MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2014) **14**, 966–975 doi: 10.1111/1755-0998.12246

Genome skimming by shotgun sequencing helps resolve the phylogeny of a pantropical tree family

PIERRE-JEAN G. MALÉ,*† LÉA BARDON,* GUILLAUME BESNARD,* ERIC COISSAC,‡ FRÉDÉRIC DELSUC,§ JULIEN ENGEL,¶ EMELINE LHUILLIER,**†† CAROLINE SCOTTI-SAINTAGNE,¶ ALEXANDRA TINAUT¶ and JÉRÔME CHAVE*

*UMR 5174 Laboratoire Évolution & Diversité Biologique, CNRS, Université Paul Sabatier, ENFA, 118 route de Narbonne, F-31062 Toulouse, France, †Department of Ecology & Evolutionary Biology, University of Toronto, Toronto, ON M5S 3G5, Canada, ‡Laboratoire d'Ecologie Alpine CNRS, UMR5553, Université Joseph Fourier, BP 53, F-38041 Grenoble Cedex 9, France, §Institut des Sciences de l'Evolution, UMR 5554-CNRS, Université Montpellier 2, Place Eugène Bataillon, Montpellier, France, ¶UMR ECOFOG, INRA, Université Antilles-Guyane, CNRS, CIRAD, AgroParisTech, Campus agronomique, BP 709, F-97387 Kourou Cedex, France, **INRA, UAR 1209 Département de Génétique Animale, INRA Auzeville, F-31326 Castanet-Tolosan, France, ††GeT-PlaGe, Genotoul, INRA Auzeville, F-31326 Castanet-Tolosan, France*

Abstract

Whole genome sequencing is helping generate robust phylogenetic hypotheses for a range of taxonomic groups that were previously recalcitrant to classical molecular phylogenetic approaches. As a case study, we performed a shallow shotgun sequencing of eight species in the tropical tree family Chrysobalanaceae to retrieve large fragments of high-copy number DNA regions and test the potential of these regions for phylogeny reconstruction. We were able to assemble the nuclear ribosomal cluster (nrDNA), the complete plastid genome (ptDNA) and a large fraction of the mitochondrial genome (mtDNA) with approximately 1000×, 450× and 120× sequencing depth respectively. The phylogenetic tree obtained with ptDNA resolved five of the seven internal nodes. In contrast, the tree obtained with mtDNA and nrDNA data were largely unresolved. This study demonstrates that genome skimming is a cost-effective approach and shows potential in plant molecular systematics within Chrysobalanaceae and other under-studied groups.

Keywords: next-generation sequencing, organellar genome, phylogenomics, tropical trees

Received 30 October 2013; revision received 9 February 2014; accepted 22 February 2014

Introduction

Over the past two decades, plant systematics has undergone a major revolution, owing to the advent of methods for inferring evolutionary relationships among species based on DNA sequences (Chase *et al.* 1993; Soltis *et al.* 2007, 2011). These advances have been obtained through DNA amplification by polymerase chain reaction (PCR) followed by classical Sanger sequencing. Organellar genome material has the advantage to be clonally inherited; hence technical issues related to heterozygosity and recombination or polyploidy are circumvented. In addition, the large amount of DNA copies for organellar genomes helps improve the efficiency of DNA extraction, a crucial stage when using degraded DNA, e.g. in herbarium samples.

In spite of this impressive progress, the phylogenetic tree of flowering plants is far from being fully resolved (Soltis *et al.* 2011). This is mostly due to four major issues. First, bursts of diversification may prevent the resolution of phylogenetic trees if based on too few phylogenetically informative sites (Magallon *et al.* 1999; Davis *et al.* 2005; Whitfield & Lockhart 2007; Davis & Anderson 2010). Second, plant groups have received an unequal consideration when it comes to molecular systematics, and even today, some lineages remain poorly investigated (e.g. Panadanaceae, Buerki *et al.* 2012; or Sapotaceae, Swenson *et al.* 2008). Third, a number of plants, especially those growing in the tropics and subtropics, produces massive amounts of secondary compounds used for defending their tissues, and extracting high-quality DNA for molecular analyses is challenging (for example, Feitosa *et al.* 2012). Moreover, PCR is notoriously fastidious on DNA extracted from herbarium specimens and is

Correspondence: Pierre-Jean Malé, Fax: +33 (0) 5 61 55 73 27; E-mail: pjg.male@gmail.com

usually only applicable for short segments of high-copy number DNA regions (Varma *et al*. 2007).

The lack of resolution of phylogenetic trees can be overcome by increasing drastically the amount of phylogenetically informative signal (Delsuc *et al*. 2005). For plants, the first application of this idea has shed light on the basal relationships among angiosperms (Goremykin *et al*. 2004; Leebens-Mack *et al*. 2005; Moore *et al*. 2007). A major advance in this direction has since been made possible through the advent of high throughput DNA sequencing (Atherton *et al*. 2010; Egan *et al*. 2012) and the production of plastid genomes at an unprecedented speed (Moore *et al*. 2006). In a few years, it was extended to all major eudicot clades (Moore *et al*. 2010) and to monocotyledons (Givnish *et al*. 2010). Most of these studies relied on the purification of plastid DNA prior to sequencing, which is a challenging task that needs specific, expensive equipment. Moreover, they were limited to the use of only plastid exons either because of the need of designing conserved primers for PCR amplification or because of errors generated in homopolymers by 454 sequencers which are more frequent in noncoding regions.

Here, we implement a rapid and cost-effective strategy for generating phylogenetically informative genomic data. In a nutshell, the idea is to perform a 'genome skimming' approach (see Cronn *et al*. 2008; Straub *et al*. 2012; McPherson *et al*. 2013) by a shallow (typically 1×) shotgun sequencing of the total genomic DNA. In addition to producing reliable and entire plastid genomes, we explored whether we could also generate full sequences of the nuclear ribosomal cluster and at least partial sequences of the mitochondrial genome. As a proof of concept, we applied this method to eight species belonging to Chrysobalanaceae, an ecologically important family in the Neotropics (Prance 1972; Prance & White 1988; Hopkins 2007). This family has long been overlooked in molecular phylogenetics (but see Yakandawala *et al*. 2010), in part because it is rich in secondary chemical compounds (Feitosa *et al*. 2012). We show that even for this family with limited available molecular systematics information the entire plastid genome, the nuclear ribosomal unit and at least a part of the mitochondrial genome can be obtained from the same sequencing run. This new information yields better-resolved phylogenetic trees than those based on transcribed sequences only.

Materials and methods

Study species

Chrysobalanaceae is a pantropical woody plant family in the order Malpighiales (Prance & White 1988; Davis *et al*. 2005; Apg III 2009). The most recent taxonomic treatment of the family includes 531 species and 18 genera (Prance & Sothers 2003). Over 80% of the species occur in the Neotropics where they are often an important component of the woody plant community, both in terms of ecology and diversity (Prance 1972; Prance & White 1988; Hopkins 2007). We focused on eight species belonging to seven of the 12 major monophyletic groups described in Bardon *et al*. (2013). We included three *Licania* species: *L. sprucei* (Hook.f.) Fritsch (clade L3), *L. heteromorpha* Benth. (clade L4), and *L. alba* (Bernoulli) Cuatrec. (clade L5). We also included *Parinari campestris* Aubl., *Chrysobalanus icaco* L., *Couepia guianensis* Benth., and two species of the genus *Hirtella* (*H. physophora* Mart. & Zucc. and *H. racemosa* Lam.).

Fresh leaf tissue samples were immediately frozen in liquid nitrogen and stored at −80 °C. We also used silica-dried leaf tissue stored at ambient temperature for one species (*L. heteromorpha*). All leaf tissues were collected from well-defined populations near Kourou, French Guiana (locations 5°04′N; 52°42′W and 5°23′N; 52°54′W).

DNA extraction and library preparation

Leaves were ground in a ZM 200 centrifugal mill (Retsch, Hahn, Germany) through a 0.5 mm sieve, in liquid nitrogen and at full speed. We then used a TissueLyser II (Qiagen, Courtaboeuf, France) to crush 10 mg of frozen ground tissue with three glass beads (4 mm in diameter) in a 2-mL microtube for two rounds of 1 min at 30 Hz separated by a 45-s pause to prevent overheating of the samples. Total DNA was then extracted using a DNeasy Plant Mini Kit (Qiagen) following the manufacturer's protocol.

Five separate extractions were performed for each plant, pooled and concentrated in a Speed Vac centrifuge (Savant Instruments, Inc., Hicksville, NY, USA) to a total volume of 100 μL. Quality and concentration of the DNA samples were assessed on a 1% agarose gel and with a Nanodrop ND-1000 (Thermo Fisher Scientific, Delaware). We also applied the PicoGreen method (Quant-iT™PicoGreen®; Molecular Probes Inc., OR, USA) to measure the concentration of double-stranded DNA (Murakami & McCaman 1999).

Library construction and sequencing was carried out on a HiSeq 2000 Illumina sequencing system (Illumina, San Diego, CA, USA). A separate library was constructed for each species using the Illumina TruSeq DNA Sample Prep v2 kit. Briefly, genomic DNA was sheared by sonication and the fragments were end-repaired. Purified fragments were A-tailed and ligated to sequencing indexed adapters. Fragments with an insert size of approximately 400 bp were gel-extracted and enriched with 10 cycles of PCR for blunting DNA molecules before library quantification and validation. Each library was then hybridized to the HiSeq 2000 flow cell using the Illumina TruSeq PE Cluster Kit v3. Bridge amplification was performed to generate clusters, and paired reads of 101 nucleotides were collected on the HiSeq 2000 sequencer, using the Illumina TruSeq SBS Kit v3 (200 cycles). *Hirtella physophora* was sequenced on 1/12th of a lane during a preliminary test. The seven other libraries were multiplexed on one lane of the HiSeq 2000 flow cell, with each library on 1/24th of a lane.

Sequence assembly

Image analysis and base calling were automatically performed using the Illumina software CASAVA 1.8. Plastid genomes (ptDNA), mitochondrial genomes (mtDNA) and nuclear ribosomal clusters (nrDNA) were reconstructed using a script based on the UNIX command *awk* (Aho *et al.* 1979) and Python scripts (OBITOOLS, freely available at http://metabarcoding.org/obitools). Bioinformatics analyses were conducted on a computer cluster of the GENOTOUL bioinformatics platform (http://bioinfo.genotoul.fr/, Toulouse, France).

Because the closest complete ptDNA, mtDNA and nrDNA genomes available in public genomic databases were from species belonging to distant clades [either euphorbioids or salicoids according to Xi *et al.* (2012), we performed *de novo* assemblies rather than read mapping on a reference. First, we reconstructed short sequences (the probes) unambiguously originating either from ptDNA or from nrDNA. All reads were BLAST-searched against the phylogenetically closest genomes available on GenBank: *Jatropha curcas* L., Euphorbiaceae (GenBank accession no.: NC_012224) for ptDNA and *Linum usitatissimum* L., Linaceae (GenBank accession number: EU307117) for nrDNA. Reads with a match of at least 90% were assembled into small contigs using the Velvet assembler (Zerbino & Birney 2008). Second, we used these probes to initiate the genome assembly process by *in silico* genome walking: using the extractreads2 program (included in the OBITOOLS package), we selected sequence reads including a 'word' of at least 80 bases shared with the probes, the newly selected reads were subsequently used as probes and this process was repeated until no new read was identified (for an overview of classical genome walking in eukaryotes, see Leoni *et al.* 2011). This approach is similar in spirit to the strategy implemented for mitochondrial genomes in the MITObim (Hahn *et al.* 2013). Finally, the selected reads were joined using COPE when overlapping (Liu *et al.* 2012) and assembled using the Velvet assembler (Zerbino & Birney 2008). A script example is avaialble from Dryad (doi: 10.5061/dryad.78p90). The few resulting contigs were manually assembled using GENEIOUS v6.0.5 (Kearse *et al.* 2012). For each accession, we assembled a complete plastid genome and a nuclear ribosomal unit comprising the complete sequence of 26S, 18S and 5.8S genes and internal transcribed spacers (ITS1 and ITS2). For the nuclear ribosomal DNA, we did not assemble the internal gene spacer (IGS) because of the complexity of this region which is rich in duplications and inversions.

We also assembled large contigs of the mitochondrial genome (mtDNA). Short repeated sequence elements can represent up to 10% of the mtDNA in plants (Kubo *et al.* 2000), which impedes the assembly of the complete mitochondrial genome based on short fragments. Moreover, ptDNA and mtDNA may share very similar sequences, sometimes as long as several kilobases (Stern & Lonsdale 1982). We discovered that since mtDNA is underrepresented in leaf cells compared to ptDNA, shared sequences were an issue only for mtDNA assembly. We thus focused on the assembly of mitochondrial genes and their flanking regions. Reads that showed a match higher than 90% with ptDNA when BLAST-searched against ptDNA were filtered out. The remaining reads were assembled into long contigs as previously described, using as a reference the mitochondrial genome of *Ricinus communis* L., Euphorbiaceae (GenBank accession number: NC_015141). Only those contigs that unambiguously contained mtDNA genes were used for subsequent analyses.

Reads were finally mapped to the obtained sequences and annotation of the consensus sequences was performed in GENEIOUS v6.0.5. The newly generated sequences are available on GenBank (see Table 1 for accession numbers).

Phylogenetic reconstructions

For each cellular compartment, sequences were aligned along with outgroup sequences using MAUVE (Darling *et al.* 2004) and MAFFT (Katoh *et al.* 2002) as provided through the GENEIOUS v6.0.5 platform. Genome alignments were subsequently manually refined and the GBLOCKS program was used to avoid spurious informative characters caused by issues in automatic genome alignment due to indels and inversions (Castresana 2000; Talavera & Castresana 2007). Conserved and flanking positions were defined as shared by at least half the sequences with a maximum of eight contiguous nonconserved positions in at least 10-nucleotides long blocks. All gap positions were discarded. We also discarded the second inverted repeat region from the ptDNA alignment to avoid considering twice the same

Table 1 Sequence length, average sequencing depth, percentage of reads (% of reads), GC percentage (GC %) and GenBank accession number obtained for nuclear ribosomal region (nrDNA), plastid genome (ptDNA) and mitochondrial genome (mtDNA) for eight Chrysobalanaceae species

	nrDNA					ptDNA					mtDNA				
Species	Sequence length	Mean depth ± SD	% of reads	GC %	GenBank accession	Sequence length	Mean depth ± SD	% of reads	GC %	GenBank accession	Sequence length	Mean depth ± SD	% of reads	GC %	GenBank accession
Chrysobalanus icaco	6390	1084 ± 401	0.62	55.7	KJ414477	162 776	369 ± 64	5.36	36.2	KJ414480	237 331	81.1 ± 31.2	1.79	44.5	KJ414487-KJ415048
Couepia guianensis	6532	745 ± 300	0.32	55.5	KJ414475	162 123	225 ± 37	2.43	36.3	KJ414482	220 825	88.3 ± 54.9	1.30	44.6	
Hirtella physophora	6527	6476 ± 1100[a]	1.13	55.7	KJ414478	162 953	1545 ± 224[a]	6.69	36.2	KJ414485	249 879	389.8 ± 144.8[a]	2.64	44.4	
Hirtella racemosa	5953	1809 ± 753	0.67	55.9	KJ414472	162 889	562 ± 111	5.64	36.2	KJ414479	254 183	123.9 ± 133.3	1.99	44.4	
Licania alba	6558	1342 ± 491	0.54	55.8	KJ414473	162 465	476 ± 81	4.77	36.3	KJ414483	285 676	150.1 ± 47.7	2.68	44.4	
Licania heteromorpha	6626	1031 ± 402	0.47	56.7	KJ414471	162 840	543 ± 92	6.29	36.2	KJ414481	254 123	136.7 ± 132.4	2.50	44.5	
Licania sprucei	6679	926 ± 407	0.48	56.3	KJ414476	162 232	642 ± 140	7.94	36.3	KJ414484	274 241	178.7 ± 78.5	3.84	44.4	
Parinari campestris	6610	1284 ± 400	0.70	56.0	KJ414474	162 635	315 ± 71	4.24	36.2	KJ414486	244 301	105.3 ± 34.4	2.18	44.5	

[a] *Hirtella physophora* was sequenced on 1/12th of a lane during a preliminary test. The other libraries were multiplexed on one lane of the HiSeq 2000 flow cell, with each library on 1/24th of a lane.

information. Alignments are available from Dryad (doi: 10.5061/dryad.78p90). The ptDNA-based tree was rooted with complete plastid Malpighiales genomes available on GenBank: *Jatropha curcas*, *Manihot esculenta*, *Populus alba* and *Ricinus communis* (GenBank accession numbers: NC_012224, NC_010433, NC_008235 and NC_016736). The mtDNA-based tree was rooted using the complete mitochondrial genome of *Ricinus communis* (NC_015141), the only complete mitochondrial genome in the Malpighiales currently available on Genbank. The nrDNA-based tree was rooted with the only complete nuclear ribosomal region in the Malpighiales available on GenBank, that of *Linum usitatissimum* (EU307117). The number of variable sites and of parsimony informative characters (PICs) was calculated using MEGA v5.05 (Tamura *et al.* 2011).

The alignments were partitioned according to the best-fit nucleotide substitution model determined using the PARTITIONFINDER software (Lanfear *et al.* 2012). Input files corresponded to the most finely partitioned scheme possible i.e. individual partitions for each codon position of each protein-coding gene, for each rRNA, for each tRNA and for each noncoding region. In PARTITIONFINDER, we used the "greedy" algorithm (heuristic search) with branch lengths estimated as "unlinked" and tested the two models implemented in RAXML v7.2.8 (Stamatakis 2006): GTR+Γ and GTR+I+Γ. The optimal partitioning schemes and substitution models (see Tables S2 and S3, Supporting information) were used in subsequent phylogenetic reconstructions.

Phylogenetic reconstruction was performed by maximum likelihood inference (ML) using the RAXML software. Branch support for the topology was assessed by bootstrapping the original data set 1000 times. We also used the Bayesian approach as implemented in MRBAYES v3.2.1 (Ronquist & Huelsenbeck 2003). In these analyses, MCMCMC were run for 100 million generations, sampling parameters and trees every 1000 generations. Two independent runs each with four chains were executed simultaneously. The convergence of the chains was assessed by the Average Standard Deviation (ASD) in split frequencies across generations between the two runs. For all analyses, ASD was lower than 0.001 after excluding 25% of the samples as the burnin. Fifty per cent majority-rule consensus trees and their associated Bayesian posterior probabilities were generated with the sumt command in MRBAYES after discarding the first 25% of the samples as the burnin of the chains. The same analyses were conducted with alignments comprising only nonintronic transcribed sequences and with a concatenated data set comprising ptDNA, mtDNA and nrDNA using *Ricinus communis* as an outgroup (GenBank accession numbers: NC_016736 and NC_015141).

Results

Genome sequencing

The total DNA samples contained at least 1.5 μg of DNA, according to both Nanodrop and PicoGreen instruments with 260/230 and 260/280 purity ratios superior to 1.78 and 1.61, respectively. For each Illumina library, between 10.6 and 17.8 million reads were obtained per 24th of a lane (see Table S1, Supporting information).

We were able to reconstruct a complete ptDNA genome for all eight sampled species. The ptDNA genomes were around 162 kb in size (average GC content: 36.2%), and ptDNA sequencing depth was over 225× for 1/24th of a HiSeq 2000 lane (mean: 488×; Table 1). The fraction of chloroplast reads varied among species between 2.4% and 7.9% of the total reads (Table 1). The gene order was conserved among the eight species and was the same as the other available ptDNA genomes of Malpighiales [as an example, see Fig. S1 (Supporting information) for the physical map of the plastid genome of *Licania alba*].

We also assembled around 250 kb of the mtDNA genome for all species. These partial genomes included all genes but *rps1*, *rps7*, *rps13*, *rps19*, *rpl10* and *rpl16* (average GC content: 44.5%). Reads unambiguously assigned to mtDNA genomes ranged between 1.4% and 3.2% of the total reads, and the sequencing depth was superior to 81× for mtDNA (mean: 132×; Table 1).

Finally, a nrDNA contig including the complete sequence of 18S, ITS1, 5.8S, ITS2 and 26S was assembled for each species (mean size: 6485 bp; average GC content: 56%). The sequencing depth was greater than 745× for nrDNA (mean: 1432×; Table 1). Between 0.32% and 1.13% of the reads matched nrDNA sequences.

Phylogenetic results

After discarding the second inverted repeat region and ambiguously aligned sites, the ptDNA genome alignment comprised 109 474 nucleotide sites of which 1900 (1.7%) were variable and only 376 were parsimony informative within Chrysobalanaceae (Table 2). Half of the variable sites were found in noncoding regions. The ML tree and the Bayesian consensus tree had almost exactly the same topology, and only differed in the placement of *C. icaco*. All nodes but two received bootstrap values ≥85% and all but the same two had Bayesian posterior probabilities equal to 1 (Fig. 1a). The trees obtained with the data set including only plastid nonintronic transcribed sequences had the same topology but the nodes were less strongly supported: only three nodes received bootstrap values ≥85% and four nodes had posterior probabilities equal to 1.

Table 2 Alignment length, number of variable sites and number of parsimony informative characters (PICs) in noncoding, protein-coding, tRNA-coding and rRNA-coding DNA for the three regions considered in this study: nuclear ribosomal region (nrDNA), plastid genome (ptDNA) and mitochondrial sequences (mtDNA). The number of variable sites and of PICs was calculated excluding outgroup sequences

	Total alignment			Noncoding DNA			Protein-coding DNA			tRNA-coding DNA			rRNA-coding DNA		
	Length	Number of variable sites	Number of PICs	Length	Number of variable sites	Number of PICs	Length	Number of variable sites	Number of PICs	Length	Number of variable sites	Number of PICs	Length	Number of variable sites	Number of PICs
nrDNA	5609	220 (3.9%)	85	401	89	30	na	na	na	na	na	na	5208	131	55
ptDNA	109 474	1900 (1.7%)	376	38 771	905	186	63 900	981	187	2281	11	3	4522	3	0
mtDNA	155 509	1009 (0.6%)	224	120 419	891	205	28 225	102	18	1451	7	0	5414	9	1

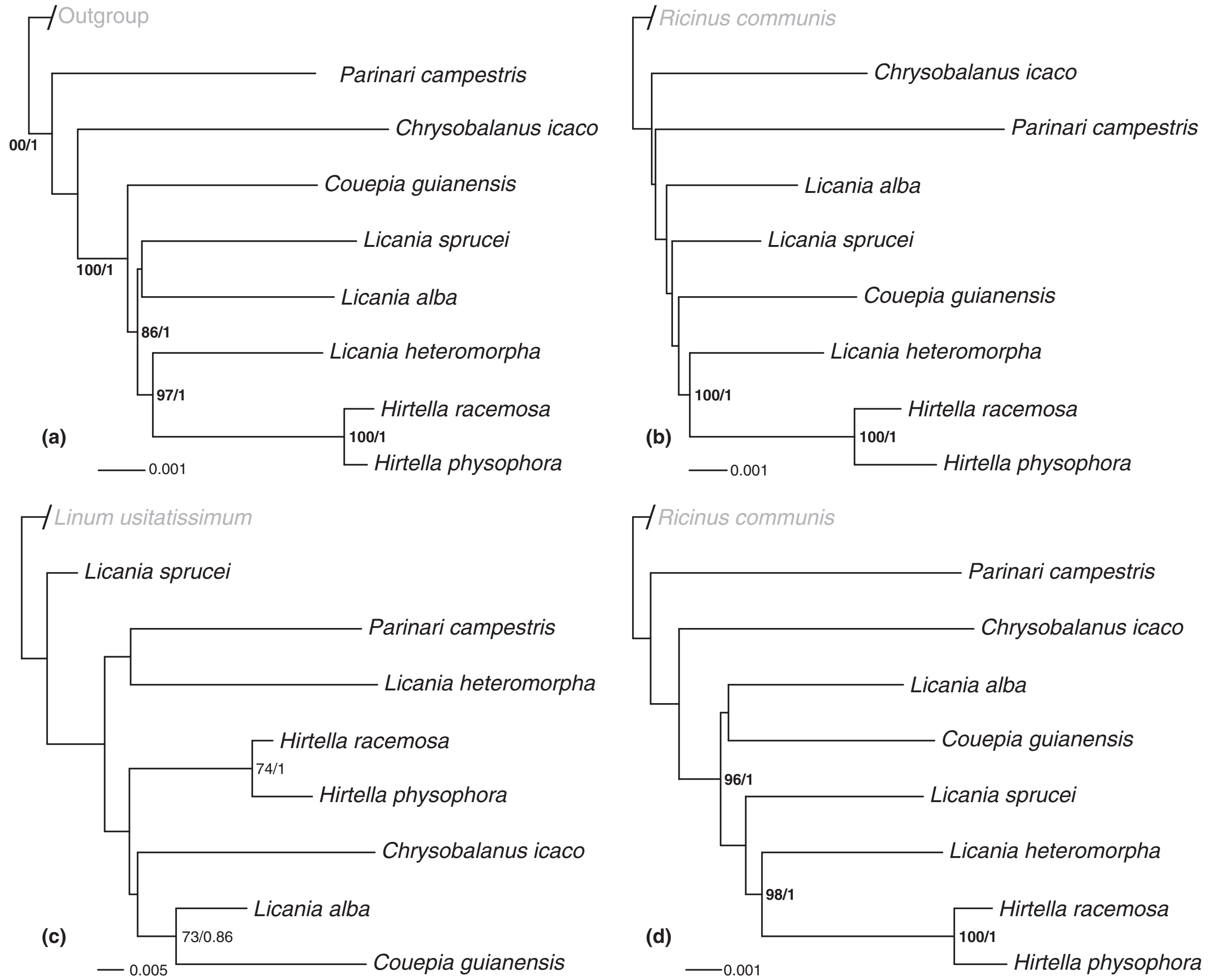

Fig. 1 Maximum likelihood phylogenetic trees for eight Chrysobalanaceae species obtained with complete plastid genome (ptDNA; panel a), partial mitochondrial genome (mtDNA; panel b), nuclear ribosomal region (nrDNA; panel c) and the concatenation of ptDNA, mtDNA and nrDNA (panel d). Branch lengths are drawn proportional to number of nucleotide substitutions per site, excepted for the branch leading to the outgroup that has been cut for scaling purposes. Bootstrap support values (using RAXML) and Bayesian posterior probabilities (from 50% majority-rule consensus trees obtained using MRBAYES) are reported when bootstrap values are superior to 70%. Bootstrap values higher than 85% are highlighted in bold. The part of the trees shaded in grey represents outgroups included in these analyses.

The mtDNA sequence alignment comprised 155 509 nucleotide sites, of which 1009 (0.6%) were variable and of these, 224 were parsimony informative (Table 2). Most of the variable sites (92%) were located in noncoding regions. The ML and Bayesian trees obtained with the data set comprising nontranscribed sequences were both poorly supported, but the few supported nodes were consistent with those supported in the ptDNA phylogeny. The trees obtained with only mitochondrial nonintronic transcribed sequences were even more poorly resolved (Fig 1b).

The alignment of the nuclear ribosomal region comprised 5609 nucleotide sites. Of these, 220 were variable and 85 parsimony informative (Table 2). With 3.9% of polymorphic sites, this nuclear region was found to be more variable than the other two genomic compartments. However, phylogenetic reconstructions, be they based on datasets including nontranscribed regions or not, produced poorly resolved trees and they only strongly supported the placement of the two closely related *Hirtella* species within the same clade (Fig. 1c).

Finally, the concatenated data set resulted in almost the same phylogenetic hypothesis as the one obtained with ptDNA alone, excepted for the placement of *L. alba* and with a generally lower node support (Fig. 1d).

Discussion

The value of shotgun sequencing for plant phylogenetics

We showed that shallow shotgun sequencing helps generate full organelle genomes for plant phylogenetic reconstruction and may be implemented for virtually all angiosperm species. This method is not based on PCR, so it alleviates much of the hurdles encountered in amplicon-based approaches. In addition, the method does not require a purification of ptDNA prior to sequencing, so laboratory work is greatly simplified. As a result, we obtained entire ptDNA genomes for eight species in a family for which little previous genomic study had been carried out (Yakandawala *et al*. 2010; Bardon *et al*. 2013). Unlike recent phylogenomic studies (Xi *et al*. 2012), we reconstructed full genomes, including noncoding regions. The sequencing depth of the ptDNA genomes was such that base-calling errors are very unlikely. This method represents about a ten-fold reduction in the assembled per-nucleotide sequencing cost compared with capillary sequencing. Such costs are likely to further drop in the future, especially if more species can be multiplexed in one run.

The Illumina library production was a critical technical step, and although 1–5 μg DNA is required as per the manufacturer's protocol, we found that much lower DNA quantities would suffice to generate good data (see also Straub *et al*. 2012; Besnard *et al*. 2013). In our study, it was relatively straightforward to obtain enough good quality DNA, and this limitation should be the focus of forthcoming technological improvements.

Towards a refined phylogeny of Chrysobalanaceae

The phylogenetic hypothesis obtained with complete ptDNA is consistent with the one published by Yakandawala *et al*. (2010) and Bardon *et al*. (2013). However, in previous analyses, almost none of the deep nodes were supported. As a result, while the monophyly of genera such as *Hirtella* and *Parinari* were suggested, the paraphyly of *Licania* could only be postulated. In the present analysis, five of the seven nodes are strongly supported in both ML and Bayesian analyses, but since Chrysobalanaceae comprise 18 genera, only five of which are included in the present study, we cannot draw definitive conclusions about the monophyly of genera in this family. However, our analysis is consistent with the existence of a Neotropical clade comprising the *Hirtella*, *Licania* and *Couepia* genera, and the basal position of Paleotropical members of the family. Through our analysis, the paraphyly of genus *Licania* is now strongly supported, and this has serious implications for the systematics of the family. A systematic revision of this large genus is now in need, and work in this direction is underway (C. A. Sothers, G. T. Prance & M. W. Chase, unpublished results). *Parinari* might be the basal-most lineage in our sampling and *Chrysobalanus*, a species-poor genus present in the coastal vegetation of both the Neotropics and Africa, might be more closely related to the Neotropical clade than to *Parinari*. At this point, our results do not allow us to discuss further the molecular systematics of this family, a question which we postpone to a forthcoming study including a more comprehensive taxon sampling.

On the relative merits and limits of mtDNA and nrDNA in plant phylogenomics

Unlike ptDNA, we were unable to fully reconstruct the complete mtDNA genome, due to its complex structure in plants, with many repeated regions. The few studies that included complete mtDNA genomes of plants (e.g. Rivarola *et al*. 2011) reported a genome size of about 500 kb. Here, we were able to assemble *c*. 250 kb of the mtDNA genome, including almost all the mitochondrial genes. The total length of mtDNA exons of our assembled regions is less than 30 kb (Table 2), a low figure compared with the ptDNA genome. Plant mtDNA genomes are highly dynamic (Galtier 2011), but they evolve mainly by intragenomic recombination through repeated sequences (Lonsdale *et al*. 1988), and appear to have a slower substitution rate than ptDNA (Wolfe *et al*. 1987; Palmer & Herbon 1988). Our results are in line with these observations with twice less variable sites in mtDNA vs. ptDNA (Table 2). Thus, the value of assembling large regions of mtDNA for phylogenetic reconstructions appears debatable, at least in Chrysobalanaceae.

The nuclear ribosomal region contained a proportion of variable characters twice higher than ptDNA. These results are in agreement with studies comparing evolutionary rates among plant cell genomes (Hamby & Zimmer 1992; Doyle 1993). In spite of this variability, the nrDNA phylogenetic tree was poorly resolved. One possible interpretation is the fast evolutionary rate of nrDNA (e.g., Grimm & Denk 2008), compared to ptDNA and mtDNA (Wolfe *et al*. 1987). This could lead to a high level of homoplasy, thus decreasing the phylogenetic information in this DNA region. The many nuclear copies of the ribosomal unit also evolve under concerted evolution, which is known to increase homoplasy (Hillis *et al*. 1991), probably because of GC-biased gene conversion (Galtier 2003; Escobar *et al*. 2011). Other phenomena such as the presence of paralogous copies and/or pseudogenes interfering with the sequencing of nrDNA (Álvarez & Wendel 2003) may further contribute to explain the poor resolution of nrDNA-based phylogenetic hypotheses in plants.

Our study confirms the importance of ptDNA genomic information to reconstruct deep nodes in the phylogeny of Chrysobalanaceae, and possibly of other plant lineages. On the other hand, mtDNA and nrDNA provided only marginally more information than ptDNA. An obvious next step would be to sequence nuclear exons through RNA sequencing (Zimmer & Wen 2012), but this is a far more technical step and one that critically depends on the question being addressed.

Implications for DNA barcoding

The present study also brings novel insight to plant DNA barcoding. Much emphasis has been placed in discovering DNA-based techniques to identify tissue samples based on sequencing small DNA fragments, an approach that has shown some promise in ecology (Vogler & Monaghan 2007; Gonzalez *et al.* 2009; Hibert *et al.* 2013; Quéméré *et al.* 2013), in the traceability of products of the food industry (Di Pinto *et al.* 2013; Pérez-Jiménez *et al.* 2013), in enforcing international regulations regarding biodiversity and biological resources (Armstrong & Ball 2005; Ardura *et al.* 2010), and in forensics (Dawnay *et al.* 2007). For plants, the consortium for the barcoding of Life has reached an agreement after much debate only in 2010, and fragments of two fragments of ptDNA exons have been chosen (*rbcL* and *matK*) to become the official barcoding markers (CBOL Plant Working Group 2009). Developing large DNA banks for these two markers is invaluable for many areas of research, but since we are now able to produce a full ptDNA genome, it is worthwhile asking if developing full ptDNA genome libraries would be advantageous over classic DNA barcode libraries (Kane *et al.* 2012). The technology described here produces ptDNA genomes at a tenth of the cost per base compared with the typical DNA barcode (*c.* 1.3 kb when concatenated). Hence, the cost of producing a full ptDNA genome currently is only about an order of magnitude larger than that of producing just the *rbcL* and *matK* barcodes. This cost gap is likely to reduce in the near future. Indeed, novel multiplexing approaches are expected, while Sanger sequencing costs are unlikely to go down now.

Since they rely on small DNA fragment sequencing, Illumina-based approaches are particularly suited to sequence heavily degraded genomes, thus facilitating a sound molecular characterization of rare or extinct species and increasing the potential use of species collections conserved in herbariums and museums (Guschanski *et al.* 2013). In addition, full ptDNA sequencing offers far more opportunities for detecting informative regions useful both in plant systematics and DNA barcoding, following the strategy of Shaw *et al.* (2007). Also, base-calling error is low for Illumina platforms, especially when the sequencing depth is 100× or higher. Finally, full ptDNA sequencing does not conflict with the currently established DNA barcoding strategy, since *rbcL* and *matK* would also be sequenced.

Conclusion

This study reports on a rapid and reliable method to generate plant organellar genomic data using a new-generation sequencing technology. We have demonstrated that this method is easy to implement even in tropical plant families, and that it yields robust data for organellar genomes. The evolutionary histories based on ptDNA and mtDNA sequences published here should not be misconstrued as definitive because they ignore several lineages in the family and the bulk of the nuclear DNA information, but they set the stage of future studies. For Chrysobalanaceae, and probably for other tropical plant families, this approach should help generate much more robust phylogenetic trees than in any previous study.

Acknowledgements

We thank P. Taberlet for his advice and support during the inception of this project. We gladly acknowledge the support from the *Plateforme Génomique* of Genopole Toulouse Midi-Pyrénées where most of the sequencing was performed, and from the Bioinformatics platform of Genopole Toulouse Midi-Pyrénées where most of the computation was performed. We also acknowledge the "*Génétique et Génomique des réponses aux stress biotique et abiotique du Tournesol*" group at the laboratory LIPM for their logistical help during the nucleic acid extraction. This work has benefited from "*Investissement d'Avenir*" grants managed by the *Agence Nationale de la Recherche* (CEBA, ref. ANR-10-LABX-25-01 and TULIP, ref. ANR-10-LABX-0041) and from an ANR grant (METABAR). This is contribution ISEM 2014-021 of the *Institut des Sciences de l'Evolution*.

References

Aho AV, Kernighan BW, Weinberger PJ (1979) Awk – a pattern scanning and processing language. *Software: Practice and Experience*, **9**, 267–279.

Álvarez I, Wendel JF (2003) Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution*, **29**, 417–434.

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

Ardura A, Linde AR, Moreira JC, Garcia-Vasquez E (2010) DNA barcoding of conservation and management of Amazonian commercial fish. *Biological Conservation*, **143**, 1438–1443.

Armstrong KF, Ball SL (2005) DNA barcodes for biosecurity: invasive species identification. *Philosophical transactions of the Royal Society of London. Series B, Biological Sciences*, **360**, 1813–1823.

Atherton RA, McComish BJ, Shepherd LD *et al.* (2010) Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform. *Plant Methods*, **6**, 22.

Bardon L, Chamagne J, Dexter KG *et al.* (2013) Origin and evolution of Chrysobalanaceae: insights into the evolution of plants in the Neotropics. *Botanical Journal of the Linnean Society*, **171**, 19–37.

Besnard G, Christin P-A, Malé P-JG *et al.* (2013) Phylogenomic and taxonomical surveys of Lecomtelleae (Poaceae), an isolated, early diverging panicoid tribe from Madagascar. *Annals of Botany*, **112**, 1057–1066.

Buerki S, Callmander MW, Devey DS *et al.* (2012) Straightening out the screw-pines: a first step in understanding phylogenetic relationships within Pandanaceae. *Taxon*, **61**, 1010–1020.

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, **17**, 540–552.

CBOL Plant Working Group (2009) A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the USA*, **106**, 12794–12797.

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

Cronn RC, Liston A, Parks M *et al.* (2008) Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Research*, **36**, e122.

Darling A, Mau B, Blattner F, Perna N (2004) Mauce: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research*, **14**, 1394–1403.

Davis CC, Anderson W (2010) A complete generic phylogeny of Malpighiaceae inferred from nucleotide sequence data and morphology. *American Journal of Botany*, **97**, 2031–2048.

Davis CC, Webb CO, Wurdack KJ, Jaramillo CA, Donoghue MJ (2005) Explosive radiation of Malpighiales supports a Mid-Cretaceous origin of modern tropical rain forests. *The American Naturalist*, **165**, E36–E65.

Dawnay N, Ogden R, McEwing R, Carvalho GR, Thorpe RS (2007) Validation of the barcoding gene COI for use in forensic genetic species identification. *Forensic Science International*, **173**, 1–6.

Delsuc F, Brinkmann H, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. *Nature Reviews. Genetics*, **6**, 361–375.

Di Pinto A, Di Pinto P, Terio V *et al.* (2013) DNA barcoding for detecting market substitution in salted cod fillets and battered cod chunks. *Food Chemistry*, **141**, 1757–1762.

Doyle JJ (1993) DNA, phylogeny, and the flowering of plant systematics. *BioSciences*, **43**, 380–389.

Egan AN, Schlueter J, Spooner DM (2012) Applications of next-generation sequencing in plant biology. *American Journal of Botany*, **99**, 175–185.

Escobar JS, Glémin S, Galtier N (2011) GC-biased gene conversion impacts ribosomal DNA evolution in vertebrates, angiosperms, and other eukaryotes. *Molecular Biology and Evolution*, **28**, 2561–2575.

Feitosa EA, Xavier HS, Randau KP (2012) Chrysobalanaceae: traditional uses, phytochemistry and pharmacology. *Revista Brasileira de Faramacognosia*, **22**, 1181–1186.

Galtier N (2003) Gene conversion drives GC content evolution in mammalian histones. *Trends in Genetics*, **19**, 65–68.

Galtier N (2011) The intriguing evolutionary dynamics of plant mitochondrial DNA. *BMC Biology*, **9**, 61.

Givnish TJ, Ames M, McNeal JR *et al.* (2010) Assembling the tree of the monocotyledons: plastome sequence phylogeny and evolution of Poales. *Annals of the Missouri Botanical Garden*, **97**, 584–616.

Gonzalez MA, Baraloto C, Engel J *et al.* (2009) Identification of Amazonian trees with DNA barcodes. *PLoS ONE*, **4**, e7483.

Goremykin VV, Hirsh-Ernst KI, Wölfl S, Hellwig FH (2004) The chloroplast genome of *Nymphaea alba*: whole-genome analyses and the problem of identifying the most basal angiosperm. *Molecular Biology and Evolution*, **21**, 1445–1454.

Grimm GW, Denk T (2008) ITS evolution in *Platanus* (Platanaceae): homoeologues, pseudogenes and ancient hybridization. *Annals of Botany*, **101**, 403–419.

Guschanski K, Krause J, Sawyer S *et al.* (2013) Next-Generation museomics disentangles one of the largest primate radiations. *Systematic Biology*, **62**, 539–554.

Hahn C, Bachmann L, Chevreux B (2013) Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Research*, **41**, e129.

Hamby RK, Zimmer EA (1992) Ribosomal RNA as a phylogenetic tool in plant systematics. In: *Molecular Systematics of Plants* (eds Soltis PS, Soltis DE, Doyle JJ), pp. 50–91. Chapman & Hall, New York, New York.

Hibert F, Taberlet P, Chave J *et al.* (2013) Unveiling the diet of elusive rainforest herbivores in Next Generation Sequencing era? The tapir as a case study. *PLoS ONE*, **8**, e60799.

Hillis DM, Moritz C, Porter CA, Baker RJ (1991) Evidence for biased gene conversion in concerted evolution of ribosomal DNA. *Science*, **251**, 308–310.

Hopkins MJG (2007) Modelling the known and unknown plant biodiversity of the Amazon Basin. *Journal of Biogeography*, **34**, 1400–1411.

Kane N, Sveinsson S, Dempewolf H *et al.* (2012) Ultra-barcoding in cacao (*Theobroma* spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. *American Journal of Botany*, **99**, 320–329.

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transformation. *Nucleic Acids Research*, **14**, 3059–3066.

Kearse M, Moir R, Wilson A *et al.* (2012) GENEIOUS Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, **28**, 1647–1649.

Kubo T, Nishizawa S, Sugawara A *et al.* (2000) The complete nucleotide sequence of the mitochondrial genome of sugar beet (*Beta vulgaris* L.) reveals a novel gene for tRNA^{Cys}(GCA). *Nucleic Acids Research*, **28**, 2571–2576.

Lanfear R, Calcott B, Ho SYW, Guindon S (2012) PARTITIONFINDER: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution*, **29**, 1695–1701.

Leebens-Mack J, Rauberson LA, Cui L *et al.* (2005) Identifying the basal angiosperm node in chloroplast genome phylogenies: sampling one's way out of the Felsenstein zone. *Molecular Biology and Evolution*, **22**, 1948–1963.

Leoni C, Volpicella M, De Leo F, Gallerani R, Ceci LR (2011) Genome walking in eukaryotes. *FEBS Journal*, **278**, 3953–3977.

Liu B, Yuan J, Yiu S-M *et al.* (2012) COPE: an accurate k-mer-based pair-end reads connection tool to facilitate genome assembly. *Bioinformatics*, **28**, 2870–2874.

Lohse M, Dreschel O, Bock R (2007) OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochonrial genomes. *Current Genetics*, **52**, 267–274.

Lonsdale DM, Brears T, Hodge TP, Melville SE, Rottman WH (1988) The plant mitochondrial genome: homologous recombination as a mechanism for generating heterogeneity. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, **319**, 149–163.

Magallon S, Crane PR, Herendeen PS (1999) Phylogenetic pattern, diversity, and diversification of Eudicots. *Annals of the Missouri Botanical Garden*, **86**, 297–372.

McPherson H, van der Merwe M, Delaney SK *et al.* (2013) Capturing chloroplast variation for molecular ecology studies: a simple next generation sequencing approach applied to a rainforest tree. *BMC Ecology*, **13**, 8.

Moore MJ, Dhingra A, Soltis PS *et al.* (2006) Rapid and accurate pyrosequencing of angiosperm plastid genomes. *BMC Plant Biology*, **6**, 17.

Moore MJ, Bell CD, Soltis PS, Soltis DE (2007) Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proceedings of the National Academy of Sciences of the USA*, **104**, 19363–19368.

Moore MJ, Soltis PS, Bell CD, Burleigh JG, Soltis DE (2010) Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proceedings of the National Academy of Sciences of the USA*, **107**, 4623–4628.

Murakami P, McCaman MT (1999) Quantitation of adenovirus DNA and virus particles with the PicoGreen fluorescent dye. *Analytical Biochemistry*, **274**, 283–288.

Palmer JD, Herbon LA (1988) Plant mitochondrial DNA evolved rapidly in structure, but slowly in sequence. *Journal of Molecular Evolution*, **28**, 87–97.

Pérez-Jiménez M, Besnard G, Dorado G, Hernandez P (2013) Varietal tracing of virgin olive oils based on plastid DNA variation profiling. *PLoS ONE*, **8**, e70507.

Prance GT (1972) Monograph of the Chrysobalanaceae. *Flora Neotropica*, **9**, 1–410.

Prance GT, Sothers CA (2003) *Species Plantarum: Flora of the World. Part 9. Chrysobalanaceae 1. Chrysobalanus to Parinari*. Australian Biological Resources Study, Canberra, Australia.

Prance GT, White F (1988) The genera of Chrysobalanaceae: a study in practical and theoretical taxonomy and its relevance to evolutionary biology. *Philosophical transactions of the Royal Society of London. Series B, Biological Sciences*, **320**, 1–184.

Quéméré E, Hibert F, Miquel C *et al.* (2013) A DNA metabarcoding study of a primate dietary diversity and plasticity across its entire fragmented range. *PLoS ONE*, **8**, e58971.

Rivarola M, Foster JT, Chan AP *et al.* (2011) Castor bean organelle genome sequencing and worldwide genetic diversity analysis. *PLoS ONE*, **6**, e21743.

Ronquist F, Huelsenbeck JP (2003) MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.

Shaw J, Lickey EB, Schilling EE, Small RL (2007) Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. *American Journal of Botany*, **94**, 275–288.

Soltis DE, Gitzendanner MA, Soltis PS (2007) A 567-taxon data set for angiosperms: the challenges posed by Bayesian analyses of large data sets. *International Journal of Plant Sciences*, **168**, 137–157.

Soltis DE, Smith SA, Cellinese N *et al.* (2011) Angiosperm phylogeny: 17 genes, 640 taxa. *American Journal of Botany*, **98**, 704–730.

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

Stern DB, Lonsdale DM (1982) Mitochondrial and chloroplast genomes of maize have a 12-kilobase DNA sequence in common. *Nature*, **299**, 702.

Straub SCK, Parks M, Weitemier K *et al.* (2012) Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. *American Journal of Botany*, **99**, 349–364.

Swenson U, Richardson JE, Bartish IV (2008) Multi-gene phylogeny of the pantropical subfamily Chrysophylloideae (Sapotaceae): evidence of generic polyphyly and extensive morphological homoplasy. *Cladistics*, **24**, 1006–1031.

Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology*, **56**, 564–577.

Tamura K, Peterson D, Peterson N *et al.* (2011) MEGA5: Molecular Evolutionary Genetics Analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, **28**, 2731–2739.

Varma A, Padh H, Shrivastava N (2007) Plant genomic DNA isolation: an art or a science. *Biotechnology Journal*, **2**, 386–392.

Vogler AP, Monaghan MT (2007) Recent advances in DNA taxonomy. *Journal of Zoological Systematics and Evolutionary Research*, **45**, 1–10.

Whitfield JB, Lockhart PJ (2007) Deciphering ancient rapid radiations. *Trends in Ecology & Evolution*, **22**, 258–265.

Wolfe KH, Li WH, Sharp PM (1987) Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Sciences of the USA*, **84**, 9054–9058.

Xi Z, Ruhfel BR, Schaefer H *et al.* (2012) Phylogenomics and *a posteriori* data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. *Proceedings of the National Academy of Sciences of the USA*, **109**, 17519–17524.

Yakandawala D, Morton CM, Prance GT (2010) Phylogenetic relationships of the Chrysobalanaceae inferred from chloroplast, nuclear, and morphological data. *Annals of the Missouri Botanical Garden*, **97**, 259–281.

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

Zimmer EA, Wen J (2012) Using nuclear gene data for plant phylogenetics: progress and prospects. *Molecular Phylogenetics and Evolution*, **65**, 774–785.

P.J.G.M. and J.C. designed the study and wrote the manuscript. J.E., P.J.G.M. and C.S.S. conducted the field sampling. E.L., P.J.G.M. and A.T. produced the data. L.B., G.B., E.C., F.D. and P.J.G.M. analysed the data. All authors commented on the manuscript.

Data Accessibility

Raw sequences are available from the SRA database under the accession no. SRP039035. DNA sequences have been deposited in the GenBank database (see Table 1 for accession numbers). An example of UNIX script, the alignment files and all tree files are available from Dryad (doi: 10.5061/dryad.78p90).

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1 Physical map of the plastid genome (ptDNA) of *Licania alba*, drawn with the OGDRAW software (Lohse *et al.* 2007).

Fig. S2 Maximum likelihood phylogenetic trees for eight Chrysobalanaceae species obtained with nonintronic transcribed sequences of plastid genome (panel a), mitochondrial genome (panel b) and nuclear ribosomal region (panel c). Branch lengths are drawn proportional to number of nucleotide substitutions per site, excepted for the branch leading to the outgroup that has been cut for scaling purposes. Bootstrap support values (using RAXML) and Bayesian posterior probabilities (from 50% majority-rule consensus trees obtained using MRBAYES) are reported when bootstrap values are superior to 70%. Bootstrap values higher than 85% are highlighted in bold. The part of the trees shaded in grey represents outgroups included in these analyses.

Table S1 Quantity and quality of total DNA used for eight Chrysobalanaceae species and number of reads obtained after sequencing. *H. physophora* was sequenced on a 12th of a HiSeq 2000 lane whereas the other species were sequenced on a 24th of a HiSeq 2000 lane.

Table S2 Optimal partitioning scheme and best-fit models identified by the PARTITIONFINDER software with model choice restricted to models available in RAXML for alignments comprising nontranscribed regions (hereafter labelled “nc”). Alpha and the proportion of invariant sites have been estimated using RAXML.

Table S3 Optimal partitioning scheme and best-fit models identified by the PARTITIONFINDER software with model choice restricted to models available in RAXML for alignments comprising only nonintronic transcribed regions. Alpha and the proportion of invariant sites have been estimated using RAXML.