A dedicated target capture approach reveals variable genetic markers across micro- and macro-evolutionary time scales in palms

Marylaure de La Harpe[1] | Jaqueline Hess[1] | Oriane Loiseau[2,3] | Nicolas Salamin[2,3] | Christian Lexer[1] | Margot Paris[4]

[1]Department of Botany and Biodiversity Research, University of Vienna, Vienna, Austria

[2]Department of Computational Biology, Biophore, University of Lausanne, Lausanne, Switzerland

[3]Swiss Institute of Bioinformatics, Lausanne, Switzerland

[4]Department of Biology, Unit Ecology and Evolution, University of Fribourg, Fribourg, Switzerland

Correspondence
Margot Paris, University of Fribourg, Fribourg, Switzerland.
Email: margotparis1@gmail.com

Funding information
Swiss National Science Foundation (SNSF), Grant/Award Number: CRSII3_147630; University of Zurich; Illumina; National Science Foundation

Abstract

Understanding the genetics of biological diversification across micro- and macro-evolutionary time scales is a vibrant field of research for molecular ecologists as rapid advances in sequencing technologies promise to overcome former limitations. In palms, an emblematic, economically and ecologically important plant family with high diversity in the tropics, studies of diversification at the population and species levels are still hampered by a lack of genomic markers suitable for the genotyping of large numbers of recently diverged taxa. To fill this gap, we used a whole genome sequencing approach to develop target sequencing for molecular markers in 4,184 genome regions, including 4,051 genes and 133 non-genic putatively neutral regions. These markers were chosen to cover a wide range of evolutionary rates allowing future studies at the family, genus, species and population levels. Special emphasis was given to the avoidance of copy number variation during marker selection. In addition, a set of 149 well-known sequence regions previously used as phylogenetic markers by the palm biological research community were included in the target regions, to open the possibility to combine and jointly analyse already available data sets with genomic data to be produced with this new toolkit. The bait set was effective for species belonging to all three palm subfamilies tested (Arecoideae, Ceroxyloideae and Coryphoideae), with high mapping rates, specificity and efficiency. The number of high-quality single nucleotide polymorphisms (SNPs) detected at both the sub-family and population levels facilitates efficient analyses of genomic diversity across micro- and macro-evolutionary time scales.

KEYWORDS
Arecaceae, phylogenomics, population genomics, sequence capture, whole genome sequencing

1 | INTRODUCTION

A particularly ambitious goal of molecular ecology and evolutionary biology is to understand the genetic underpinnings of biological diversification at both micro- and macro-evolutionary time scales (Futuyma & Kirkpatrick, 2017). The necessity of bridging these two evolutionary fields is widely appreciated by the molecular ecology community (Bragg, Potter, Bi, & Moritz, 2016; de La Harpe et al., 2017; Glenn & Faircloth, 2016). However, achieving it is far from

Christian Lexer and Margot Paris should be considered joint senior author.

trivial in any organismal group (Rolland, Silvestro, Litsios, Faye, & Salamin, 2018) and developing molecular genetic toolkits able to span both micro- and macro-evolutionary time scales would thus be particularly helpful for studies of evolutionary radiation. We focus here on palms (Arecaceae), a highly diverse (>2,600 species) plant family including (a) several rapid radiations in the Neo- and Paleotropics (Couvreur & Baker, 2013), (b) many species filling important ecological niches in tropical rain forests at low to intermediate altitudes (Balslev et al., 2011; Balslev, Bernal, & Fay, 2016) and (c) several economically important taxa, such as the African and American oil palm (*Elaeis guineensis* and *Elaeis oleifera*) and the date palm (*Phoenix dactylifera*).

While markers traditionally used in molecular ecology and evolutionary genetics have generally been limited to either short or long evolutionary time scales (de La Harpe et al., 2017), the sequencing of whole genomes or transcriptomes promises to overcome these limitations. In this context, whole genome sequencing (WGS) is a powerful approach to catalogue genomic variation, but its cost is still prohibitive for most research groups working with hundreds of population samples or organisms with large genomes (McCartney-Melstad, Mount, & Shaffer, 2016; Suren et al., 2016). Thus, sequence capture is becoming a popular approach to reduce sequencing cost by targeting a scalable number of genomic regions of interest (Glenn & Faircloth, 2016). The main advantage of this method is its flexibility, as the number, size, location and nature of targeted genomic regions can be adapted to specific research questions in any non-model organism (de La Harpe et al., 2017; Lemmon & Lemmon, 2013). Target capture has proven to be effective for studies covering micro- and/or macro-evolutionary time scales, but its development is challenging. First, the identification of genomic regions spanning different evolutionary rates is necessary to maximize information from shallow to deep phylogenetic distances. At the same time, target regions should ideally be conserved enough to successfully genotype a large number of species. Because of this challenge, ultra-conserved elements or organellar capture has been widely used for phylogenomic analyses in deeply divergent lineages (Faircloth, Branstetter, White, & Brady, 2015; Lemmon & Lemmon, 2013; Prum et al., 2015; Siepel et al., 2005). Sequence capture is also increasingly used in plant phylogenomics, and probes were specifically designed for various genera including *Bartsia* (Uribe-Convers, Settles, & Tank, 2016), *Helianthus* (Stephens, Rogers, Mason, Donovan, & Malmberg, 2015), *Inga* (Nicholls et al., 2015), *Sarracenia* (Stephens, Rogers, Heyduk, et al., 2015) and the order Zingiberales (Sass, Iles, Barrett, Smith, & Specht, 2016) while a potentially universal angiosperm probe kit was recently made available (Buddenhagen et al., 2016). More shallow time scales or intraspecific analyses mostly relied on exon (Suren et al., 2016; Syring et al., 2016; Zhou & Holliday, 2012) or RAD-derived marker captures (Schmid et al., 2017; Suchan et al., 2016). Recently, Heyduk, Trapnell, Barrett, and Leebens-Mack (2016) developed a target capture set for palms based on a total of 837 conserved exons from 176 nuclear genes. Based on phylogenetic inferences, this capture set demonstrated a high utility to resolve relationships among anciently diverged species, but ambiguities remain at low taxonomic levels (e.g., for closely related species within the *Sabal* genus, Heyduk et al., 2016). Therefore, a newly improved capture set covering both micro- and macro-evolutionary time scales would represent a great opportunity for research projects on the ecology, evolutionary biology, systematics and many other aspects of this emblematic, economically and ecologically important plant family.

A second major challenge in target capture development is the selection of genomic regions or genes without pervasive copy number polymorphism, which can be difficult in plants that display many genomic duplications. The genomes of *E. guinensis*, *E. oleifera* and *P. dactylifera* revealed abundant segmental duplications, suggesting that their ancestor was a polyploid (Singh et al., 2013). Despite their palaeotetraploid origin, chromosome counts showed that most extant palm species are diploid with chromosome numbers between 2n = 26 and 2n = 36 (Bennett & Leitch, 2012; Roser, Johnson, & Hanson, 1997). Several methods have been described for the detection of putative paralogs. A first approach is based on the use of known sequence information such as genome or transcriptome assemblies when available for the study system (Bragg et al., 2016; Chamala et al., 2015; Heyduk et al., 2016). Putative orthologs are identified for each species using independent reciprocal BLASTx searches, and only genes identified as single copy in all species tested are retained as targets. If a reference genome is available, a second possible strategy is to identify copy number variable regions based on whole genome sequencing of one or ideally several closely related individuals. Duplications and losses can be detected through changes in coverage or in the orientation or size distribution of paired-end reads, and a plethora of methods is available for this purpose (reviewed in Zhao, Wang, Wang, Jia, & Zhao, 2013). Such regions can then be removed from bait design. A third, elegant approach uses sequence information to detect an apparent excess of heterozygous sites when reads from different paralogs are mapped to the same region of the reference genome (Djedatin, Monat, Engelen, & Sabot, 2017).

In this article, we present the development of a target capture kit allowing analyses across evolutionary time scales in palms. We chart a pipeline of in silico and wet lab steps to obtain a set of target enrichment capture probes able to efficiently interrogate a large number of genic and intergenic regions for genomic variation across a range of different evolutionary distances, making use of available and newly created genome resources. Our strategy was to first sequence the whole genome of one sample of *Geonoma undata*, a widely distributed Neotropical species (Henderson, 2011), in order to (a) detect single or low-copy genes using coverage and heterozygosity information, (b) detect non-genic regions conserved between *G. undata* and the reference genome of the oil palm *Elaeis guinensis* (Singh et al., 2013) and (c) obtain *G. undata* sequence information for the design of the probes. Using this information, a set of 3,920 single or low-copy genes and 133 non-genic conserved regions (putative neutral regions) were chosen as target regions. In addition, 131 well-known sequence regions previously used as phylogenetic markers by palm systematists, biogeographers and evolutionary biologists were added to the target regions. The goal was to give the evolutionary and conservation biology and systematic communities the possibility to combine and jointly analyse already available phylogenetic data sets with genomic data to be produced with this new kit. We tested the efficiency, specificity and reproducibility of the kit and its utility for SNP discovery across micro- and macro-evolutionary time scales including species from three palm sub-families with up to 87 Myr divergence from *G. undata* (Baker & Couvreur, 2013; Roncal, Blach-Overgaard, Borchsenius, Balslev, & Svenning, 2011), as well as intraspecific and population samples of *G. undata*. Finally, we assessed the utility of the kit for phylogenetic and population genetic analyses.

2 | MATERIALS AND METHODS

2.1 | Whole genome sequencing of *Geonoma undata*

We performed whole genome sequencing (WGS) of the species *G. undata*, using the specimen IO260_A collected in the Peruvian Amazon (S05°49′05.5, W77°49′20.1). Voucher deposit details of the samples used in this manuscript are presented in Supporting information Table S1. Leaves were dried in silica gel before DNA extraction with the DNeasy® Plant mini Kit (Qiagen, Venlo, the Netherlands). Library preparation was performed using the Illumina TruSeq® DNA PCR-Free Sample Preparation Kit before sequencing in 1/2 of an Illumina HiSeq3000 lane (paired-end 2 × 126 bp). Reads were trimmed with the program condetri v2.2 (Smeds & Kunstner, 2011) using 20 as high-quality threshold parameter, before mapping with bowtie2 v2.2.5 (Langmead & Salzberg, 2012) to the African oil palm genome (*Elaeis guineensis*, NCBI assembly accession GCF_000442705.1), the closest sequenced species to *Geonoma* (around 67 My of divergence, Baker & Couvreur, 2013). gatk v3.3 (McKenna et al., 2010) was used for realignment around indels and base recalibration using default parameters and gatk best practices for data pre-processing of March 2015, before SNP calling with gatk unifiedgenotyper. We used the iterative read mapping and realignment strategy described in Gan et al. (2011) to obtain pseudoreference sequences of *G. undata* for each genomic region covered by our sequencing effort. This method consists of incorporating *G. undata*-specific variation into the *E. guineensis* annotated reference genome, making use of the *E. guineensis* high-quality reference assembly and annotation. Each of the three iterations consisted of: (a) mapping the reads onto the reference/pseudoreference genome using bowtie2 and the very-sensitive-local option (mapping onto the *E. guineensis* reference genome for the first iteration and then to the latest *G. undata* pseudoreference created during the previous iteration), (b) calling the SNPs with gatk unifiedgenotyper v3.3 using only reads with mapping quality >20, (c) filtering high-quality variants with $Q > 30$ and a minimum coverage of 5× with vcftools v0.1.13 (Danecek et al., 2011) and (d) building a new pseudoreference sequence with gatk fastaalternatereferencemaker. The *G. undata* pseudoreference genome contains the plastid genome. With this strategy, multi-allelic sites caused by paralogs mapping at a unique location in the genome were not modified in the new pseudoreference genome. The iterative alignment to preliminary versions of a consensus genome has the advantage to extend subsequent reference-mapping to highly divergent regions, to build sequences into insertions and to help resolving ambiguous calls (Gan et al., 2011; Sarver et al., 2017; Thompson et al., 2015).

Reads were finally mapped to the pseudoreference genome with bowtie2 and the very-sensitive-local option, realigned around indels and base recalibrated with gatk. SNPs were called with gatk unifiedgenotyper using the options --genotype_likelihoods_model SNP and --output_mode EMIT_ALL_SITES in order to obtain calls at both variant and invariant sites. Coverage, mapping statistics, heterozygosity and nucleotide divergence to the oil palm genome were calculated for each annotated exon using bedtools v2.24.0 (Quinlan & Hall, 2010) and vcftools v0.1.13 (Danecek et al., 2011).

2.2 | Genome size estimation of *G. undata*

We used findGSE (Sun, Ding, Piednoel, & Schneeberger, 2018) to estimate the genome size from the quality-trimmed *G. undata* WGS reads. To obtain k-mer spectra, we ran Jellyfish (Marcais & Kingsford, 2011) with $k = 17$ to $k = 21$ in increments of 2. Based on observed k-mer spectra (Supporting information Figure S1), we set exp_hom = 12 and ran findGSE in heterozygous mode for each value of *k*. We also explored higher values of k but the algorithm was unable to run in heterozygous mode for those. Genome size estimates and repeat content estimates are given as averages of estimates with $k = 17–21$.

2.3 | Low-copy gene selection and target choice

For the target selection, we focussed on the 22,957 genes distributed along the 16 assembled chromosomes of the *E. guinenesis* reference genome (77% of the 29,818 annotated genes) and avoided genes located on unplaced low-quality extra scaffolds. A total of 3,920 target genes were selected using preferentially the following criteria: (a) low-copy signature in the *G. undata* genome, (b) available description and known functions, (c) wide range of rates of molecular evolution, avoiding overly conserved genes with low phylogenetic signal and overly variable genes potentially corresponding to paralogs, partial or pseudo-genes, (d) average exonic size of 1,218 bp (range: 225–7,710 bp) avoiding genes composed of many small exons and (e) uniform distribution along the chromosomes. We used both coverage and heterozygosity estimates described in paragraph 2.1 to detect putative duplication signals in *G. undata*. The rationale of these methods is the detection of genes exhibiting an excess of read coverage (see Zhao et al., 2013 for a review) or heterozygous sites (Djedatin et al., 2017) due to the mapping of paralogs in the same region of the reference genome. A total of 2,137 out of the 22,957 genes tested (9.3%) were considered as putative multi-copy in *G. undata*, with coverage and heterozygosity higher than expected compared to the rest of the genome (i.e., higher than the average over all genes +1 SD: minimum coverage of 41×, and minimum heterozygosity of 0.026). Rate of molecular evolution (*E*) was estimated for each gene as the rate of nucleotide substitutions since divergence between *G. undata* gene sequences and the *E. guineensis* reference genome ($E = K/2T$, Hartl & Clark, 2007). We estimated the average number of nucleotide substitutions per site (*K*) using the Jukes and Cantor substitution model correction (1969) and used the divergence time (*T*) of 66.71 Myr estimated between the two species by Baker and Couvreur (2013). The physical distance between selected genes was 159,793 bp on average.

In addition to the 3,920 selected genes, we added known markers commonly used for palm phylogenetics. First, we added 123 genes out of the 175 genes described in Heyduk et al. (2016). We did not retain 52 of the Heyduk et al. (2016) genes because of their multi-copy signals in *G. undata* or their location in the unplaced scaffolds of the *E. guineensis* reference genome. Then, we added eight common low-copy nuclear markers previously used for phylogenetic analyses of palms: PRK (Lewis & Doyle, 2002), MS (Lewis & Doyle, 2001; 2002), RBP2 (Roncal, Francisco-Ortega, Asmussen, & Lewis, 2005), CISP4 (Bacon, Feltus, Paterson, & Bailey, 2008), WRKY2, WRKY7, WRKY19 (Meerow et al., 2009) and PHYB (Ludeña et al., 2011). Two of these markers are located in unplaced scaffolds (CISP4 and PRK). For the four markers PRK, RBP2, CISP4 and PHYB, both exonic and intronic regions were used as target. The description of the selected genes is presented in Supporting information Table S1, and their characteristics compared to all annotated genes are presented in Supporting information Figure S2.

Finally, a set of 133 non-genic markers was added as target. They were conserved between *G. undata* and *E. guineensis* (i.e., *G. undata reads* were successfully mapped onto the *E. guineensis* reference genome at these target locations), distributed on all annotated chromosomes (5–14 per chromosome) and did not present multi-copy signatures in *G. undata*. As their average distance to the closest annotated gene was 55,848 bp (range: 2,636 to 394,122 bp), we considered them as putatively neutral markers.

2.4 | Sample description and DNA extraction

Five genera representing three palm sub-families were used to evaluate the efficiency of the bait capture method at the macro-evolutionary (="intergeneric samples") level (Figure 1). The Arecoideae sub-family was represented by the species *Cocos nucifera* and two species from the Geonomateae tribe: *G. undata* (sample IO260_A for which whole genome sequencing was carried out) and *Asterogyne guianensis*. *Ceroxylon alpinum* represented the Ceroxyloideae sub-family, and *Licuala merguensis* represented the Coryphoideae sub-family.

To test the efficiency of the method at the intraspecific level, we sequenced five *G. undata* samples from five different populations covering the geographic range of the species (="inter-population samples"). Finally, for two of the Colombian *G. undata* populations, five samples were also sequenced in this study (="intra-population samples"). Voucher deposit details are presented in Supporting information Table S1 for all concerned samples. DNA was extracted using the DNeasy® plant mini kit (Qiagen, Venlo, the Netherlands) following the supplier's instructions. DNA quality was evaluated with agarose gels and a NanoDrop™ spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). DNA was quantified with a Qubit® Fluorometer v 2.2 (Invitrogen, Thermo Fisher Scientific, Waltham, MA, USA), and 500 ng of DNA was used for library preparation.

2.5 | Library preparation, dual-indexed sequencing and capture

DNA samples were fragmented to 400-bp fragments with a Bioruptor® ultrasonicator (Diagenode, Liège, Belgium) with six cycles of 30 s. ON, and 90 s. OFF. Sample cleaning, end repair and A-tailing steps were performed with a KAPA LTP library preparation kit (Roche, Basel, Switzerland) following the supplier's protocol with the exception of reaction volumes, which were divided by two to reduce costs. Adaptor ligation and Adaptor fill-in reactions were carried out following Meyer and Kircher (2010). One fifth (4 µl) of the ligated fragment solution was amplified for eight cycles using the KAPA HiFi Hotstart ReadyMix (Roche, Basel, Switzerland) and a set of dual-index primers, as recommended by Kircher, Sawyer, and Meyer (2012) to avoid inaccuracies in multiplex sequencing. Libraries were quantified with a Qubit® FLUOROMETER v 2.2 before pooling in equimolar ratio.

Target capture was conducted on pooled dual-indexed libraries following myBait® Custom Target Capture Kits protocol (Arbor Biosciences, Ann Arbor, MI, USA), with 18-hr incubation time at 65°C and 12 cycles of post-capture PCRs. The library pooling, the entire target capture protocol and the sequencing were repeated twice independently to evaluate the reproducibility of the method. An initial amount of 1.2 µg of pooled libraries (including DNA from 64 samples in total) was used as template for each target capture hybridization reaction. The pooled target capture reactions were quantified with a Qubit® FLUOROMETER v 2.2 before sequencing with an Illumina HiSeq3000 sequencer in paired-end 2 × 150 bp mode. A fraction only of the Illumina lanes were used for our sample sequencing in order to obtain an average of 1 million read per sample and per target capture reaction. The replicates were sequenced in different Illumina lanes.

2.6 | Read trimming and mapping

After quality checking of the raw Illumina data with FASTQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), the reads were trimmed with the program CONDETRI v2.2 (Smeds & Kunstner, 2011) using 20 as high-quality threshold parameter. Trimmed reads were mapped to the *G. undata* pseudoreference genome described above with BOWTIE2 v2.2.5 (Langmead & Salzberg, 2012) and the

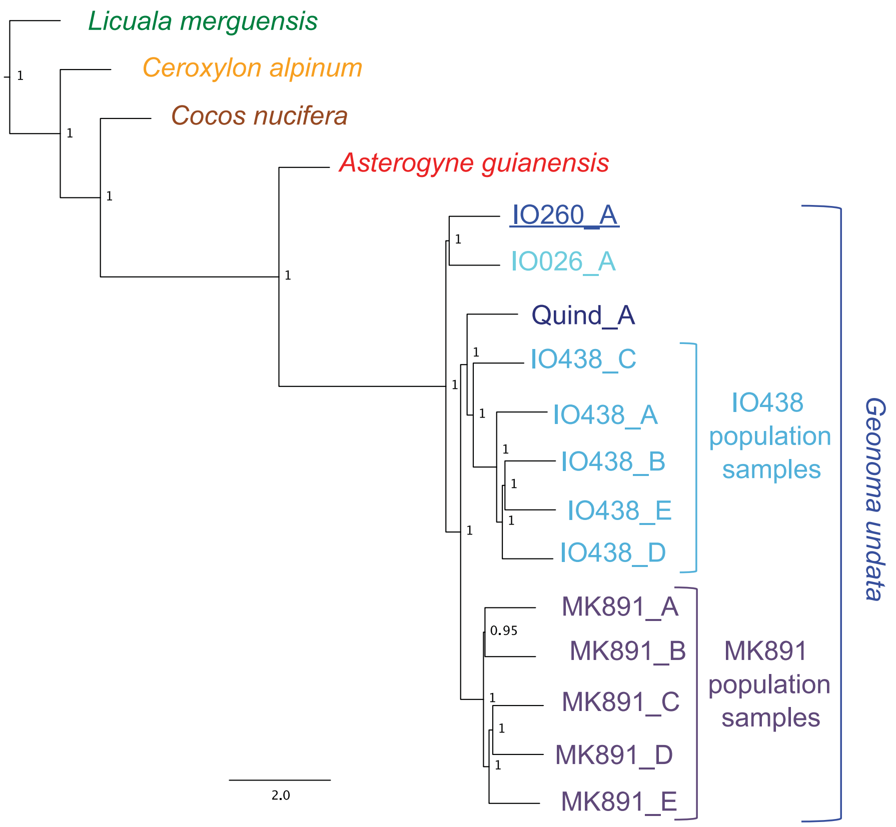

FIGURE 1 Coalescent-based phylogeny inferred from gene trees of the 17 studied samples, with branch support. The *G. undata* sample IO260_A used for whole genome sequencing is underlined. Each species is represented by one colour, and *G. undata* populations are represented by different shades of blue

very-sensitive-local option. Only reads mapping at a unique location in the genome were kept for analyses.

In order to test for the suitability of the method for different palm sub-families, all samples were also mapped to the two available palm genomes of *Elaeis guineensis* (NCBI assembly accession GCF_000442705.1) and *Phoenix dactylifera* (NCBI assembly accession GCF_000413155.1). For both genomes, the plastid genome was included in the reference. Filtered reads were mapped with BOWTIE2 v2.2.5 using the same parameters as described above.

2.7 | Specificity and global efficiency of the capture

Specificity (i.e., the proportion of IN-target reads) and efficiency (i.e., the proportion of baits covered by at least three reads) of the bait capture set were calculated for each sample based on uniquely mapped reads using BEDTOOLS v2.24.0 (Quinlan & Hall, 2010). Target regions were defined as the bait locations and their flanking 500 bases. To check the effect of sequencing effort on efficiency, uniquely mapped reads of the two replicates for each sample were merged and downsampled using Picard tools v1.119 (https://broadinstitute.github.io/picard). Downsampling reduced read number in increments of 100,000 reads from 100,000 to 1 million reads and in increments of 200,000 reads from 1 million reads to the maximum read number possible per sample. A total of 10 iterations were carried out per sample and downsampled read number.

Repeatability of the method was evaluated for all five intergeneric samples used for validation purposes here by comparing the two independent replicates for the average coverage per bait using BEDTOOLS v2.24.0 (Quinlan & Hall, 2010).

2.8 | Factors influencing capture efficiency per bait

The effects of several factors on bait efficiency were tested with a generalized linear model (GLM) using the package MASS in R v3.2.0 (R Core Team, 2015). Bait efficiency was calculated for each sample as the number of uniquely mapped reads mapping at least partially on the corresponding bait target region. Predictor variables were related to baits (GC content and bait density), to the targeted genomic regions (rates of molecular evolution) or to the sequenced samples (divergence time to the pseudoreference genome of *G. undata*). Only baits corresponding to exonic regions were used for these analyses, and intronic and non-genic regions were not included. GC content was calculated per bait with EMBOSS suite v6.6.0 (Rice, Longden, & Bleasby, 2000). Bait density represents the number of baits located 500 bp around the corresponding bait and was calculated with BEDTOOLS v2.24.0 (Quinlan & Hall, 2010). Rates of molecular evolution were estimated previously as the rate of nucleotide substitutions since divergence between the two species *G. undata* sample and *E. guineensis* (see paragraph 2.3). As a model with Poisson error distribution showed evidence of overdispersion, we used a negative-binomial family for the GLM analyses as recommended in Richards (2008). Significance of the GLM was calculated using type II analysis-of-variance using a likelihood ratio test.

2.9 | Variant detection

PCR duplicates were first masked with PICARD TOOLS v1.119 (https://broadinstitute.github.io/picard), and reads were realigned around indels and base-recalibrated using GATK v3.8 (McKenna et al., 2010). Both duplicates per sample were merged before variant calling. SNPs were then called for target regions and their surrounding 1,000 bp using UNIFIEDGENOTYPER of GATK v3.8 using the EMIT_ALL_SITES option. Samples were assigned within four groups of five samples: the intergeneric samples, the *G. undata* inter-population samples including five samples from five different populations, and the two *G. undata* intra-population samples. For the choice of *G. undata* inter-population samples, the individuals with the highest coverage were used to represent each population. For each group of five samples, we filtered sites with the following parameters: minimum quality >20, no indel and minimum depth of 10×. Variant sites were further filtered with a minimum of two alleles, and a minimum minor allele count of two. Two different missing data thresholds were tested: no missing data allowed and 20% missing allowed, thus representing one missing sample out of the five samples in each tested group. All filtering and depth calculations were performed with VCFTOOLS v0.1.13 (Danecek et al., 2011).

2.10 | Phylogenetic and population genetic analyses

To test the suitability of the detected SNPs both at micro- and macroscales, we estimated phylogenetic relationships using all samples and performed population genetic analyses with all *G. undata* samples (13 inter- and intra-population samples). Phylogenetic inference was performed using three maximum-likelihood methods and one coalescent-based method. First, we reconstructed a maximum-likelihood tree from the concatenated alignment of all sequences using the program RAXML v8.2.0 (Stamatakis, 2014) under the GTRGAMMA substitution model with 100 rapid bootstrap replicates. We then performed a second RAXML analysis on the alignment of all SNPs, applying the Lewis ascertainment bias correction to take into account the absence of constant sites. We also inferred phylogenetic relationships using an approximate maximum-likelihood approach developed for large sequence alignments implemented in the program FASTTREE v2.1.8 (Price, Dehal, & Arkin, 2010) with default parameters. However, concatenation does not deal with gene tree incongruences, which arises in the presence of incomplete lineage sorting and may result in high support for a wrong topology (Kubatko & Degnan, 2007). Therefore, we also used ASTRAL, a gene tree-based coalescent approach well suited for large genomics data sets (Mirarab & Warnow, 2015). We ran ASTRAL v5.6.1 with default parameters on the set of individual gene trees previously estimated with RAXML (under similar parameters to the concatenation analysis). Phylogenetic trees were visualized with the program FIGTREE v1.4.2 (https://tree.bio.ed.ac.uk/software/figtree/).

Then, a multi dimensional scaling (MDS) of the identity by descent matrix was used for model-free clustering of the *G. undata* inter- and intra-population samples. The identity by descent matrix was estimated using PLINK v1.07 (Purcell et al., 2007, https://pngu.mgh.harvard.edu/purcell/plink/). Finally, genomewide ancestry and potential admixture were estimated with ADMIXTURE v1.2365 using default parameters (Alexander, Novembre, & Lange, 2009).

3 | RESULTS

3.1 | Sequencing of a *G. undata* genome

More than 152.5 million reads were obtained for the WGS sample IO260_A, including 92.3% of high-quality reads that remained paired after trimming. A total of 44.1% of the trimmed reads mapped on the *E. guineensis* reference genome. When mapped onto the *G. undata* pseudoreference (see Section 2), mapping efficiency increased by 27.5% and reached 56.2% of overall alignment rate. In total, 94.5% of the annotated genes were covered with at least 3× coverage, highlighting the efficiency of the mapping strategy to recover gene sequences. Median coverage calculated with VCFTOOLS was 9×.

Our genome size estimates for the WGS-sequenced palm individual IO260_A resulted in an inferred genome size of approximately 3 giga bases (Gb; 3,042,066,005 bases). This genome size is consistent with the *C*-value estimate of 3.6 Gb for the *Geonoma interrupta* species, the only Geonoma species included in the Kew *C*-value database (Bennett & Leitch, 2012). Estimated repetitive content of the genome is 78.3% with an inferred homozygous k-mer coverage of 9.3×. This inferred value is highly concordant with the observed median coverage of 9×.

3.2 | Target capture Kit specification

A total of 59,264 baits of 120 bp each (“long baits”) were designed by the company Arbor Biosciences (Ann Arbor, MI, USA; formerly MYcroarray) to cover the 4.184 selected targets with a tiling of 2. The cumulative target size of our kit was 4,287,662 bp in total. The kit was named PopcornPalm in reference to our associated project “using POpulation genomics, Phylogenetics and COmmunity ecology to understand Radiations in Neotropical mountains” and is available under this name at Arbor Biosciences (Ann Arbor, MI, USA). The file with the sequences of the 59,264 baits (full PopcornPalm bait set) is available under Dryad number https://doi.org/10.5061/dryad.3v9v238.

In addition to this set and in order to give more flexibility to the palm community, we propose two additional bait sets PopcornPalm57K and PopcornPalm54K containing, respectively, 57,061 baits and 54,090 baits (Dryad number https://doi.org/10.5061/dryad.3v9v238). These two additional bait sets offer the palm community the possibility to (a) combine our PopcornPalm kit with the previous palm set of 2,909 baits presented in Heyduk et al. (2016) and already used for several palm projects and (b) use additionally a different commercial company for bait synthesis, including companies providing only 57 K bait kit synthesis instead of 60 K for Arbor Biosciences. The baits not included in the PopcornPalm57K and

PopcornPalm54K sets were identified based on their low bait efficiency ranking for the five intergeneric samples, counting the number of samples with only 3 reads or less for each bait. These two new bait sets include 3,990 and 3,910 target genes, respectively, instead of the 4,051 genes targeted in the full version of the kit. The presence/absence of each gene in these two additional bait sets is indicated in Supporting information Table S2.

3.3 | Mapping rates in different palm reference genomes

Overall mapping rates on the pseudoreference genome were high for all samples (Supporting information Table S3), ranging from 88.4% for *Asterogyne guianensis* to 92.9% for *G. undata*. Mapping to the *E. guineensis* and *P. dactylifera* genomes was also very successful for all samples (rates ranged from 80.9% to 93.8%). For each sample, mapping rates varied according to the reference, with higher mapping rates always observed when using the most closely related reference genome (Supporting information Table S3).

Mapping rates to the pseudoreference genome were not correlated with divergence time (Pearson's correlation: $t = -0.11$, $df = 8$, p-value = 0.92). For example, the second highest mapping rate was observed for *Licuala merguensis*, the most distant species to the *G. undata* pseudoreference genome (87 Myr, Baker & Couvreur, 2013). This samples also had higher mapping rates compared with the other samples for the two other reference genomes used. This result suggests that divergence time to the pseudoreference genome was not the main factor affecting mapping rate, and that other sample-specific parameters might be crucial for mapping efficacy.

3.4 | Specificity and global efficiency of the capture bait set

Specificity of the capture was high, with 82.3% of IN-target reads on average (range: 79.8%–84.2%). Percentage of IN-target reads significantly decreased with the genetic distance to the pseudoreference genome *G. undata* used to design the baits ($t = -3.72$, $df = 8$, p-value = 0.015, Pearson's correlation coefficient = –0.74). On average, only 0.02% of the reads mapped to the plastid genome (Supporting information Table S3).

Global efficiency of the method was also high for all samples with an average of 89.2% of the 59,264 baits covered with at least three reads. For the *G. undata* replicated samples IO260_A, 97% of the bait regions with low efficiency (only 3 reads or less) were recovered with the WGS sequencing and analyses of the same sample. This suggests that inefficiency of these regions was not due to deletions in the *G. undata* genome or to lower mapping efficiency.

After merging the two replicates per sample, downsampling of uniquely mapped reads was performed to evaluate the effect of sequencing effort on capture efficiency (Figure 2). Efficiency values were, on average, 58.8% with 100,000 reads per sample, which is the lowest number of reads tested. A rapid increase in efficiency was observed with read number of up to 1 million reads, and efficiency levelled off drastically after this read number. For all sequencing efforts tested, efficiency decreased with divergence time to the pseudoreference genome used to design the baits. For example, 85% global efficiency was reached with 900,000 uniquely mapped reads of *L. merguensis* and with only 400,000 reads of *G. undata*. These values can help to obtain optimal capture efficiency while reducing the cost of palm phylogenetic and population genetic studies.

To avoid bias due to sequencing effort, we compared global efficiency statistics on different genomic features using the downsampled bam files containing 1 million uniquely mapped reads per sample. For each sample, the number of mapping reads per bait was averaged between the 10 iterations of downsampling. Global efficiency for exons (96.3% of the baits) ranged from 87.2% for *L. merguensis* to 91.2% for *G. undata* and was not correlated with divergence time to the pseudoreference genome ($t = -1.62$, $df = 8$, p-value = 0.14). In all species, efficiency was lower for the baits covering introns (70.2% on average, range: 66.6%–72.7%). For baits designed on putative neutral regions outside genes, efficiency remained high in *G. undata* (96.3%), but significantly decreased with divergence time to the pseudoreference genome ($t = -5.49$, $df = 8$, p-value <0.001, Pearson's correlation coefficient = –0.89). For the more distant species *L. merguensis*, 40% of the non-genic regions were recovered despite the 87 Myr estimated divergence time to the pseudoreference genome used to design the baits (Figure 2).

3.5 | Factors influencing bait efficiency

Correlations of bait efficiency (i.e., the number of mapping reads per bait) between replicates were highly significant for all species, with p-values <0.001 and correlation coefficients ranging from 0.94 for *G. undata* to 0.98 for *L. merguensis* (Supporting information Figure S3). This result indicates the high reproducibility of our target capture method for the three palm sub-families tested. Correlation of bait efficiency between species decreased with divergence time (Supporting information Figure S4), with minimum correlation coefficients of 0.48 between *C. nucifera* and *L. merguensis* (Supporting information Figure S3).

All predictor variables tested (GC content, bait density, rates of molecular evolution of the gene and divergence time to the pseudoreference genome of *G. undata*) had significant effects on bait efficiency ($p < 0.001$ for all variables). Bait efficiency decreased with GC content lower than 20% and higher than 45% and in genes with higher rates of molecular evolution (Figure 3). As expected, bait efficiency increased with bait density, indicating that genes with long exons had higher sequencing quality. Finally, despite its statistical significance, divergence time to *G. undata* had little effect on bait efficiency (Figure 3). This result confirms the high global efficiency observed for all the samples tested.

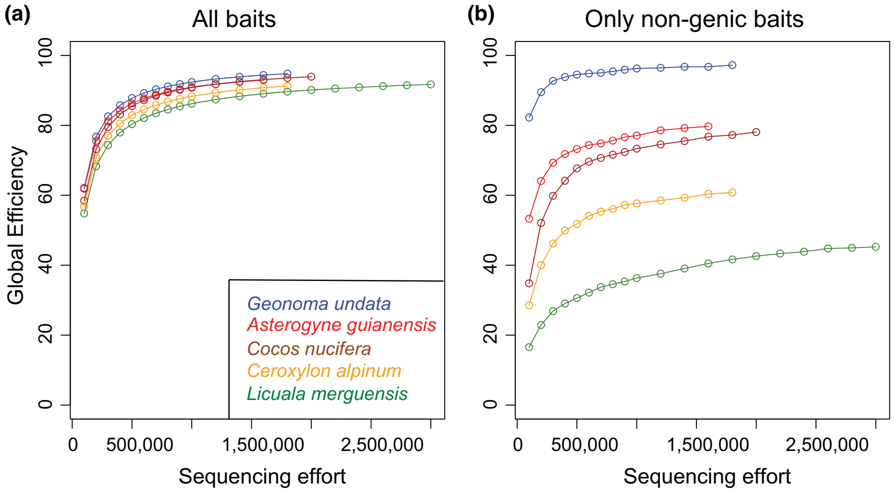

FIGURE 2 Effect of sequencing effort on global efficiency (i.e., the proportion of baits covered by at least three reads) for (a) all 59,264 baits, and (b) only 1,666 baits located in non-genic putatively neutral regions. Sequencing effort corresponds to the different subsampling of uniquely mapped reads. Colour codes for species correspond to Figure 1

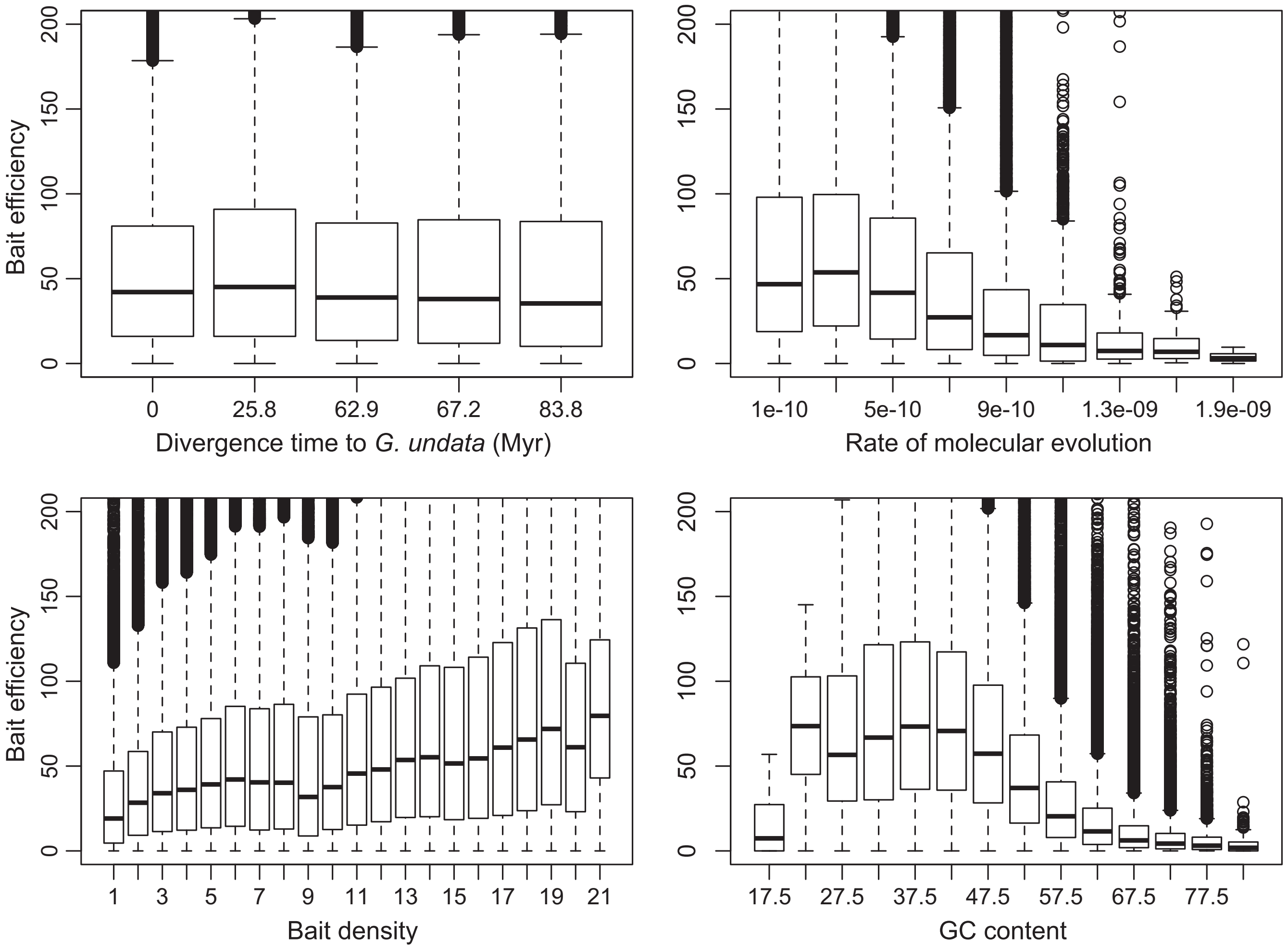

FIGURE 3 Effects of the divergence times to *G. undata*, rate of molecular evolution, bait density and GC content on bait efficiency (i.e., the number of read mapping on bait location). To gain visibility, the Y-axis was scaled to 200 and a part of the outliers is not visible in the figure (circles and close circles forming the black lines)

3.6 | Variant detection at the macro- and microscales

More than 3.8 million sites and 494,000 high-quality SNPs were obtained for the five intergeneric samples for a data set with a maximum of 20% of missing data per site, which represents only 4.4% missing data on average per sample (Table 1). The number of high-quality SNPs was >20,000 in both *G. undata* population data sets (5 samples per population). When considering the inter-population data set of five samples from five different populations, the number of SNPs doubled and reached 42,795 (Table 1). On average, 27% of the positions and 34.7% of the SNPs were located in regions surrounding baits, represented mainly by introns, UTRs and gene flanking regions. From 250 to 5,164, SNPs were detected in targeted non-genic regions.

Mean depth per SNP varied from 38.1× for the population IO348 to 50× for the intergeneric samples, suggesting that the sequencing effort was well adapted to high-quality data sets.

3.7 | Data informativeness for phylogenetic and population genetic analyses

The phylogeny inferred with the ASTRAL-II was strongly supported both at deep (support of 1 on average for intergeneric branches) and shallow scales (support values ranging from 0.95 to 1 for *G. undata* intraspecific branches; Figure 1). The phylogenetic trees inferred by RAXML were highly supported for the two concatenated data sets (all sequences or SNPs only, Supporting information Figures S5 and S6), exhibiting high bootstrap values at interspecific (bootstrap values of 100) and inter-population nodes (bootstrap values between 96 and 100 among *G. undata* populations). The phylogeny inferred with FASTTREE also had high local support values for interspecific and inter-population splits (local support values of 1 for these levels; Supporting information Figure S7). In all trees obtained with a maximum-likelihood method, within-population relationships had lower support (Supporting information Figures S5, S6 and S7). All methods produced similar topologies at the interspecific level.

At the intraspecific and population levels, MDS of identity by descent clustered the *G. undata* samples by geography and population origin. The two samples collected in Peru and Ecuador were separated from the Colombian samples on the 2nd axis (Figure 4a). The Colombian samples clustered by populations, with the exception of one sample of population IO438_C that appeared genetically intermediate between the other samples collected in the same location (Chamuscado, Colombia) and the sample Quind_A from Armenia (Colombia). The two sampling locations are located along the same mountain chain in western Colombia and are separated by ca. 280 km. Admixture analysis shows the same pattern and distinguished the five populations of *G. undata* represented by one to five samples (Figure 4b; Supporting information Figure S6). The sample IO438_C exhibited the same pattern as in the MDS figure and was detected with a distinct ancestry compared to the four remaining samples from this population. Results from Admixture suggest that this sample is admixed, exhibiting ancestry patterns from two different Colombian populations (Figure 4c).

4 | DISCUSSION

Genomewide patterns of speciation and radiation (i.e., genetic diversification across the full range of possible eco-evolutionary time scales) are starting to be characterized successfully in several widely used model systems such as cichlid fishes, sticklebacks, *Heliconius* butterflies, *Arabidopsis*, sunflowers or New World lupins (Badouin et al., 2017; Brawand et al., 2014; Dasmahapatra et al., 2012; Jones et al., 2012; Lamichhaney et al., 2015; Nevado, Atchison, Hughes, & Filatov, 2016; Novikova et al., 2016). Nevertheless, most current studies of this type are limited by a lack of molecular genetic data able to span a wide range of evolutionary distances (de La Harpe et al., 2017), and even high-profile genomic studies are clearly affected by sampling trade-offs. The palm family is no exception to this general pattern (Baker & Dransfield, 2016; Barrett, Bacon, Antonelli, Cano, & Hofmann, 2016). In fact, the need for greater genetic resolution at micro-evolutionary time scales (i.e., population genetics) is perfectly exemplified by palms: species delimitation has proved extremely challenging in many palm genera because of homoplasious morphological characters and/or high intraspecific morphological variation (Baker, Dransfield, & Hedderson, 2000; Hahn, 2002; Henderson, 2011). In the neotropical genus *Geonoma*, for example, 20% of the 68 species are in fact considered as species complexes and currently tentatively divided into 90 subspecies (Henderson, 2011), but only few studies at the population genetic level have been carried out (but see Roncal, Francisco-Ortega, &

TABLE 1 SNP detection and description at the intergeneric, inter-population and intra-population levels

	No. positions	% positions IN-bait	No. SNPs	% SNPs IN-bait	No. SNP in non-genic targeted regions	Average SNP depth
No missing data						
Intergeneric samples (5 ind.)	3,815,768	68.0	494,186	59.2	2,576	50
Inter-population samples (5 ind. from 5 different populations)	3,841,127	73.6	34,627	66.9	724	46.1
Population MK891 (5 ind.)	2,741,018	80.8	16,219	72.8	261	48.2
Population IO438 (5 ind.)	3,159,533	78.6	16,774	70.7	250	41.4
Maximum 20% missing data allowed						
Intergeneric samples (5 ind.)	4,896,285	63.5	634,554	55.1	5,164	45.9
Inter-population samples (5 ind. from 5 different populations)	4,724,849	67.9	42,795	61.6	985	41.9
Population MK891 (5 ind.)	3,562,856	76.2	20,561	68.2	321	42.8
Population IO438 (5 ind.)	3,765,068	75.3	20,009	67.5	278	38.1

Lewis, 2007; Borchsenius, Lozada, & Knudsen, 2016). Therefore, the development of genomic markers spanning a wide range of applications provides a unique opportunity to bring new insights into the mechanisms underlying the origin and maintenance of palm diversity both at the macro- and micro-evolutionary scales. Furthermore, one ambitious goal of the palm community is to reconstruct a species-level tree including all ca. 2,600 palm species (Baker & Dransfield, 2016; Faurby, Eiserhardt, Baker, & Svenning, 2016). A kit covering both high and low taxonomic levels (including closely related species within genera) will be highly valuable to achieve this goal.

4.1 | SNP detection and utility at both micro- and macroscales

We have presented a target capture kit for palms covering 4,051 genes and 133 non-genic putatively neutral regions. The high efficiency (i.e., the proportion of successful baits per sample) of our approach for three palm sub-families demonstrates the utility of the kit for species that diverged up to 87 Myr. With such high levels of efficiency for all three palm sub-families tested, our kit is expected to be useful for the entire palm (Arecaceae) family. We did not test its utility for other plant families and therefore recommend performing preliminary tests to evaluate the possibility to use this kit across families.

The selection of genes spanning different evolutionary rates, in addition to the successful sequencing of regions surrounding the targeted exons (e.g., introns and UTRs), facilitated the detection of more than 600,000 and 20,000 high-quality SNPs at the intergeneric and population levels, respectively. The phylogenetic trees inferred by different methods tested were all highly supported at the interspecific level. In addition, we were able to detect fine-scale genetic structure and gene flow at the inter- and intra-population levels. The set of 4,184 markers for target capture developed here is therefore a powerful new genomic tool for both phylogenomic and population genomic analyses in palms.

4.2 | WGS of a *Geonoma* palm genome to improve target choice and bait design

Oil and date palm genomes experienced a paleopolyploidization event with abundant oriented homeologous duplicated sequences between chromosomes (Singh et al., 2013). Genomes of diploid palms are also highly variable in size, ranging from 0.9 Gb for *Phoenix canariensis* to 13.9 Gb for *Pinanga subintegra* (Bennett & Leitch, 2012). These values highlight potentially pervasive copy number polymorphism in palm genomes. Methods of choice for CNV (copy number variation) detection usually incorporate a combination of depth of coverage, read orientation and fragment size deviation analysis of next-generation sequencing (NGS) data and require either a contiguously mappable reference genome or sample-reference pairing (Zhao et al., 2013). In palms, the high-quality reference genomes in addition to the multiple WGS data sets available for several samples of these reference species represent highly valuable resources for CNV analyses (Gros-Balthazard et al., 2017). However, these reference palm species have relatively small genome sizes (*C*-values of 0.95 Gb for *P. dactylifera* and 1.9 Gb for *E. guineensis*) compared to the 85 other species measured in Bennett and Leitch (2012; average of 3.6 Gb, range: 0.9–13.9 Gb). The limitations of smaller reference genomes for CNV detection were previously highlighted in conifers, as nine of the targets selected as low copy from the Loblolly pine

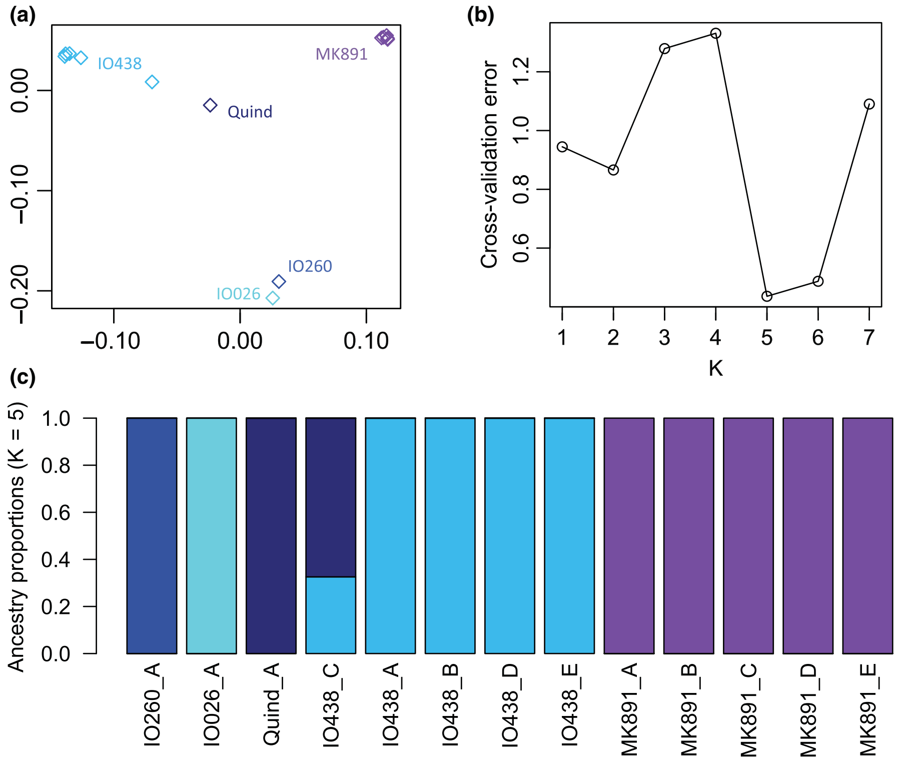

FIGURE 4 Intraspecific analyses. (a) multi dimensional scaling (MDS) graph for the 13 *G. undata* samples studied. Samples are colour coded by population, following the same colour code as in Figure 1. Population names are indicated in the proximity of the corresponding sample(s) using the same colour code. (b) Admixture cross-validation errors. (c) Ancestry proportions detected by Admixture for *K* = 5. Colour code is similar to Figure 1

genome were found highly repetitive in whitebark pine and attracted 54% of the IN-bait reads for this species (Syring et al., 2016). In our study, 8.5% of the targeted genes previously identified as low copy in palm reference genomes (Heyduk et al., 2016) were detected as putative multi-copy regions in *G. undata*. In total, 9.3% of the tested genes showed putative duplication patterns in *Geonoma* and were not considered as interesting targets for the capture. This result indicates that the sequencing of more palm species with different genome sizes would be greatly beneficial for the detection of gene families undergoing expansion in some palm genera and for avoiding them in marker development. The recent coconut reference genome (Xiao et al., 2017), in addition to our *G. undata* WGS sample IO260_A described in this study, is first step towards this end.

4.3 | Quality of the capture

For all species tested, we replicated the library preparation, target capture and sequencing to assess the reproducibility of the method. Correlation coefficients between replicate read coverage across baits were extremely high for all species (range: 0.94–0.98). Correlation between replicates was previously shown to be very high for human exome capture (>0.98%, Bainbridge et al., 2010; Bodi et al., 2013), and similar values were obtained for the three palm sub-families here, despite up to 87 Myr of divergence with *G. undata*, the species used to construct the baits.

The method was highly efficient for all three palm sub-families tested, with 86.2%–92.4% of bait regions successfully recovered per sample with a sequencing effort of 1 million uniquely mapped reads (about 1.1–1.2 million raw reads per sample). A drastic decrease in efficiency with divergence time was previously observed in other organisms (Bragg et al., 2016; Hedtke, Morgan, Cannatella, & Hillis, 2013) with an up to twofold decrease at 80 Myr divergence in skink lizards (Bragg et al., 2016). The effect was much smaller in palms with only 6.7% decrease in efficiency for the most distant species belonging to the Coryphoideae sub-family. Divergence time to *G. undata* affected mainly the recovery of non-genic targeted regions. It is however surprising to recover 40% of these non-genic regions in *L. merguensis* despite the 87 Myr divergence time with *G. undata*. Some of these non-genic regions might be more conserved in palms than expected, and neutrality tests should be performed before investigating processes such as gene flow, population connectivity or demography.

Specificity of the target capture kit ranged from 80% to 84% of IN-target reads for all species tested. Such specificity values are high compared to most previously published exome capture kits in palms (IN-target reads percentage ranging from 8.47% to 74.81% in Sabal species, Heyduk et al., 2016) or in non-model species (Puritz & Lotterhos, 2017). The only minor disadvantage of high specificity is the lack of OFF-target reads mapping to the plastid genome (average of 0.02% of the reads per samples only). The recovery of organelle genomes using OFF-target reads is common practice as organelles are traditional markers for phylogenetics (Heyduk et al., 2016; Singhal, Grundler, Colli, & Rabosky, 2017; Weitemier et al., 2014). Using 44 phylogenetically disparate reptile taxa, Singhal et al. (2017) showed that mitochondrial genome coverage was negatively correlated with capture specificity per sample. In our case, all species showed low levels of plastid genome recovery independently of specificity statistics or divergence time to *G. undata*. Plastid genomes can be of great interest for macro-evolutionary and biogeographic studies and were recently presented as a potentially universal “extended barcode” overcoming the limitations of traditionally used barcodes (Coissac, Hollingsworth, Lavergne, & Taberlet, 2016). One simple solution for recovering plastid genomes would be to use a genome skimming approach consisting of whole genome sequencing at low coverage (Straub et al., 2011, 2012). This method can be combined with target capture sequencing at low cost and time as the dual-indexed libraries are already prepared for all samples and already pooled at an equimolar ratio. The library preparations can be sequenced for genome skimming in the same sequencing run as the target capture reactions or in different sequencing runs when the target sequencing specificity statistics are of interest. In summary, our results point to the considerable amenability of target capture sequencing approaches to research questions aimed at bridging micro- and macro-evolutionary time scales.

ACKNOWLEDGEMENTS

We thank Michael Kessler from University of Zurich, Switzerland, Maria Jose Sanin from Medellin University CES, Colombia, and Henrik Balsev from Aarhus Botanical garden, Denmark, for sharing samples. We thank Alison Devault and Jacob Enk from Arbor Bioscience for bait design, valuable discussions and troubleshooting. Benoit Villain and Jonathan Rolland helped with DNA extractions and library preparations. We thank Thomas Gruet for IT and bioinformatics help. We thank Muriel Fragniere and Remy Bruggmann from the Next Generation Sequencing Platform in Bern for Illumina sequencing and demultiplexing, and the Genetic Diversity Centre (GDC) from ETH Zurich for computer resources. Finally, we thank the anonymous reviewers for their helpful comments. This work was funded by the Swiss National Science Foundation (SNF, Grant Number: CRSII3_147630).

AUTHOR CONTRIBUTION

Funding was secured by N.S. and C.L. The study was designed by M.P. in discussion with O.L., M.d.l.H., N.S. and C.L. Sample collection was coordinated by O.L. and N.S. The targets were selected and the molecular work was performed by M.d.l.H., O.L. and M.P. The data were analysed by M.d.l.H., J.H. and M.P.. The manuscript was written by M.P. with significant input from all co-authors. C.L. and M.P. should be considered joint senior author.

DATA ACCESSIBILITY

Whole genome sequencing and targeted sequencing sequence reads generated as part of this manuscript are available in NCBI (project

PRJNA482221). The three different probe set files are available in Dryad (https://doi.org/10.5061/dryad.3v9v238).

ORCID

Christian Lexer http://orcid.org/0000-0002-7221-7482
Margot Paris http://orcid.org/0000-0001-7328-3820

REFERENCES

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*, 1655–1664. https://doi.org/10.1101/gr.094052.109

Bacon, C. D., Feltus, F. A., Paterson, A. H., & Bailey, C. D. (2008). Novel nuclear intron-spanning primers for Arecaceae evolutionary biology. *Molecular Ecology Resources*, *8*, 211–214. https://doi.org/10.1111/j.1471-8286.2007.01928.x

Badouin, H., Gouzy, J., Grassa, C. J., Murat, F., Staton, S. E., Cottret, L., … Legrand, L. (2017). The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature*, *546*, 148. https://doi.org/10.1038/nature22380

Bainbridge, M. N., Wang, M., Burgess, D. L., Kovar, C., Rodesch, M. J., D'Ascenzo, M., … Jeddeloh, J. A. (2010). Whole exome capture in solution with 3 Gbp of data. *Genome Biology*, *11*, R62.

Baker, W. J., & Couvreur, T. L. P. (2013). Global biogeography and diversification of palms sheds light on the evolution of tropical lineages. I. Historical Biogeography. *Journal of Biogeography*, *40*, 274–285. https://doi.org/10.1111/j.1365-2699.2012.02795.x

Baker, W. J., & Dransfield, J. (2016). Beyond Genera Palmarum: Progress and prospects in palm systematics. *Botanical Journal of the Linnean Society*, *182*, 207–233.

Baker, W. J., Dransfield, J., & Hedderson, T. A. (2000). Phylogeny, character evolution, and a new classification of the calamoid palms. *Systematic Botany*, *25*, 297–322. https://doi.org/10.2307/2666644

Balslev, H., Bernal, R., & Fay, M. F. (2016). Palms - emblems of tropical forests. *Botanical Journal of the Linnean Society*, *182*, 195–200. https://doi.org/10.1111/boj.12465

Balslev, H., Kahn, F., Millan, B., Svenning, J. C., Kristiansen, T., Borchsenius, F., … Eiserhardt, W. L. (2011). Species diversity and growth forms in tropical American palm communities. *Botanical Review*, *77*, 381–425. https://doi.org/10.1007/s12229-011-9084-x

Barrett, C. F., Bacon, C. D., Antonelli, A., Cano, A., & Hofmann, T. (2016). An introduction to plant phylogenomics with a focus on palms. *Botanical Journal of the Linnean Society*, *182*, 234–255. https://doi.org/10.1111/boj.12399

Bennett, M. D., Leitch, I. J. (2012) Angiosperm DNA C-values database (release 8.0, Dec. 2012) http://www.kew.org/cvalues/.

Bodi, K., Perera, A. G., Adams, P. S., Bintzler, D., Dewar, K., Grove, D. S., … Singh, S. (2013). Comparison of commercially available target enrichment methods for next-generation sequencing. *J Biomol Tech*, *24*, 73–86. https://doi.org/10.7171/jbt.13-2402-002

Borchsenius, F., Lozada, T., & Knudsen, J. T. (2016). Reproductive isolation of sympatric forms of the understorey palm *Geonoma macrostachys* in western Amazonia. *Botanical Journal of the Linnean Society*, *182*, 398–410.

Bragg, J. G., Potter, S., Bi, K., & Moritz, C. (2016). Exon capture phylogenomics: Efficacy across scales of divergence. *Molecular Ecology Resources*, *16*, 1059–1068. https://doi.org/10.1111/1755-0998.12449

Brawand, D., Wagner, C. E., Li, Y. I., Malinsky, M., Keller, I., Fan, S., … Turner-Maier, J. (2014). The genomic substrate for adaptive radiation in African cichlid fish. *Nature*, *513*, 375.

Buddenhagen, C., Lemmon, A. R., Lemmon, E. M., Bruhl, J., Cappa, J., Clement, W. L., … Mitchell, N. (2016). Anchored phylogenomics of angiosperms I: Assessing the robustness of phylogenetic estimates. *bioRxiv*, 086298.

Chamala, S., García, N., Godden, G. T., Krishnakumar, V., Jordon-Thaden, I. E., De Smet, R., … Soltis, P. S. (2015). MARKERMINER 1.0: A new application for phylogenetic marker development using angiosperm transcriptomes. *Applications in Plant Sciences*, *3*, 1400115. https://doi.org/10.3732/apps.1400115

Coissac, E., Hollingsworth, P. M., Lavergne, S., & Taberlet, P. (2016). From barcodes to genomes: Extending the concept of DNA barcoding. *Molecular Ecology*, *25*, 1423–1428. https://doi.org/10.1111/mec.13549

Core Team, R. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/

Couvreur, T. L. P., & Baker, W. J. (2013). Tropical rain forest evolution: Palms as a model group. *BMC Biology*, *11*, 48. https://doi.org/10.1186/1741-7007-11-48

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … McVean, G. (2011). The variant call format and VCFTOOLS. *Bioinformatics*, *27*, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Dasmahapatra, K. K., Walters, J. R., Briscoe, A. D., Davey, J. W., Whibley, A., Nadeau, N. J., … Salazar, C. (2012). Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, *487*, 94–98. https://doi.org/10.1038/nature11041

de la Harpe, M., Paris, M., Karger, D. N., Rolland, J., Kessler, M., Salamin, N., & Lexer, C. (2017). Molecular ecology studies of species radiations: Current research gaps, opportunities and challenges. *Molecular Ecology*, *26*, 2608–2622. https://doi.org/10.1111/mec.14110

Djedatin, G., Monat, C., Engelen, S., & Sabot, F. (2017). DuplicationDetector, a light weight tool for duplication detection using NGS data. *Current Plant Biology*, *9–10*, 23–28. https://doi.org/10.1016/j.cpb.2017.07.001

Faircloth, B. C., Branstetter, M. G., White, N. D., & Brady, S. G. (2015). Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. *Molecular Ecology Resources*, *15*, 489–501. https://doi.org/10.1111/1755-0998.12328

Faurby, S., Eiserhardt, W. L., Baker, W. J., & Svenning, J. C. (2016). An all-evidence species-level supertree for the palms (Arecaceae). *Molecular Phylogenetics and Evolution*, *100*, 57–69. https://doi.org/10.1016/j.ympev.2016.03.002

Futuyma, D. J., & Kirkpatrick, M. (2017). *Evolution*, 4th ed. Sunderland, MA: Sinauer Associates.

Gan, X., Stegle, O., Behr, J., Steffen, J. G., Drewe, P., Hildebrand, K. L., … Kahles, A. (2011). Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature*, *477*, 419–423. https://doi.org/10.1038/nature10414

Glenn, T. C., & Faircloth, B. C. (2016). Capturing Darwin's dream. *Molecular Ecology Resources*, *16*, 1051–1058. https://doi.org/10.1111/1755-0998.12574

Gros-Balthazard, M., Galimberti, M., Kousathanas, A., Newton, C., Ivorra, S., Paradis, L., … Santoni, S. (2017). The discovery of wild date palms in Oman reveals a complex domestication history involving centers in the Middle East and Africa. *Current Biology*, *27*, 2211–2218+. https://doi.org/10.1016/j.cub.2017.06.045

Hahn, W. J. (2002). A phylogenetic analysis of the Arecoid Line of palms based on plastid DNA sequence data. *Molecular Phylogenetics and Evolution*, *23*, 189–204. https://doi.org/10.1016/S1055-7903(02)00022-2

Hartl, D. L., & Clark, A. G. (2007). *Principles of Population Genetics*, 4th ed. Sunderland, MA: Sinauer Associates, Inc.

Hedtke, S. M., Morgan, M. J., Cannatella, D. C., & Hillis, D. M. (2013). Targeted enrichment: Maximizing orthologous gene comparisons across deep evolutionary time. *Plos One*, *8*, 7. https://doi.org/10.1371/journal.pone.0067908

Henderson, A. (2011). A revision of Geonoma (Arecaceae). *Phytotaxa*, *17*, 1–271. https://doi.org/10.11646/phytotaxa.17.1.1

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

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., … Birney, E. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, *484*, 55–61. https://doi.org/10.1038/nature10944

Jukes, T. H., & Cantor, C. R. (1969). Evolution of protein molecules. In H. N. Munro (Ed.), *Mammalian Protein Metabolism* (pp. 21–132). New York: Academic Press.

Kircher, M., Sawyer, S., & Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Research*, *40*, 1. https://doi.org/10.1093/nar/gkr771

Kubatko, L. S., & Degnan, J. H. (2007). Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology*, *56*, 17–24. https://doi.org/10.1080/10635150601146041

Lamichhaney, S., Berglund, J., Almén, M. S., Maqbool, K., Grabherr, M., Martinez-Barrio, A., … Grant, B. R. (2015). Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature*, *518*, 371–375. https://doi.org/10.1038/nature14181

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*, 357–U354.

Lemmon, E. M., & Lemmon, A. R. (2013). High-throughput genomic data in systematics and phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, *44*, 99–121. https://doi.org/10.1146/annurev-ecolsys-110512-135822

Lewis, C. E., & Doyle, J. J. (2001). Phylogenetic utility of the nuclear gene malate synthase in the palm family (Arecaceae). *Molecular Phylogenetics and Evolution*, *19*, 409–420. https://doi.org/10.1006/mpev.2001.0932

Lewis, C. E., & Doyle, J. J. (2002). A phylogenetic analysis of tribe Areceae (Arecaceae) using two low-copy nuclear genes. *Plant Systematics and Evolution*, *236*, 1–17.

Ludeña, B., Chabrillange, N., Aberlenc-Bertossi, F., Adam, H., Tregear, J. W., & Pintaud, J. C. (2011). Phylogenetic utility of the nuclear genes AGAMOUS 1 and PHYTOCHROME B in palms (Arecaceae): An example within Bactridinae. *Annals of Botany*, *108*, 1433–1444. https://doi.org/10.1093/aob/mcr191

Marcais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, *27*, 764–770. https://doi.org/10.1093/bioinformatics/btr011

McCartney-Melstad, E., Mount, G. G., & Shaffer, H. B. (2016). Exon capture optimization in amphibians with large genomes. *Molecular Ecology Resources*, *16*, 1084–1094. https://doi.org/10.1111/1755-0998.12538

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*, 1297–1303. https://doi.org/10.1101/gr.107524.110

Meerow, A. W., Noblick, L., Borrone, J. W., Couvreur, T. L., Mauro-Herrera, M., Hahn, W. J., … Schnell, R. J. (2009). Phylogenetic analysis of seven WRKY genes across the palm Subtribe Attaleinae (Areceaceae) identifies Syagrus as sister group of the Coconut. *Plos One*, *4*, 10.

Meyer, M., & Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc 2010*, pdb prot5448.

Mirarab, S., & Warnow, T. (2015). ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, *31*, 44–52. https://doi.org/10.1093/bioinformatics/btv234

Nevado, B., Atchison, G. W., Hughes, C. E., & Filatov, D. A. (2016). Widespread adaptive evolution during repeated evolutionary radiations in New World lupins. *Nature Communications*, *7*, 12384. https://doi.org/10.1038/ncomms12384

Nicholls, J. A., Pennington, R. T., Koenen, E. J., Hughes, C. E., Hearn, J., Bunnefeld, L., … Kidner, C. A. (2015). Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the neotropical rain forest genus Inga (Leguminosae: Mimosoideae). *Frontiers in Plant Science*, *6*, 710. https://doi.org/10.3389/fpls.2015.00710

Novikova, P. Y., Hohmann, N., Nizhynska, V., Tsuchimatsu, T., Ali, J., Muir, G., … Holm, S. (2016). Sequencing of the genus Arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nature Genetics*, *48*, 1077. https://doi.org/10.1038/ng.3617

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FASTTREE 2-approximately maximum-likelihood trees for large alignments. *Plos One*, *5*, 3. https://doi.org/10.1371/journal.pone.0009490

Prum, R. O., Berv, J. S., Dornburg, A., Field, D. J., Townsend, J. P., Lemmon, E. M., & Lemmon, A. R. (2015). A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature*, *526*, 569–U247.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., … Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*, 559–575. https://doi.org/10.1086/519795

Puritz, J. B., & Lotterhos, K. E. (2017). Expressed exome capture sequencing (EecSeq): A method for cost-effective exome sequencing for all organisms with or without genomic resources. *Biorxiv*, 223735.

Quinlan, A. R., & Hall, I. M. (2010). BEDTOOLS: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: The European molecular biology open software suite. *Trends in Genetics*, *16*, 276–277. https://doi.org/10.1016/S0168-9525(00)02024-2

Richards, S. A. (2008). Dealing with overdispersed count data in applied ecology. *Journal of Applied Ecology*, *45*, 218–227. https://doi.org/10.1111/j.1365-2664.2007.01377.x

Rolland, J., Silvestro, D., Litsios, G., Faye, L., & Salamin, N. (2018). Clownfishes evolution below and above the species level. *Proceedings of the Royal Society B: Biological Sciences*, *285*, 1863. https://doi.org/10.1098/rspb.2017.1796

Roncal, J., Blach-Overgaard, A., Borchsenius, F., Balslev, H., & Svenning, J. C. (2011). A dated phylogeny complements macroecological analysis to explain the diversity patterns in geonoma (Arecaceae). *Biotropica*, *43*, 324–334. https://doi.org/10.1111/j.1744-7429.2010.00696.x

Roncal, J., Francisco-Ortega, J., Asmussen, C. B., & Lewis, C. E. (2005). Molecular phylogenetics of tribe geonomeae (Arecaceae) using nuclear DNA sequences of phosphoribulokinase and RNA polymerase II. *Systematic Botany*, *30*, 275–283. https://doi.org/10.1600/0363644054223620

Roncal, J., Francisco-Ortega, J., & Lewis, C. E. (2007). An evaluation of the taxonomic distinctness of two *Geonoma macrostachys* (Arecaceae) varieties based on intersimple sequence repeat (ISSR) variation. *Botanical Journal of the Linnean Society*, *153*, 381–392. https://doi.org/10.1111/j.1095-8339.2007.00619.x

Roser, M., Johnson, M. A. T., & Hanson, L. (1997). Nuclear DNA amounts in palms (Arecaceae). *Botanica Acta*, *110*, 79–89. https://doi.org/10.1111/j.1438-8677.1997.tb00614.x

Sarver, B. A., Keeble, S., Cosart, T., Tucker, P. K., Dean, M. D., & Good, J. M. (2017). Phylogenomic insights into mouse evolution using a pseudoreference approach. *Genome Biology and Evolution*, *9*, 726–739. https://doi.org/10.1093/gbe/evx034

Sass, C., Iles, W. J. D., Barrett, C. F., Smith, S. Y., & Specht, C. D. (2016). Revisiting the Zingiberales: Using multiplexed exon capture to resolve ancient and recent phylogenetic splits in a charismatic plant lineage. *PeerJ*, *4*, e1584. https://doi.org/10.7717/peerj.1584

Schmid, S., Genevest, R., Gobet, E., Suchan, T., Sperisen, C., Tinner, W., & Alvarez, N. (2017). HyRAD-X, a versatile method combining exome capture and RAD sequencing to extract genomic information from ancient DNA. *Methods in Ecology and Evolution*, *8*, 1374–1388. https://doi.org/10.1111/2041-210X.12785

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., … Weinstock, G. M. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, *15*, 1034–1050. https://doi.org/10.1101/gr.3715005

Singh, R., Ong-Abdullah, M., Low, E. T. L., Manaf, M. A. A., Rosli, R., Nookiah, R., … Azizi, N. (2013). Oil palm genome sequence reveals divergence of interfertile species in Old and New Worlds. *Nature*, *500*, 335-+.

Singhal, S., Grundler, M., Colli, G., & Rabosky, D. L. (2017). Squamate Conserved Loci (SqCL): A unified set of conserved loci for phylogenomics and population genetics of squamate reptiles. *Molecular Ecology Resources*, *17*, e12–e24.

Smeds, L., & Kunstner, A. (2011). CONDETRI—A content dependent read trimmer for Illumina Data. *Plos One*, *6*, 10. https://doi.org/10.1371/journal.pone.0026314

Stamatakis, A. (2014). RAXML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stephens, J. D., Rogers, W. L., Heyduk, K., Cruse-Sanders, J. M., Determann, R. O., Glenn, T. C., & Malmberg, R. L. (2015). Resolving phylogenetic relationships of the recently radiated carnivorous plant genus Sarracenia using target enrichment. *Molecular Phylogenetics and Evolution*, *85*, 76–87.

Stephens, J. D., Rogers, W. L., Mason, C. M., Donovan, L. A., & Malmberg, R. L. (2015). Species tree estimation of diploid Helianthus (Asteraceae) using target enrichment. *American Journal of Botany*, *102*, 910–920.

Straub, S. C., Fishbein, M., Livshultz, T., Foster, Z., Parks, M., Weitemier, K., … Liston, A. (2011). Building a model: Developing genomic resources for common milkweed (*Asclepias syriaca*) with low coverage genome sequencing. *BMC Genomics*, *12*, 211. https://doi.org/10.1186/1471-2164-12-211

Straub, S. C., Parks, M., Weitemier, K., Fishbein, M., Cronn, R. C., & Liston, A. (2012). Navigating the tip of the Genomic Iceberg: Next-generation sequencing for plant systematics. *American Journal of Botany*, *99*, 349–364. https://doi.org/10.3732/ajb.1100335

Suchan, T., Pitteloud, C., Gerasimova, N. S., Kostikova, A., Schmid, S., Arrigo, N., … Alvarez, N. (2016). Hybridization capture using RAD Probes (hyRAD), a new tool for performing genomic analyses on collection specimens. *Plos One*, *11*, 3. https://doi.org/10.1371/journal.pone.0151651

Sun, H. Q., Ding, J., Piednoel, M., & Schneeberger, K. (2018). findGSE: Estimating genome size variation within human and Arabidopsis using k-mer frequencies. *Bioinformatics*, *34*, 550–557. https://doi.org/10.1093/bioinformatics/btx637

Suren, H., Hodgins, K. A., Yeaman, S., Nurkowski, K. A., Smets, P., Rieseberg, L. H., … Holliday, J. A. (2016). Exome capture from the spruce and pine giga-genomes. *Molecular Ecology Resources*, *16*, 1136–1146. https://doi.org/10.1111/1755-0998.12570

Syring, J. V., Tennessen, J. A., Jennings, T. N., Wegrzyn, J., Scelfo-Dalbey, C., & Cronn, R. (2016). Targeted capture sequencing in whitebark pine reveals range-wide demographic and Adaptive patterns despite challenges of a large, repetitive genome. *Frontiers in Plant Science*, *7*, 484. https://doi.org/10.3389/fpls.2016.00484

Thompson, O. A., Snoek, L. B., Nijveen, H., Sterken, M. G., Volkers, R. J., Brenchley, R., … Hajnal, A. (2015). Remarkably divergent regions punctuate the genome assembly of the *Caenorhabditis elegans* Hawaiian Strain CB4856. *Genetics*, *200*, 975.

Uribe-Convers, S., Settles, M. L., & Tank, D. C. (2016). A Phylogenomic approach based on PCR target enrichment and high throughput sequencing: Resolving the diversity within the South American species of *Bartsia* L. *(Orobanchaceae). Plos One*, *11*, 2. https://doi.org/10.1371/journal.pone.0148203

Weitemier, K., Straub, S. C., Cronn, R. C., Fishbein, M., Schmickl, R., McDonnell, A., & Liston, A. (2014). Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences*, *2*, 1400042. https://doi.org/10.3732/apps.1400042

Xiao, Y., Xu, P., Fan, H., Baudouin, L., Xia, W., Bocs, S., … Li, J. (2017). The genome draft of coconut (*Cocos nucifera*). *Gigascience*, *6*, 1–1. https://doi.org/10.1093/gigascience/gix095

Zhao, M., Wang, Q. G., Wang, Q., Jia, P. L., & Zhao, Z. M. (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: Features and perspectives. *BMC Bioinformatics*, *14*, S1. https://doi.org/10.1186/1471-2105-14-S11-S1

Zhou, L. C., & Holliday, J. A. (2012). Targeted enrichment of the black cottonwood (*Populus trichocarpa*) gene space using sequence capture. *BMC Genomics*, *13*, 703. https://doi.org/10.1186/1471-2164-13-703

SUPPORTING INFORMATION

Additional supporting information may be found online