An introduction to plant phylogenomics with a focus on palms

CRAIG F. BARRETT¹,*, CHRISTINE D. BACON², ALEXANDRE ANTONELLI²,3, ÁNGELA CANO⁴ and TOBIAS HOFMANN²

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


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
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 Szöllösi 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, ginger, 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 Trachycarpaeae; 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 genomes 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.
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 Arecoidae), J. Tregar, unpubl. data; *Geonoma undata* Klotzsch (Arecoidae), C. Lexer, unpubl. data; *Mauritia flexuosa* L.f. (Calamoideae), Tregar 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.,...
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).

**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.

**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>
<thead>
<tr>
<th>Method</th>
<th>Cost</th>
<th>Technical requirements</th>
<th>Bioinformatics</th>
<th>Wasted data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole genome sequencing</td>
<td>extremely high*</td>
<td>extremely high</td>
<td>extremely high</td>
<td>little</td>
</tr>
<tr>
<td>Genome skimming</td>
<td>low</td>
<td>low</td>
<td>low</td>
<td>vast majority</td>
</tr>
<tr>
<td>Transcriptomes (phylogenetics)</td>
<td>high</td>
<td>moderate</td>
<td>high</td>
<td>majority</td>
</tr>
<tr>
<td>Transcriptomes (gene expression)</td>
<td>very high</td>
<td>moderate</td>
<td>high</td>
<td>little</td>
</tr>
<tr>
<td>Sequence capture</td>
<td>moderate–high**</td>
<td>high</td>
<td>high</td>
<td>little</td>
</tr>
</tbody>
</table>

*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 \(-80\, ^\circ\text{C}\) 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; Penalba 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. Blumentiel 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), Genious (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 $10^{-1}$ 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); Trimomatic (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,
lelizing phylogenetic analyses, many challenges offer their own computational resources. Miller, Pfeiffer & Schwartz (2010), https://www.phylo.org and many research-oriented institutions modest to medium-sized analyses [e.g. CIPRES; some freely available online clusters can often help to divide the workload and speed up 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 Eugeissonae 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 (bootstrap 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>
<thead>
<tr>
<th>Taxonomic level</th>
<th>Raw reads</th>
<th>Contigs</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chamaedoreeae</td>
<td>Tribe</td>
<td>620 000</td>
<td>84 855</td>
</tr>
<tr>
<td><strong>Sabal</strong></td>
<td>Inter-specific</td>
<td>166 208</td>
<td>11 588</td>
</tr>
<tr>
<td>Geonoma macrostachys</td>
<td>Intra-specific</td>
<td>1 737 895</td>
<td>43 026</td>
</tr>
</tbody>
</table>

*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 [Bio-perl.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 phylogenomics 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 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


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


**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.


**Travers JS, 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.


