905.hg1804s66.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30:** 2114–2120.

Bourgis F, Kilaru A, Cao X, Ngando-Ebongue GF, Drira N, Ohlrogge JB, Arondel V. 2011. Comparative transcriptome and metabolite analysis of oil palm and date palm mesocarp that differ dramatically in carbon partitioning. *Proceedings of the National Academy of Sciences of the United States of America* **108:** 12527–12532.

Bousquet J, Strauss SH, Doerksen AH, Price RA. 1992. Extensive variation in evolutionary rate of *rbcL* gene-sequences among seed plants. *Proceedings of the National Academy of Sciences of the United States of America* **89:** 7844–7848.

Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R, Chitsaz H, Chou WC, Corbeil J, Del Fabbro C, Docking TR, Durbin R, Earl D, Emrich S, Fedotov P, Fonseca NA, Ganapathy G, Gibbs RA, Gnerre S, Godzaridis E, Goldstein S, Haimel M, Hall G, Haussler D, Hiatt JB, Ho IY, Howard J, Hunt M, Jackman SD, Jaffe DB, Jarvis ED, Jiang H, Kazakov S, Kersey PJ, Kitzman JO, Knight JR, Koren S, Lam TW, Lavenier D, Laviolette F, Li YR, Li ZY, Liu BH, Liu Y, Luo R, MacCallum I, MacManes MD, Maillet N, Melnikov S, Naquin D, Ning Z, Otto TD, Paten B, Paulo OS, Phillippy AM, Pina-Martins F, Place M, Przybylski D, Qin X, Qu C, Ribeiro FJ, Richards S, Rokhsar DS, Ruby JG, Scalabrin S, Schatz MC, Schwartz DC, Sergushichev A, Sharpe T, Shaw TI, Shendure J, Shi YJ, Simpson JT, Song H, Tsarev F, Vezzi F, Vicedomini R, Vieira BM, Wang J, Worley KC, Yin SY, Yiu SM, Yuan JY, Zhang GJ, Zhang H, Zhou S, Korf IF. 2013. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* **2:** 10.

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

Carvalhais LC, Dennis PG, Tyson GW, Schenk PM. 2012. Application of metatranscriptomics to soil environments. *Journal of Microbiological Methods* **91:** 246–251.

Chamala S, Garcia N, Godden GT, Krishnakumar V, Jordon-Thaden IE, De Smet R, Barbazuk WB, Soltis DE, Soltis PS. 2015. Markerminer 1.0: a new application for phylogenetic marker development using angiosperm transcriptomes. *Applications in Plant Sciences* **3:** 1400115.

Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods* **10:** 563–569.

Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research* **38:** 1767–1771.

Comer JR, Zomlefer WB, Barrett CF, Davis JI, Stevenson DW, Heyduk K, Leebens-Mack JH. 2015. Resolving relationships within the palm subfamily Arecoideae (Arecaceae) using plastid sequences derived from next-generation sequencing. *American Journal of Botany* **102:** 888–899.

Couvreur TLP, Baker WJ. 2013. Tropical rain forest evolution: palms as a model group. *BMC Biology* **11:** 48.

Couvreur TLP, Forest F, Baker WJ. 2011. Origin and global diversification patterns of tropical rain forests: inferences from a complete genus-level phylogeny of palms. *BMC Biology* **9:** 44.

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

Cruaud A, Gautier M, Galan M, Foucaud J, Saune L, Genson G, Dubois E, Nidelet S, Deuve T, Rasplus JY. 2014. Empirical assessment of RAD sequencing for interspecific phylogeny. *Molecular Biology and Evolution* **31:** 1272–1274.

Daniel R. 2005. The metagenomics of soil. *Nature Reviews Microbiology* **3:** 470–478.

Davies KTJ, Bennett NC, Tsagkogeorga G, Rossiter SJ, Faulkes CG. 2015. Family wide molecular adaptations to underground life in African mole-rats revealed by phylogenomic analysis. *Molecular Biology and Evolution* **32:** 3089–3107.

Day WHE. 1983. Computationally difficult parsimony problems in phylogenetic systematics. *Journal of Theoretical Biology* **103:** 429–438.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology and Evolution* **24:** 332–340.

Delmont TO, Eren AM, Maccario L, Prestat E, Esen OC, Pelletier E, Le Paslier D, Simonet P, Vogel TM. 2015. Reconstructing rare soil microbial genomes using in situ enrichments and metagenomics. *Frontiers in Microbiology* **6:** 358.

van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. 2014. Ten years of next-generation sequencing technology. *Trends in Genetics* **30:** 418–426.

Dransfield J, Uhl NW, Asmussen CB, Baker WJ, Harley MM, Lewis CE. 2005. A new phylogenetic classification of the palm family, Arecaceae. *Kew Bulletin* **60:** 559–569.

Dransfield J, Uhl NW, Asmussen CB, Baker WJ, Harley MM, Lewis CE. 2008. *Genera palmarum – the evolution and classification of palms*. Kew: Royal Botanic Gardens.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* **7:** 214.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* **29:** 1969–1973.

Dunn CW, Luo X, Wu ZJ. 2013. Phylogenetic analysis of gene expression. *Integrative and Comparative Biology* **53:** 847–856.

Eaton DAR, Ree RH. 2013. Inferring phylogeny and introgression using RADseq data: an example from flowering plants (*Pedicularis*: Orobanchaceae). *Systematic Biology* **62:** 689–706.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, deWinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* **323:** 133–138.

Eisen JA. 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Research* **8:** 163–167.

Ekblom R, Jochen BWW. 2014. A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications* **7:** 1026–1042.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* **6:** e19379.

Emerson KJ, Merz CR, Catchen JM, Hohenlohe PA, Cresko WA, Bradshaw WE, Holzapfel CM. 2010. Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **107:** 16196–16200.

Ewing B, Green P. 1998. Base-calling of automated sequencer traces using PHRED. II. Error probabilities. *Genome Research* **8:** 186–194.

Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology* **61:** 717–726.

Fang YJ, Wu H, Zhang TW, Yang M, Yin YX, Pan LL, Yu XG, Zhang XW, Hu SNA, Al-Mssallem IS, Yu J. 2012. A complete sequence and transcriptomic analyses of date palm (*Phoenix dactylifera* L.) mitochondrial genome. *PLoS ONE* **7:** e37164.

Folk RA, Mandel JR, Freudenstein JV. 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:** apps.1500039. doi: 10.3732/apps.1500039.

Gatesy J, Springer MS. 2014. Phylogenetic analysis at deep timescales: unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. *Molecular Phylogenetics and Evolution* **80:** 231–266.

Gaut BS, Muse SV, Clark WD, Clegg MT. 1992. Relative rates of nucleotide substitution at the *rbcL* locus of monocotyledonous plants. *Journal of Molecular Evolution* **35:** 292–303.

Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A. 2005. Galaxy: a platform for interactive large-scale genome analysis. *Genome Research* **15:** 1451–1455.

Gilbert JA, Hughes M. 2011. Gene Expression Profiling: Metatranscriptomics. *Methods in Molecular Biology* **733:** 195–205.

Glenn TC. 2011. Field guide to next-generation DNA sequencers. *Molecular Ecology Resources* **11:** 759–769.

González VL, Andrade SCS, Bieler R, Collins TM, Dunn CW, Mikkelsen PM, Taylor JD, Giribet G. 2015. A phylogenetic backbone for Bivalvia: an RNA-seq approach. *Proceedings of the Royal Society of London B: Biological Sciences* **282:** 20142332.

Gremme G, Steinbiss S, Kurtz S. 2013. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE-ACM Transactions on Computational Biology and Bioinformatics* **10:** 645–656.

Grover CE, Salmon A, Wendel JF. 2012. Targeted sequence capture as a powerful tool for evolutionary analysis. *American Journal of Botany* **99:** 312–319.

Haridas S, Breuill C, Bohlmann J, Hsiang T. 2011. A biologist's guide to *de novo* genome assembly using next-generation sequence data: a test with fungal genomes. *Journal of Microbiological Methods* **86:** 368–375.

Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution* **27:** 570–580.

Heyduk K, Trapnell DW, Barrett CF, Leebens-Mack J. 2015. Phylogenomic analyses of species relationships in the genus *Sabal* (Arecaceae) using targeted sequence capture. *Biological Journal of the Linnean Society* **117:** 106–120.

Hillis DM, Moritz C. 1990. *Molecular systematics*. Sunderland: Sinauer.

Hodges E, Rooks M, Xuan ZY, Bhattacharjee A, Gordon DB, Brizuela L, McCombie WR, Hannon GJ. 2009. Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nature Protocols* **4:** 960–974.

Hohenlohe PA, Amish SJ, Catchen JM, Allendorf FW, Luikart G. 2011. Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources* **11:** 117–122.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics* **6:** e1000862.

Hohenlohe PA, Day MD, Amish SJ, Miller MR, Kamps-Hughes N, Boyer MC, Muhlfeld CC, Allendorf FW, Johnson EA, Luikart G. 2013. Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. *Molecular Ecology* **22:** 3002–3013.

Huang YY, Matzke AJM, Matzke M. 2013. Complete sequence and comparative analysis of the chloroplast genome of coconut palm (*Cocos nucifera*). *PLoS ONE* **8:** e74736.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17:** 754–755.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Jiao YN, Li JP, Tang HB, Paterson AH. 2014. Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. *Plant Cell* **26:** 2792–2802.

Kembel SW, O'Connor TK, Arnold HK, Hubbell SP, Wright SJ, Green JL. 2014. Relationships between phyllosphere bacterial communities and plant functional traits in a neotropical forest. *Proceedings of the National Academy of Sciences of the United States of America* **111:** 13715–13720.

Koren S, Phillippy AM. 2015. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current Opinion in Microbiology* **23:** 110–120.

Lanfear R, Ho SYW, Davies TJ, Moles AT, Aarssen L, Swenson NG, Warman L, Zanne AE, Allen AP. 2013. Taller plants have lower rates of molecular evolution. *Nature Communications* **4:** 1879.

Larget BR, Kotha SK, Dewey CN, Ane C. 2010. BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics* **26:** 2910–2911.

Laver T, Harrison J, O'Neill PA, Moore K, Farbos A, Paszkiewicz K, Studholme DJ. 2015. Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification* **3:** 1–8.

Leaché AD, Banbury BL, Felsenstein J, Nieto-Montes de Oca A, Stamatakis A. 2015a. Short tree, long tree, right tree, wrong tree: new acquisition bias corrections for inferring SNP phylogenies. *Systematic Biology.* **64:** 1032–1047.

Leaché AD, Chavez AS, Jones LN, Grummer JA, Gottscho AD, Linkem CW. 2015b. Phylogenomics of phrynosomatid lizards: conflicting signals from sequence capture versus restriction site associated DNA sequencing. *Genome Biology and Evolution* **7:** 706–719.

Lemmon AR, Emme SA, Lemmon EM. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology* **61:** 727–744.

Lemmon EM, Lemmon AR. 2013. High-throughput genomic data in systematics and phylogenetics. *Annual Review of Ecology, Evolution, and Systematics* **44:** 99–121.

Lewis CE, Doyle JJ. 2001. Phylogenetic utility of the nuclear gene malate synthase in the palm family (Arecaceae). *Molecular Phylogenetics and Evolution* **19:** 409–420.

Li RQ, Zhu HM, Ruan J, Qian WB, Fang XD, Shi ZB, Li YR, Li ST, Shan G, Kristiansen K, Li SG, Yang HM, Wang J, Wang J. 2010. *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Research* **20:** 265–272.

Liu L. 2008. BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics* **24:** 2542–2543.

Liu L, Pearl DK. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Systematic Biology* **56:** 504–514.

Liu L, Yu LL, Kubatko L, Pearl DK, Edwards SV. 2009a. Coalescent methods for estimating phylogenetic trees. *Molecular Phylogenetics and Evolution* **53:** 320–328.

Liu L, Yu LL, Pearl DK, Edwards SV. 2009b. Estimating species phylogenies using coalescence times among sequences. *Systematic Biology* **58:** 468–477.

Liu LA, Yu LL, Edwards SV. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology* **10:** 302.

Mardis ER. 2013. Next-generation sequencing platforms. *Annual Review of Analytical Chemistry* **6:** 287–303.

Mariac C, Scarcelli N, Pouzadou J, Barnaud A, Billot C, Faye A, Kougbeadjo A, Maillol V, Martin G, Sabot F, Santoni S, Vigouroux Y, Couvreur TLP. 2014. Cost-effective enrichment hybridization capture of chloroplast genomes at deep multiplexing levels for population genetics and phylogeography studies. *Molecular Ecology Resources* **14:** 1103–1113.

Matasci N, Hung L-H, Yan Z, Carpenter EJ, Wickett NJ, Mirarab S, Nguyen N, Warnow T, Ayyampalayam S, Barker M, Burleigh J, Gitzendanner MA, Wafula E, Der JP, dePamphilis CW, Roure B, Philippe H, Ruhfel BR, Miles NW, Graham SW, Mathews S, Surek B, Melkonian M, Soltis DE, Soltis PS, Rothfels C, Pokorny L, Shaw JA, DeGironimo L, Stevenson DW, Villarreal J, Chen T, Kutchan TM, Rolf M, Baucom RS, Deyholos MK, Samudrala R, Tian Z, Wu X, Sun X, Zhang Y, Wang J, Leebens-Mack J, Wong GK-S. 2014. Data access for the 1,000 Plants (1KP) project. *GigaScience* **3:** 17.

Maxam AM, Gilbert W. 1977. New method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America* **74:** 560–564.

McCormack JE, Faircloth BC, Crawford NG, Gowaty PA, Brumfield RT, Glenn TC. 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome Research* **22:** 746–754.

Miller MA, Pfeiffer W, Schwartz T. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. Proceedings of the Gateway Computing Environments Workshop (GCE), 14 Nov. 2010, New Orleans, 1–8.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA. 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research* **17:** 240–248.

Miller W, Drautz DI, Janecka JE, Lesk AM, Ratan A, Tomsho LP, Packard M, Zhang Y, McClellan LR, Qi J, Zhao F, Gilbert MTP, Dalen L, Arsuaga JL, Ericson PGP, Huson DH, Helgen KM, Murphy WJ, Gotherstrom A, Schuster SC. 2009. The mitochondrial genome sequence of the Tasmanian tiger (*Thylacinus cynocephalus*). *Genome Research* **19:** 213–220.

Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. 2014. ASTRAL: Genome-scale coalescent-based species tree estimation. *Bioinformatics* **30:** I541–I548.

Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H. 1986. Specific enzymatic amplification of DNA *in vitro*: the polymerase chain reaction. *Cold Spring Harbor Symposium on Quantitative Biology* **51:** 263–273.

Nagarajan N, Pop M. 2013. Sequence assembly demystified. *Nature Reviews Genetics* **14:** 157–167.

Nater A, Burri R, Kawakami T, Smeds L, Ellegren H. 2015. Resolving evolutionary relationships in closely related species with whole-genome sequencing data. *Systematic Biology* **64:** 1000–1017.

Noble WS. 2009. A quick guide to organizing computational biology projects. *PLoS Computational Biology* **5:** e1000424.

Osborne JM, Bernabeu MO, Bruna M, Calderhead B, Cooper J, Dalchau N, Dunn SJ, Fletcher AG, Freeman R, Groen D, Knapp B, McInerny GJ, Mirams GR, Pitt-Francis J, Sengupta B, Wright DW, Yates CA, Gavaghan DJ, Emmott S, Deane C. 2014. Ten simple rules for effective computational research. *PLoS Computational Biology* **10:** e1003506.

Pante E, Abdelkrim J, Viricel A, Gey D, France SC, Boisselier MC, Samadi S. 2015. Use of RAD sequencing for delimiting species. *Heredity* **114:** 450–459.

Papadopoulou A, Taberlet P, Zinger L. 2015. Metagenome skimming for phylogenetic community ecology: a new era in biodiversity research. *Molecular Ecology* **24:** 3515–3517.

Patel RK, Jain M. 2012. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS ONE* **7:** e30619.

Peñalba JV, Smith LL, Tonione MA, Sass C, Hykin SM, Skipwith PL, McGuire JA, Bowie RCK, Moritz C. 2014. Sequence capture using PCR-generated probes: a cost-effective method of targeted high-throughput sequencing for nonmodel organisms. *Molecular Ecology Resources* **14:** 1000–1010.

Pendleton M, Sebra R, Pang AWC, Ummat A, Franzen O, Rausch T, Stutz AM, Stedman W, Anantharaman T, Hastie A, Dai H, Fritz MHY, Cao H, Cohainl A, Deikusl G, Durrett RE, Blanchard SC, Altman R, Chin CS, Guo Y, Paxinos EE, Korbe JO, Darne RB, McCombie WR, Kwok PY, Mason CE, Schadt EE, Bashirl A. 2015. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature Methods* **12:** 780–U140.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. 2012. Double digest RADseq: An inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PLoS ONE* **7:** e37135.

Ponciano JM, Burleigh JG, Braun EL, Taper ML. 2012. Assessing parameter identifiability in phylogenetic models using data cloning. *Systematic Biology* **61:** 955–972.

Prum RO, Berv JS, Dornburg A, Field DJ, Townsend JP, Lemmon EM, Lemmon AR. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature* **526:** 569–573.

Quero H, Yáñez E. 2000. El complejo *Brahea–Erythea* (Palmae: Coryphideae). Proyecto CONABIO L216. Informe final. Available at: http://www.conabio.gob.mx/institucion/cgi-bin/datos.cgi?Letras=L&Numero=216. Last accessed 10/8/15.

Ramirez KS, Leff JW, Barberán A, Bates ST, Betley J, Crowther TW, Kelly EF, Oldfield EE, Shaw EA, Steenbock C, Bradford MA, Wall DH, Fierer N. 2014. Biogeographic patterns in belowground diversity in New York City's Central Park are similar to those observed globally. *Proceedings of the Royal Society B: Biological Sciences* **281:** 20141988.

Reitzel AM, Herrera S, Layden MJ, Martindale MQ, Shank TM. 2013. Going where traditional markers have not gone before: utility of and promise for RAD sequencing in marine invertebrate phylogeography and population genomics. *Molecular Ecology* **22:** 2953–2970.

Roch S, Warnow T. 2015. On the robustness to gene tree estimation error (or lack thereof) of coalescent-based species tree methods. *Systematic Biology* **64:** 663–676.

Roncal J, Borchsenius F, Asmussen-Lange CB, Balslev H. 2010. Divergence times in tribe Geonomateae (Arecaceae) coincide with Tertiary geological events. In: Seberg O, Pedersen G, Barfod AS Davis JI, eds. *Diversity, phylogeny, and evolution of the monocotyledons*. Århus: Aarhus University Press, 245–265.

Roncal J, Francisco-Ortega J, Asmussen CB, Lewis CE. 2005. Molecular phylogenetics of tribe Geonomeae (Arecaceae) using nuclear DNA sequences of phosphoribulokinase and RNA polymerase II. *Systematic Botany* **30:** 275–283.

Roncal J, Henderson A, Borchsenius F, Cardoso SRS, Balslev H. 2012. Can phylogenetic signal, character displacement, or random phenotypic drift explain the morphological variation in the genus *Geonoma* (Arecaceae)? *Biological Journal of the Linnean Society* **106:** 528–539.

Rubin BER, Ree RH, Moreau CS. 2012. Inferring phylogenies from RAD sequence data. *PLoS ONE* **7:** e33394.

Ruhfel BR, Gitzendanner MA, Soltis PS, Soltis DE, Burleigh JG. 2014. From algae to angiosperms-inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Evolutionary Biology* **14:** 23.

Sambrook J, Fritsch EF, Maniatis T. 1989. *Molecular cloning*. New York: Cold Spring Harbor Laboratory Press.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74:** 5463–5467.

Scarcelli N, Barnaud A, Eiserhardt W, Treier UA, Seveno M, d'Anfray A, Vigouroux Y, Pintaud JC. 2011. A set of 100 chloroplast DNA primer pairs to study population genetics and phylogeny in monocotyledons. *PLoS ONE* **6:** e19954.

Scarcelli N, Mariac C, Couvreur TLP, Faye A, Richard D, Sabot F, Berthouly-Salazar C, Vigouroux Y. 2016. Intra-individual polymorphism in chloroplasts from NGS data: where does it come from and how to handle it? *Molecular Ecology Resources* **16:** 434–445.

Schadt EE, Turner S, Kasarskis A. 2010. A window into third-generation sequencing. *Human Molecular Genetics* **19:** R227–R240.

Schatz MC, Witkowski J, McCombie WR. 2012. Current challenges in de novo plant genome sequencing and assembly. *Genome Biology* **13:** 243.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei FS, Pasternak S, Liang CZ, Zhang JW, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du FY, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen WZ, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L, Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, Kohlberg S, Sgro J, Delgado B, Mead K, Chinwalla A, Leonard S, Crouse K, Collura K, Kudrna D, Currie J, He RF, Angelova A, Rajasekar S, Mueller T, Lomeli R, Scara G, Ko A, Delaney K, Wissotski M, Lopez G, Campos D, Braidotti M, Ashley E, Golser W, Kim H, Lee S, Lin JK, Dujmic Z, Kim W, Talag J, Zuccolo A, Fan C, Sebastian A, Kramer M, Spiegel L, Nascimento L, Zutavern T, Miller B, Ambroise C, Muller S, Spooner W, Narechania A, Ren LY, Wei S, Kumari S, Faga B, Levy MJ, McMahan L, Van Buren P, Vaughn MW, *et al.* 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326:** 1112–1115.

Simmons MP, Gatesy J. 2015. Coalescence vs. concatenation: sophisticated analyses vs. first principles applied to rooting the angiosperms. *Molecular Phylogenetics and Evolution* **91:** 98–122.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. *Genome Research* **19:** 1117–1123.

Singh R, Ong-Abdullah M, Low ETL, Manaf MAA, Rosli R, Nookiah R, Ooi LCL, Ooi SE, Chan KL, Halim MA, Azizi N, Nagappan J, Bacher B, Lakey N, Smith SW, He D, Hogan M, Budiman MA, Lee EK, DeSalle R, Kudrna D, Goicoechea JL, Wing RA, Wilson RK, Fulton RS, Ordway JM, Martienssen RA, Sambanthamurthi R. 2013. Oil palm genome sequence reveals divergence of interfertile species in Old and New Worlds. *Nature* **500:** 335–339.

Sjölander K. 2004. Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics* **20:** 170–179.

Soltis PS, Soltis DE, Doyle JJ. 1992. *Molecular systematics of plants*. New York: Chapman and Hall.

Song S, Liu L, Edwards SV, Wu SY. 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proceedings of the National Academy of Sciences of the United States of America* **109:** 14942–14947.

de Sousa F, Bertrand YJK, Nylinder S, Oxelman B, Eriksson JS, Pfeil BE. 2014. Phylogenetic properties of 50 nuclear loci in *Medicago* (Leguminosae) generated using multiplexed sequence capture and next-generation sequencing. *PLoS ONE* **9:** e109704.

Springer MS, Gatesy J. 2016. The gene tree delusion. *Molecular Phylogenetics and Evolution* **94:** 1–33.

Staats M, Erkens RHJ, van de Vossenberg B, Wieringa JJ, Kraaijeveld K, Stielow B, Geml J, Richardson JE, Bakker FT. 2013. Genomic treasure troves: complete genome sequencing of herbarium and insect museum specimens. *PLoS ONE* **8:** e69189.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22:** 2688–2690.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30:** 1312–1313.

Stamatakis A, Aberer AJ. 2013. Novel parallelization schemes for large-scale likelihood-based phylogenetic inference. *IEEE 27th International Parallel and Distributed Processing Symposium (IPDPS 2013)*: 1195–1204.

Steele PR, Hertweck KL, Mayfield D, McKain MR, Leebens-Mack J, Pires JC. 2012. Quality and quantity of data recovered from massively parallel sequencing: examples in Asparagales and Poaceae. *American Journal of Botany* **99:** 330–348.

Stevens PF. 2001 onwards**.** Angiosperm Phylogeny Website. Version 12, July 2012. Available at: http://www.mobot.org/MOBOT/research/APweb. Last accessed 10/7/2015.

Stevens PF, Davis HM. 2005. The Angiosperm Phylogeny Website – a tool for reference and teaching in a time of change. *Proceedings of the American Society for Information Science and Technology* **42:** 1. doi: 10.1002/meet.14504201249.

Straub SCK, Fishbein M, Livshultz T, Foster Z, Parks M, Weitemier K, Cronn RC, Liston A. 2011. Building a model: developing genomic resources for common milkweed (*Asclepias syriaca*) with low coverage genome sequencing. *BMC Genomics* **12:** 211.

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

Streicher JW, Schulte JA, Wiens JJ. 2015. How should genes and taxa be sampled for phylogenomic analyses with missing data? An empirical study in iguanian lizards *Systematic Biology* **65:** 128–145.

Stull GW, Moore MJ, Mandala VS, Douglas NA, Kates HR, Qi X, Brockington SF, Soltis PS, Soltis DE, Gitzendanner MA. 2013. A targeted enrichment strategy for massively parallel sequencing of angiosperm plastid genomes. *Applications in Plant Sciences* **1:** apps1200497.

Szöllősi GJ, Davin AA, Tannier E, Daubin V, Boussau B. 2015. Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. *Philosophical Transactions of the Royal Society B-Biological Sciences* **370:** 20140335.

Travers KJ, Chin CS, Rank DR, Eid JS, Turner SW. 2010. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Research* **38:** e159.

Uhl NW, Dransfield J. 1987. *Genera palmarum: a classification of palms based on the work of Harold E. Moore, Jr*. Lawrence: Allen Press.

Uthaipaisanwong P, Chanprasert J, Shearman JR, Sangsrakru D, Yoocha T, Jomchai N, Jantasuriyarat C, Tragoonrung S, Tangphatsornruang S. 2012. Characterization of the chloroplast genome sequence of oil palm (*Elaeis guineensis* Jacq.). *Gene* **500:** 172–180.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu DY, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304:** 66–74.

Weitemier K, Straub SCK, Cronn R, Fishbein M, McDonnell A, Schmickl R, Liston A. 2014. Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences* **2:** 1400042.

Westbrook CJ, Karl JA, Wiseman RW, Mate S, Koroleva G, Garcia K, Sanchez-Lockhart M, O'Connor DH, Palacios G. 2015. No assembly required: Full-length MHC class I allele discovery by PacBio circular consensus sequencing. *Human Immunology* **76:** 891–896.

Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, Ayyampalayam S, Barker MS, Burleigh JG, Gitzendanner MA, Ruhfel BR, Wafula E, Der JP, Graham SW, Mathews S, Melkonian M, Soltis DE, Soltis PS, Miles NW, Rothfels CJ, Pokorny L, Shaw AJ, DeGironimo L, Stevenson DW, Surek B, Villarreal JC, Roure B, Philippe H, dePamphilis CW, Chen T, Deyholos MK, Baucom RS, Kutchan TM, Augustin MM, Wang J, Zhang Y, Tian ZJ, Yan ZX, Wu XL, Sun X, Wong GKS, Leebens-Mack J. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences of the United States of America* **111:** E4859–E4868.

Wu SY, Song S, Liu L, Edwards SV. 2013. Reply to Gatesy and Springer: the multispecies coalescent model can effectively handle recombination and gene tree heterogeneity. *Proceedings of the National Academy of Sciences of the United States of America* **110:** E1180–E1180.

Wyman SK, Jansen RK, Boore JL. 2004. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* **20:** 3252–3255.

Wysocki WP, Clark LG, Kelchner SA, Burke SV, Pires JC, Edger PP, Mayfield DR, Triplett JK, Columbus JT, Ingram AL, Duvall MR. 2014. A multi-step comparison of short-read full plastome sequence assembly methods in grasses. *Taxon* **63:** 899–910.

Xi ZX, Liu L, Rest JS, Davis CC. 2014. Coalescent versus concatenation methods and the placement of *Amborella* as sister to water lilies. *Systematic Biology* **63:** 919–932.

Yang M, Zhang XW, Liu GM, Yin YX, Chen KF, Yun QZ, Zhao DJ, Al-Mssallem IS, Yu J. 2010. The complete chloroplast genome sequence of date palm (*Phoenix dactylifera* L.). *PLoS ONE* **5:** e12762.

Zedane L, Hong-Wa C, Murienne J, Jeziorski C, Baldwin BG, Besnard G. 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.

Zerbino DR, Birney E. 2008. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Research* **18:** 821–829.