Title
Museomics illuminate the history of an extinct, paleoendemic plant lineage (Hesperelaea, Oleaceae) known from an 1875 collection from Guadalupe Island, Mexico

Permalink
https://escholarship.org/uc/item/1615g4rh

Journal
Biological Journal of the Linnean Society, 117(1)

ISSN
0024-4066

Authors
Zedane, L
Hong-Wa, C
Murienne, J
et al.

Publication Date
2016

DOI
10.1111/bij.12509

Peer reviewed
Museomics illuminate the history of an extinct, paleoendemic plant lineage (*Hesperelaea*, Oleaceae) known from an 1875 collection from Guadalupe Island, Mexico

LOUBAB ZEDANE†, CYNTHIA HONG-WA†, JÉRÔME MURIENNE, CÉLINE JEZIORSKI, BRUCE G. BALDWIN and GUILLAUME BESNARD

1CNRS-UPS-ENFA, UMR5174, EDB (Laboratoire Evolution & Diversité Biologique), Université Paul Sabatier, 118 route de Narbonne, 31062, Toulouse, France
2Missouri Botanical Garden, PO Box 299, St. Louis, MO, 63166-0299, USA
3INRA, UAR1209, département de Génétique Animale, INRA Auzeville 31326, Castanet-Tolosan, France
4GeT-PlaGe, Genotoul, INRA Auzeville 31326, Castanet-Tolosan, France
5Jepson Herbarium and Department of Integrative Biology, University of California, Berkeley, CA, 94720-2465, USA

Received 21 October 2014; revised 29 January 2015; accepted for publication 30 January 2015

Museum collections are essential for understanding biodiversity and next-generation sequencing methods (NGS) offer new opportunities to generate genomic data on specimens of extinct species for phylogenetic and other studies. *Hesperelaea* is a monotypic Oleaceae genus that was collected only once, 140 years ago on Guadalupe Island, Mexico. This lineage is almost certainly extinct, and has been considered an insular paleoendemic of unknown relationship within subtribe *Oleinaceae*. Here, a genome skimming approach was attempted on the *H. palmeri* specimen to generate genomic data in order to interpret the biogeographic history of *Hesperelaea* in a phylogenetic framework. Despite highly degraded DNA, we obtained the complete plastome, the nuclear ribosomal DNA cluster (nrDNA), and partial sequences of low-copy genes. Six plastid regions and nrDNA internal transcribed spacers were used for phylogenetic estimations of subtribe *Oleinaceae*, including data from previous studies. Bayesian and maximum likelihood phylogenies strongly place *Hesperelaea* within an American lineage that includes *Forestiera* and *Priogymnanthus*. Molecular dating suggests an Early Miocene divergence between *Hesperelaea* and its closest relatives. Our study thus confirms that *Hesperelaea* was a paleoendemic lineage that likely predates Guadalupe Island, and provides a notable example of the high potential of NGS for analyzing historical herbarium specimens and revolutionizing systematics. © 2015 The Linnean Society of London, Biological Journal of the Linnean Society, 2016, 117, 44–57.


INTRODUCTION

Museum collections have played an important role for biodiversity inventories, taxonomy, and comparative biology in general over the last 3 centuries (Suarez & Tsutsui, 2004; Bebber et al., 2010). Today, with the development of technologies in biochemistry and molecular biology, these collections can also be seen as a putative source of biomolecules (such as DNA) for systematic studies, especially for rare species or critical taxa occurring in remote areas. With the advent of the polymerase chain reaction (PCR), historical museum specimens have been included in phylogenetic or population genetic work and this has allowed the investigation of important topics such as the placement of extinct species in phylogenies (e.g.
Wallander & Albert, 2000; Sebastian, Schaefer & Renner, 2010), the reconstruction of historic intercontinental movements of crops and their associated pests (Ames & Spooner, 2008; Schaefer & Renner, 2010; Yoshida et al., 2013), and the origin of herbicide resistance alleles in weeds (Délye, Deulvot & Chauvel, 2013). The main problem with using museum specimens for these purposes is the low quality of DNAs that can be extracted, especially from poorly preserved samples that were collected several decades or centuries ago (Wandeler, Hoek & Keller, 2007). The best preserved specimens are usually chosen for molecular work, but the amplification by PCR of short DNA segments (e.g. often less than 400 bp) requires fastidiousness and considerable expense, with somewhat unpredictable levels of success. The problem could be tackled by using next-generation sequencing methods (NGS), which have already revolutionized phylogenetic investigations by allowing the simultaneous generation of a large quantity of sequences for the different genomes present in an organism (Glenn, 2011; Harrison & Kider, 2011). In particular, NGS methods have been used to assemble repeated DNA regions such as organelar genomes and nuclear ribosomal units (Cronn et al., 2008; Straub et al., 2012; Kocher et al., 2014; Malé et al., 2014). As these procedures do not rely on PCR amplification, they can potentially be applied to poorly preserved DNA (Bi et al., 2013). NGS is thus expected to facilitate the molecular analysis of museum specimens, and especially extinct species, thereby allowing detailed sampling of lineages that have been difficult or impossible to include in phylogenetic studies.

Here, we applied NGS to a sample of an extinct monotypic genus (Hesperelaea A. Gray, Oleaceae) collected only once, in 1875. Obtaining genetic information from the specimen of this species was considered a challenge due to the poor quality of the DNA that was extracted (Wallander & Albert, 2000). This species represents an iconic, insular tree of the Californian Floristic Province (Raven & Axelrod, 1978) whose origin has remained unknown based on its unusual morphology and lack of a known fossil record (Axelrod, 1967b; see below); thus, its past history needs to be analyzed in a phylogenetic framework. In this study, we apply a genome skimming strategy (Straub et al., 2012) in order to reconstruct a complete plastid genome, a nearly complete nuclear ribosomal DNA unit, and sequences of nuclear low-copy genes of Hesperelaea palmeri A. Gray. These DNA markers were combined with available datasets to estimate phylogeny of subtribe Oleinae. Based on these results, we discuss the biogeographic history of Hesperelaea in the Californian Islands. Lastly, the implications of our study for improving the sampling in phylogenetic investigations, for both species and DNA characters, are briefly discussed.

THE STUDY SPECIES, ITS PAST HABITAT AND CURRENT SYSTEMATIC POSITION

Hesperelaea palmeri was a tree species endemic to Guadalupe Island, Baja California (Mexico), about 350 km southwest of Ensenada. Guadalupe Island has a total area of 244 km² and is composed of ocecinic shield volcanoes (olivine basalt and trachyte rocks) that formed on an extinct oceanic ridge (Batiza, 1977). Based on the oldest lava flow, the island is estimated to be 7 ± 2 million years old (Hubbs, 1967; Engel & Engel, 1970). It is the highest and most remote of the California Islands (Thorne, 1969) and has never been in contact with the mainland based on depth (~3.6 km) of intervening seafloor (Moran, 1996). Several authors (Thorne, 1969; Wallace, 1985; Moran, 1996) have discussed the disharmonic flora of Guadalupe Island and compared it to the other California Islands that are known to have a high percentage of endemics, especially the Channel Islands (Thorne, 1969). Guadalupe Island is characterized by a relatively low native diversity of vascular plants (c. 156 species) but a relatively high level (21.8% of species) of single-island endemism. A case of within-island diversification on Guadalupe Island was documented in Deinandra Greene, Asteraceae (Carquist, 1965; Baldwin, 2007).

Since the early 19th century, the original habitat of Guadalupe Island has been heavily impacted by human-related activities, in particular as a result of intense grazing by feral goats, the invasion of exotic weeds, and intensive use of local woody resources that in turn resulted in soil erosion (Moran, 1996). Recent restoration projects allowed an extermination of goats and the vegetation of the island seems to be recovering (Garcillan, Ezcurra & Vega, 2008). Before that, the island was described as a ‘biological cemetery’ because of many examples of early extinction or extirpation at the end of the 19th century or the early 20th century (e.g. birds such as endemic taxa of Bewick’s wren, spotted towhee, caracara, flicker, and storm petrel; at least 26 native plant species including endemics such as Castilleja guadalupensis Brandegee and Pogogyne tenuiflora A. Gray; Kaeding, 1905; de la Luz, Rebman & Oberbauer, 2003). Yet, genetic diversity higher than in mainland relatives was recently reported for a pollen-dispersed plastid genome in the Guadalupe cypress [Hesperocyparis guadalupensis (S. Watson) Bartel] suggesting that populations of some endemic trees did not suffer severe genetic consequences from population contractions resulting from human activities (Escobar et al., 2011).
One individual of *Hesperelaea palmeri* was collected in 1875 (Fig. S1) by Edward Palmer on the eastern side of the island in a canyon covered by trees reaching a height of up to 8 m (Gray, 1876; Watson, 1876). The original vegetation of the surrounding area was probably a mesic shrub/herbland (de la Luz et al., 2003). The taxon was sampled once, and was considered to be very rare by E. Palmer since only three living individuals were recorded among many dead ones (in Watson, 1876). Given the very small size of this population, *H. palmeri* could have been termed an 'extremely narrow endemic' and has been considered undoubtedly extinct since the end of the 19th century (Brandegee, 1900; Eastwood, 1929; Moran, 1996). The 11 duplicates of the Palmer specimen (*E. Palmer SL*; Fig. S1) conserved in eight herbaria (BM, CM, K, LE, MO, NY, P, YU) are thus the only known remains of this taxon.

*Hesperelaea* is a monotypic and putatively paleoendemic genus of the olive family (Oleaceae, order Lamiales). This family encompasses more than 600 species in 25 genera (Green, 2004). Most of them are trees or shrubs, and a few are woody climbers. The family Oleaceae has a worldwide distribution, occurring in tropical, subtropical, and temperate climates (Stevens, 2001). Several species are of major economic interest, such as the Mediterranean olive (*Olea europaea* L.) and the ash tree (*Fraxinus excelsior* L.), but numerous other Oleaceae taxa (e.g. in *Chionanthus* L., *Forsythia Vahl*, *Fraxinus* L., *Jasminum* L., *Ligustrum* L., *Osmanthus* Lour., and *Syringa* L.) are widely used as ornamentals or fragrant species. Taxa of Oleaceae are currently classified in five tribes, namely *Fontanesieae*, *Forsythieae*, *Jasminieae*, *Myxopyreae*, and *Oleae* (Wallander & Albert, 2000). *Oleae* has a tetraploid origin (Taylor, 1945) and four subtribes have been recognized in this lineage: *Ligustrinae*, *Fraxininae*, *Schreberinae*, and *Oleinae* (Wallander & Albert, 2000). *Hesperelaea* belongs to tribe *Oleae*, subtribe *Oleinae* (Wallander & Albert, 2000). The phylogenetic relationships of Oleaceae have been investigated using plastid and nuclear ribosomal regions with various degrees of resolution, and recent studies have mainly focused on specific lineages such as *Fraxinus* (Jeandroz, Roy & Bousquet, 1997; Wallander, 2008; Area et al., 2012; Hinsinger et al., 2013, 2014), *Syringa–Ligustrum* (Kim & Jansen, 1998; Li, Alexander & Zhang, 2002), *Olea* (Besnard et al., 2009), *Osmanthus* (Yuan et al., 2010; Guo et al., 2011), and *Noronhia–Chionanthus* (Hong-Wa & Besnard, 2013). These studies revealed that several Oleaceae genera (e.g. *Chionanthus, Ligustrum, Olea*, and *Osmanthus*) are paraphyletic or polyphyletic, and there is a need for taxonomic revisions, in particular of subtribe *Oleinae* (e.g. Besnard et al., 2009; Yuan et al., 2010; Guo et al., 2011; Hong-Wa & Besnard, 2013). In addition, many Oleaceae species are rare or micro-endemic, and some are known only from one or two old herbarium records (for *Noronhia* Stadtm. ex Thouars, see Hong-Wa & Besnard, 2014) making their molecular characterization difficult. Wallander & Albert (2000) placed *Hesperelaea* within subtribe *Oleinae* in the Oleaceae phylogeny but its position within *Oleinae* has remained elusive due to a lack of DNA information (only 423 bp from *rps16*, GenBank accession: AF225245). The use of more information is thus necessary to re-evaluate the systematic treatment of *Hesperelaea* as a monotypic genus, and to infer its origin in light of phylogeny, fossils, and hypotheses on the biogeography of the Californian biota (Axelrod, 1958, 1967c).

### MATERIAL AND METHODS

**DNA EXTRACTION AND LIBRARY PREPARATION**

Genomic DNA was extracted using the DNeasy Plant Mini kit (Qiagen, Valencia, CA, USA) from a leaf fragment (5 mg) taken from the fragment packet of the specimen (*E. Palmer SL*, MO-992430) deposited at the Missouri Botanical Garden herbarium. Quantification of the DNA concentration based on NanoDrop analysis (Thermo Fisher Scientific, DE, USA) indicated that the solution was unclean and not properly assessed (Fig. S2). The quantification using Quant-iTTm PicoGreen® (Molecular Probes Inc., OR, USA; Murakami & McCaman, 1999) also revealed a very low concentration of double-stranded DNA (2.2 ng µL⁻¹). First, we tried to generate short plastid DNA segments (e.g. less than 300 bp) using a PCR approach as reported by Besnard et al. (2009), but this failed on the 15 tested fragments, probably because the DNA was too fragmented. Thirty-three ng of double-stranded DNA were then used for shotgun sequencing with the Illumina technology (Illumina Inc., San Diego, CA, USA). The sequencing was performed at the GENOPOLE Toulouse Midi-Pyrénées. The library was constructed using the Illumina TruSeq Nano DNA LT Sample Prep kit following the instructions of the supplier, except that the DNA library was generated without prior DNA sonication because the DNA was presumably highly degraded. Purified fragments were A-tailed and ligated to sequencing indexed adapters. No size selection was performed on the DNA, and fragments with an insert size of c. 50–250 bp were enriched with eight cycles of PCR before library quantification and validation (Fig. S2). The library was multiplexed with 23 other libraries (generated in other projects). The pool of libraries was then hybridized to the HiSeq 2000 flow cell using the Illumina TruSeq PE Cluster Kit v.3.
Bridge amplification was performed to generate clusters, and paired-end reads of 100 nucleotides were collected on the HiSeq 2000 sequencer using the Illumina TruSeq SBS Kit v.3 (200 cycles).

**READ ASSEMBLY, SEQUENCING DEPTH AND GENE ANNOTATION**

All paired-end reads were used for the reconstruction of a complete plastid genome, a nuclear ribosomal cluster [including external and internal transcribed spacers (ETS, ITS1 and ITS2) and 18S, 5.8S and 26S ribosomal RNA genes], and five low-copy gene regions.

Reads corresponding to the plastid genome (ptDNA) and the nuclear ribosomal cluster (nrDNA) were filtered using the approach described by Besnard et al. (2013). Plastid and nuclear ribosomal sequences of the olive tree (EMBL accessions: FN996972 and L499289) were used as probes for reconstructing the plastid genome and the nuclear ribosomal cluster. First, using the program extractread2 (included in the OBITools package, http://metabarcoding.org/obitools), we selected sequence reads including a ‘word’ of at least 90 bases common with the probe. The newly selected reads were used as a probe and the previous process was repeated until no new reads were identified. Second, the set of selected reads was assembled using VELVET (Zerbino & Birney, 2008) in contigs that were then sorted with GENEIOUS v6.1.7 (Kearse et al., 2012) by comparing them with the plastid genome or the nuclear ribosomal cluster of the olive tree. Finally, all reads were mapped onto the obtained sequences using GENEIOUS to check the assembly quality and to assess the depth of sequencing of all investigated regions. Duplicated paired-end reads were filtered using a script based on BWA (http://bio-bwa.sourceforge.net) until no new reads were identified. Second, the set of selected reads was assembled using VELVET (Zerbino & Birney, 2008) in contigs that were then sorted with GENEIOUS v6.1.7 (Kearse et al., 2012) by comparing them with the plastid genome or the nuclear ribosomal cluster of the olive tree. Finally, all reads were mapped onto the obtained sequences using GENEIOUS to check the assembly quality and to assess the depth of sequencing of all investigated regions. Duplicated paired-end reads were filtered using a script based on BWA (http://bio-bwa.sourceforge.net) and SAMtools (Li et al., 2009). All of these computations were done on a computer cluster of the Geno- tool bioinformatics platform (Toulouse, France). Annotation of the consensus sequence of the plastid genome and nuclear ribosomal cluster was performed with GENEIOUS by using the annotated sequences of olive available in GenBank: *Olea europaea* subsp. *cuspi data* (Wall. ex G. Don) Cif. (FN650747) for the ptDNA, and *O. e.* subsp. *europaea* (AJ588193, AJ865373) for the nrDNA.

Partial nuclear sequences were assembled following the approach described by Besnard et al. (2013, 2015) using reference sequences of five low-copy genes (i.e. *Phantastica*; Hinsinger et al., 2013; FAD6, *CUL4, OEW, OCO*; Besnard & El Bakkali, 2014) available for a few *Oleaceae* species in GenBank. The depth of sequencing on these nuclear genes was also assessed with GENEIOUS. These data were not used to estimate phylogeny because the taxon sampling (essentially *Fraxinus* spp. or *Olea* spp.) is not comprehensive enough for the purpose of our study.

The newly generated sequences of *Hesperelaeea* are available on GenBank (LN515488, LN515489, LN681359, LN809935 to LN809938).

**PHYLOGENETIC ANALYSES**

The phylogenetic position of *H. palmeri* within *Oleinae* was then investigated using both nrDNA and ptDNA sequences. We re-used sequences generated by Hong-Wa & Besnard (2013) but we also considered two additional ptDNA segments (*rps16, atpB–rbcl*) that were analyzed by Wallander & Albert (2000). Compared to the phylogenies presented in Hong-Wa & Besnard (2013), a few additional species (e.g. *Chionanthus filiformis* (Vell.) P.S. Green, *Priotymnum hasselianus* (Chodat) P.S. Green, *Forestiera spp.*) were included in our dataset (sequences retrieved from GenBank), but fewer (six) accessions of *Noronhia* were selected to represent the worldwide geographic distribution of the genus. Finally, a subset of 45 taxa representative of the main known lineages of subtribe *Oleinae* was considered (Table S1). At least one accession of each *Oleinae* genus was included in the ptDNA and nrDNA phylogenies except for *Priotymnum* P.S. Green, which was missing in the nrDNA analysis (we unsuccessfully tried to amplify ITS1 from a recent herbarium sample (Zardini & Vera 41142, MNHN-P-P03868535)). For *H. palmeri*, we included six intergenic regions (*rps16, atpB–rbcl*, *trnL–trnF, trnK–matK, trnS–trnG, and trnT–trnL*) extracted from the complete plastid genome, and the ITS sequence (IT1-5.8S-ITS2) from the nuclear ribosomal cluster.

Nucleotide sequences were aligned using Muscle (Edgar, 2004) with default parameters. Identification and removal of the poorly aligned and gapped positions was performed using the program Gblock v0.91b (Castresana, 2000) with the options −t = d and −b5 = h, and with the additional options −b4 = 5 for *trnK–matK* and *trnS–trnG*, and −b2 = 30 for *trnL–trnF*. Inverted regions were identified by visual inspection of the alignment and recoded as missing data (i.e. positions 5923–5948 in *Noronhia foveolata* (E. May) Hong-Wa & Besnard, positions 5053–5055 in *Noronhia emarginata* Thouars, and positions 4558–4597 in *Chionanthus filiformis*; the complete alignment is available at TreeBase (http://purl.org/phylo/treebase/phylows/study/TB2:16997)). All markers were concatenated using FASConCAT (Kück & Meusemann, 2010). PartitionFinder v.1.0.1 (Lanfear et al., 2012) was used to infer the best-fit partitioning scheme and substitution model using the Bayesian Information Criterion (Schwarz, 1978).
and the greedy algorithm. We defined nine starting data blocks (ITS1, 5.8S, ITS2, atpB–rbcL, trnK–matK, rps16, trnL–trnF, trnS–trnG and trnT–trnL), and we set PARTITIONFINDER to use the nucleotide models available in BEAST (Drummond et al., 2012).

For all the phylogenetic analyses, the position of *H. palmeri* within Oleinae was investigated using nrDNA, ptDNA, or a combined nrDNA and ptDNA dataset. Maximum likelihood (ML) analyses were performed with RAxML v7.4 (Stamatakis, 2014) using the rapid bootstrap algorithm (Stamatakis, Hoover & Rougemont, 2008), automated bootstopping option, and a GTR+Γ model that was applied to each partition. Bayesian analyses were conducted using MrBayes v3.2.3 (Ronquist et al., 2012) using the best-fit partitioning scheme and substitution models specified by PARTITIONFINDER. Two parallel runs of four chains (one cold and three hot) for 20 million generations were used for the analyses. All other parameters were set to default. The first 2 500 000 generations were discarded as burn-in, and trees were sampled every 1000 generations. Trees were visualized using FigTree v1.4.2 (Rambaut, 2014).

**Molecular dating**

The molecular dating analyses were performed with BEAST v1.8.0 (Drummond et al., 2012) using the combined dataset of ptDNA and nrDNA. As for the MrBayes analyses, we used the best-fit partitioning scheme and substitution models specified by PARTITIONFINDER. A relaxed lognormal molecular clock model was applied with a birth-death speciation process prior. The divergence time between *Fraxininae* and Oleinae was calibrated based on fossil evidence from southeastern North America (Suzuki, 1982; Call & Dilcher, 1992) that was dated to the Middle Eocene (38.0–47.8 Mya; Cohen, Finney & Gibbard, 2013). We thus used 38 Mya as a minimum age. Unlike Hinsinger et al. (2013), we did not use the end of the Middle Eocene (here 47.8 Mya) as a maximum age because it corresponds to the upper uncertainty of the minimum age. We thus used a uniform distribution between 38 and 55 Mya to model the calibration of this node. As suggested by Hinsinger et al. (2013), the divergence between *Fraxinus americana* L. and *F. angustifolia* Vahl. was calibrated based on an Upper Miocene fossil attributed to a species related to *F. angustifolia* (Palamarev, 1989). A uniform distribution between 12 and 38 Mya was applied to this node. The divergence of subgenus *Olea* occurred at least 23 Mya (Muller, 1981; Palamarev, 1989; Terral et al., 2004). This node was calibrated using a uniform distribution between 23 and 38 Mya. Finally, the divergence between *Olea europaea* subsp. *europaea* (Mediterranean olive) and *O. e*. subsp. *cuspidata* (African olive) occurred at least 3.2 Mya (Palamarev, 1989; Terral et al., 2004). This node was calibrated between 3.2 and 23 Mya following a uniform distribution. Analyses were conducted for two independent MCMC runs. Each run consisted of 40 million generations with a sampling frequency every 1000 generations. The two independent runs were then combined with LogCombiner v1.8.0 (in the BEAST package) and the first 4 million generations were discarded as burn-in. The tree posterior distributions were then summarized using TreeAnnotater v1.8.0 (also available in the BEAST package), and the maximum clade credibility tree was visualized using FigTree v1.4.2 (Rambaut, 2014).

**RESULTS**

A total of 10 694 511 of paired-end 100-bp reads were generated with HiSeq technology from the DNA library of *H. palmeri*. A complete chloroplast genome sequence, a nearly complete nuclear ribosomal cluster, and partial nuclear low-copy regions were assembled.

**ASSEMBLY OF THE COMPLETE PLASTID GENOME**

The size of the complete plastid genome of *H. palmeri* is 155 820 base pairs (bp) with an average GC content of 37.8%. The sequence length and GC content are very close to those reported for olive (EMBL accession no FN996972: 155 886 bp, GC content = 37.8%). Overall, 276 035 paired-end reads (2.58% of the total reads) matched to the ptDNA sequence, with an average size of inserts of 118 bp (see Fig. S3). The percentage of ptDNA duplicated paired-end reads was estimated to be 7.98%. After removing duplicates, the mean sequencing depth was estimated to be 329.8× (± 144.9). One hundred thirty plastid genes were annotated (i.e. 85 proteins coding genes, 37 tRNA genes, and eight rRNA genes). The gene order in the *H. palmeri* plastid genome is identical to that described for olive (Mariotti et al., 2010; Besnard et al., 2011) indicating no major re-organization in this genome between these two species of Oleinae.

**ASSEMBLY OF THE NUCLEAR RIBOSOMAL DNA CLUSTER**

An nrDNA contig of 8075 bp with an average GC content of 54.5% was then assembled. This region includes the complete sequence of the 5′ETS, 18S, ITS1, 5.8S, ITS2 and 26S subregions. The segment
used for phylogenetic analyses (including the complete ITS1, 5.8S and ITS2; Hong-Wa & Besnard, 2013) was 757-bp long and shows a GC content of 57.46%. This latter value is similar to those reported in Forestiera Poir. (GC content ranging from 55.89 to 57.67%), which belongs to the same lineage (see below). The non-transcribed region of the intergenic spacer (IGS) remains incomplete due to short repeated and inverted elements (e.g. Maggini et al., 2008) making the assembly very difficult with our approach. Overall, 8281 paired-end reads (c. 0.08% of the total reads) matched to the nrDNA region including the complete sequence of the 18S, ITS1, 5.8S, ITS2 and 26S subregions (5822 bp) with an average size of inserts of 123 bp (see Fig. S3). The percentage of duplicated paired-end reads was 5.63% (532 paired-end reads). After removing duplicates, sequencing depth was estimated to be 254.9× (± 126.9).

ASSEMBLY OF LOW-COPY NUCLEAR GENES

For the five nuclear low-copy regions (with a combined length of about 3750 bp), we isolated a total of 29 paired-end reads (plus one duplicated read that was excluded). The mean sequencing depth was 1.55×. The coverage of the sequences was 66.9% meaning that several parts of these five genes remained unsequenced. Predicted coding sequences did not show any stop codon suggesting that our sequences may encode functional proteins. The sequence and summary of assembly results are given for each gene in Supporting Information (Material S1 and Table S2).

PHYLOGENETIC ESTIMATIONS AND DATING OF LINEAGE DIVERGENCE

The characteristics of the best-partitioning scheme and substitution model inferred by PartitionFinder are: GTR+I+Γ for the subset ITS1 and ITS2, K80+I+Γ for 5.8S, GTR+I+Γ for atpB-rbcL and trnK-matK, GTR+I+Γ for rps16 and trnL-trnF, GTR+I+Γ for trnS-trnG, and finally GTR+Γ for trnT-trnL. Trees obtained from ptDNA and nrDNA sequences are given in Figure 1 and Supporting Information (Supporting Information, Figs S5 to S7). For all of these genomic regions, the topologies obtained with ML vs. Bayesian inference (BI) were mostly congruent. We regard maximum likelihood bootstrap (MLBS) and Bayesian posterior probability (BPP) values of 100–85% and 100–95%, respectively, as strong, 84–75% and 94–85% as moderate, and 74–50% and 84–70% as low support. Overall, the ptDNA and nrDNA topologies were very similar to our recently published phylogenies of subtribe Oleinae (Hong-Wa & Besnard, 2013), but as shown in our previous work, the plastid and nuclear DNA topologies are discordant. In particular, while the monophyly of subtribe Oleinae was strongly supported in the ptDNA tree, the two sampled species of Fraxinus were imbedded in the Oleinae clade in the nrDNA tree (Fig. 1).

Both ptDNA and nrDNA topologies corroborate polyphyly of several genera [i.e. Chionanthus (at least three lineages), Osmanthus (two lineages), and Olea (two lineages)] as previously reported in other studies (Besnard et al., 2009; Yuan et al., 2010; Guo et al., 2011). Our analyses also indicate at least three separate Oleinae lineages in America that include the following taxa: (i) Chionanthus filiformis (Brazil), Ch. panamensis (Standl.) Stern (Mexico), and Haenianthus spp. (Greater Antilles), (ii) Forestiera spp. (primarily distributed in North America), Priogymnanthus hasslerianus (Paraguay), and Hesperelaea palmeri (Guadalupe Island), and (iii) Chionanthus virginicus L. and Cartreana americana (L.) Nesom (North America), but this last lineage is supported only in the ptDNA tree (placement of these two species remains unresolved based on the nrDNA data).

Based on both ptDNA and nrDNA data, Hesperelaea belongs to an American lineage that includes all sampled species of Forestiera and both species of Priogymnanthus (the latter genus was unsampled for nrDNA). The BI and ML trees reconstructed with ptDNA markers indicate that this clade is strongly supported (Fig. 1A) and sister to Noronhia, an African-Malagasy lineage. The support for this deeper node, however, is low in the ML analysis (MLBS = 73%). Relationships among Forestiera, Hesperelaea, and Priogymnanthus are not resolved in the ptDNA tree; a polytomy is observed with three main lineages that correspond to the three genera. Based on nrDNA sequences (Fig. 1B), the Forestiera–Hesperelaea clade (Priogymnanthus being unavailable for nrDNA) is also supported (BPP = 98%; MLBS = 57%) and placed in a clade that includes species of Osmanthus, Philllyrea L., Picconia A. DC., Notelaea Vent., Nestegis Raf., and Olea subgenus Tetrapius (Lour.) P.S. Green plus a few species of Chionanthus. Many nodes in this large clade, however, are unresolved or weakly supported (Fig. 1B).

The dating analysis was conducted based on all sequence data (ptDNA and nrDNA; Fig. 2). The topology obtained from Beast was very similar to those obtained with ML and BI using MrBayes (Figs S8 and S9). The Beast analysis showed the age of the Oleinae ancestor at 38.99 Mya [95% highest probability density (HPD): 31.84–46.88]. The last shared ancestor of Hesperelaea and Priogymnanthus was estimated at 19.74 Mya [95% HPD: 12.49–27.49], while the ancestor of Forestiera, Hesperelaea

© 2015 The Linnean Society of London, Biological Journal of the Linnean Society, 2016, 117, 44–57
Figure 1. Maximum likelihood topologies inferred with RAXML from analyses using (A) combined plastid DNA regions (rps16, atpB-rbcL, trnL-trnF, trnK-matK, trnS-trnG and trnT-trnL) and (B) nuclear ribosomal DNA (ITS). The trees are rooted with subtribe Schreberinae. Bold values above branches are Bayesian posterior probabilities (BPP%) and the values below branches denote maximum likelihood bootstrap support (MLBS%). The subtribes Schreberinae and Fraxininae are distinguished with blue and green shaded areas, respectively. In addition, we indicated three monophyletic groups of Oleinae (Olea s.s., Forestiera and Noronhia) in the two topologies. American lineages/species of Oleinae are highlighted with pink boxes.
Figure 2. Calibrated phylogenetic trees of the Oleinae obtained from the BEAST analysis of the complete dataset (ptDNA and nrDNA). Branch lengths are proportional to time and scale is given in millions of years. Bayesian support (posterior probabilities) of nodes is indicated below branches. The median age of each node is given above branches. The 95% highest probability density (HPD) for the estimated divergence times is indicated with grey bars.

© 2015 The Linnean Society of London, Biological Journal of the Linnean Society, 2016, 117, 44–57
and *Priogymnanthus* was estimated at 21.63 Mya [95% HPD: 14.89–29.00].

**DISCUSSION**

**SEQUENCE QUALITY AND PHYLOGENETIC SIGNALS**

A complete plastid genome and a nearly complete nuclear ribosomal cluster were generated for *Hesperelaea palmeri* in spite of a highly degraded DNA extract (Fig. S2). The GC contents of these molecules were very similar to those reported in other *Oleinae* species, suggesting that the deamination rate (which may lead to a decrease of the GC content) has been low for the *Hesperelaea* herbarium-specimen DNA, as reported recently for historical museum DNA from both insects and other plants (Staats et al., 2013). The high sequencing depth for chloroplast and ribosomal DNA also allowed minimizing errors because most sites were sequenced more than 250 times. In addition, a few reads of nuclear low-copy genes were recovered. We estimated that the sequencing depth of the low-copy nuclear gene was 1.55x. Such a low sequencing depth for the low-copy nuclear genes is similar to that reported for museum specimens of birds using the same approach (Besnard et al., 2015). With a twenty-fold increase of the sequencing depth and using a reference genome of a related genus (e.g. *Olea* and/or *Fraxinus*), it could be possible to assemble nuclear contigs of several thousand nucleotides.

Our ability to generate high-quality ptDNA and nrDNA sequence data for *Hesperelaea* allowed us to investigate the phylogenetic placement of this genus within the *Oleinae*. Phylogenies estimated separately with ptDNA and nrDNA data corroborate polyphyly of several *Oleinae* genera, as previously shown by other authors (Besnard et al., 2009; Yuan et al., 2010; Guo et al., 2011; Hong-Wa & Besnard, 2013). The tree topologies for subtribe *Oleinae* based on ptDNA vs. nrDNA data, however, show deep incongruence, as very similarly found in a previous study on this group (Hong-Wa & Besnard, 2013). These topologies may differ for multiple reasons, which remain to be clearly identified but are not central to the concerns of our study. The particularly anomalous position of *Fraxinus* in the ITS tree possibly reflects evolutionary properties of nuclear ribosomal DNA (Material S2) that may potentially result in disruption of phylogenetic signal. In addition, the short length and rapid evolution of the ITS region may not always provide sufficient signal for molecular phylogenetic resolution, especially for deep divergences (Malé et al., 2014). For this reason, the deeper nodes of the phylogeny based only on nrDNA data (Fig. 1B) should be taken with caution.

**PHYLOGENETICS OF HESPERELAEA CORROBORATE THE UNIQUENESS OF AN EXTINCT LINEAGE**

In the present study, the ptDNA and nrDNA trees allow for the identification of three separate *Oleinae* lineages in America (but the third is not resolved in the nrDNA tree): *Hesperelaea–Forestiera–Priogymnanthus*, *Chionanthus panamensis–Chionanthus filiformis–Haenianthus* spp., and *Cartrema americana–Chionanthus virginicus*. These three lineages are distinctly related based on both ptDNA and nrDNA markers but their placements differ in the two phylogenetic hypotheses (Fig. 1). Irrespective of these differences, our results indicate the presence of three divergent *Oleinae* lineages in the New World since the Late Oligocene or the Early Miocene (Figs 1 and 2).

In the previous work of Wallander & Albert (2000), the phylogenetic positions of *Hesperelaea* and *Priogymnanthus* were not resolved within subtribe *Oleinae* due to a lack of DNA information. Here, our phylogenetic analyses indicate that these two genera plus *Forestiera* belong to a monophyletic lineage with strong to moderate support. Relationships among the genera *Hesperelaea*, *Priogymnanthus*, and *Forestiera* remain unresolved, however, and this polytomy may reflect either an ancient, rapid radiation or just a lack of information in our data (Figs 1 and 2). The affinity between *Forestiera* (distributed in North and Central America, plus one species in Ecuador; Cornejo, 2006) and *Priogymnanthus* (South America) was already suspected by Green (1994) based on morphological traits but, to our knowledge, a relationship of *Hesperelaea* with these two genera has never been suggested.

According to the phylogenetic dating, the stem lineage of *Hesperelaea*, *Priogymnanthus*, and *Forestiera* diverged during the Early Miocene (Fig. 2). Marked climatic and geological changes since that time (e.g. Axelrod, 1992), including onset of the summer-drying trend that ultimately gave rise to the Mediterranean-like conditions that characterize Guadalupe Island and other areas of the California Floristic Province (CA-FP), and extensive environmental heterogeneity in general have been implicated in diversification and persistence of CA-FP lineages (Raven & Axelrod, 1978; Lancaster & Kay, 2013). Such ancient divergence of the *Hesperelaea* lineage relative to the age of Guadalupe Island indicates that *Hesperelaea* can be considered an insular paleoeometric (see below), as befits its taxonomic status as a monotypic genus (Gray, 1876). The extinction of *Hesperelaea* hence represents the loss of the last known member of a unique, old lineage that evolved independently for over ten million years. The loss of such a relict taxon is unfortunately not surprising in remote, insular environments, which often represent a sanctuary for peculiar lineages prone to extinction.
especially from impacts of invasive alien species (Donlan & Wilcox, 2008; Sax & Gaines, 2008).

**On the biogeographic history of Hesperelaea**

As noted above, divergence among Hesperelaea, Forestiera, and Priogymnanthus very likely preceded the formation of Guadalupe Island (7 ± 2 Mya; Hubbs, 1967; Engel & Engel, 1970). The colonization of Guadalupe Island by the Hesperelaea ancestor thus may have occurred from the American continent as early as the Late Miocene, but it could have reached Guadalupe Island after migration via one or more islands (or temporarily exposed seamounts) of the southern Californian region. Considering the long divergence time between Hesperelaea and its closest relatives, Forestiera and Priogymnanthus, the ancestral Hesperelaea lineage may have occurred on the American continent long before its arrival on Guadalupe Island. The flora of the California Islands is known for examples of woody endemics (e.g. Lyonothamnus A. Gray, Rosaceae) that are represented only as fossils from the Miocene and Pliocene of continental western North America (Axelrod, 1967a, 1967bb; Erwin & Schorn, 2000). The general trend toward a cooler and drier climate since the Mid-Miocene was associated with expansion of drought tolerant vegetation and the loss of many woody plant lineages in western North America, while some tree lineages were able to persist only in insular environments (Axelrod, 1967b, 1992).

Of the two genera most closely related to Hesperelaea, only Forestiera is found in North America, and only one modern taxon in this genus [F. pubescens Nutt. var. parviflora (A. Gray) G.L. Nesom] occurs in the CA-FP, in coastal and interior ranges (extending to the semi-arid southwest USA and northwest Mexico; Nesom, 2009) while another species (F. macrocarpa Brandegee) occurs in the subtropical Cape Region of southern Baja California Sur (Wiggins, 1980). The range of F. pubescens extends even further eastward to eastern Texas and Oklahoma, where var. pubescens occurs (Nesom, 2009). Although taxon sampling and strength of phylogenetic resolution within Forestiera were not sufficient to address historical biogeography of the genus, the current centre of diversity of Forestiera is in Meso-America, the northern Caribbean, and the southeastern USA (Gray, 1860; Standley, 1924; Johnston, 1957). Based on those distributional considerations, we suspect that the CA-FP distribution of F. pubescens does not reflect an ancestral area for the species or for the genus Forestiera. The hypothesis of an Early or Middle Pliocene migration to California from the south or the east is consistent with reported pollen records (W.S. Ting in Axelrod, 1967b) and a macrofossil of F. buchananensis Condit, which has been suggested to be related to F. pubescens (Condit, 1944; Axelrod, 1980). During the Early Miocene (Burdigalian), another fossil assigned to Forestiera was reported in California (at Tehachapi; Axelrod, 1939), but its precise taxonomic assignment remains in question. In light of the high support for monophyly of Forestiera, we hypothesize that the ancestor of Hesperelaea descended from a mainland population that diverged prior to the crown-group diversification of Forestiera, or at least prior to diversification of the taxa of Forestiera sampled here. The mainland representatives of that Hesperelaea lineage then probably declined during the Late Miocene or Pliocene, as similarly suggested for many woody taxa in western North America such as Lyonothamnus spp. (Axelrod, 1967b; Erwin & Schorn, 2000).

**Perspectives on the use of NGS for adding value to herbarium specimens and developing phylogenomics of the olive tribe**

In our study, we had a unique opportunity to generate genomic data from an extinct monotypic genus. DNA extraction required the sampling of one leaf (from the fragment packet) but the generation of genomic data enhanced the value of the specimen and compensated for this minimal damage to the collection (Suarez & Tsutsui, 2004; Wandeler et al., 2007; Besnard et al., 2014). Our results demonstrate the high potential of NGS for the analysis of museum specimens of extinct or very rare species in order to evaluate their taxonomic status or genetic attributes. This approach also may help more generally with the inventory and identification of taxa that are only known from museum collections. For example, these new technologies could accelerate the discovery of undescribed species already represented in herbaria (Bebber et al., 2010). Indeed, many unusual specimens remain unidentified even at the genus or family level, and on Guadalupe Island in particular, one such plant taxon also collected by E. Palmer in 1875 (in sterile condition) that was presumably endemic to the island (‘Planta sp.’ in Moran, 1996) deserves to be analyzed with our approach.

Although we generated a huge quantity of DNA information on an old herbarium sample of a critical Oleaceae species, phylogenetic reconstructions based on these data were limited due to the lack of Oleaceae genomic sequences in public databases. Indeed, only ten complete plastid genomes of Oleaceae are presently available (on January 23rd, 2014) and these data are not representative of the diversity of this family; to date, nine of those sequences of Oleaceae have been generated on Olea subgenus Olea (Lee et al., 2007; Mariotti et al., 2010; Besnard et al., 2011). Similarly, the nuclear ribosomal cluster is only
available for the Mediterranean olive tree, while ITS and ETS regions have been extensively used in phylogenetic reconstructions in several Oleaceae genera (Jeandroz et al., 1997; Li et al., 2002; Wallander, 2008; Besnard et al., 2009; Yuan et al., 2010; Hinsinger et al., 2013; Hong-Wa & Besnard, 2013). To reconstruct robust phylogenetic trees of Oleinae, the genome skimming approach needs to be used with a sampling that represents all known major lineages (Wallander & Albert, 2000; Hong-Wa & Besnard, 2013). In addition, many Oleinae species are rare and a few have been collected only once, more than 100 years ago [e.g. for Noronhia see Hong-Wa & Besnard, (2014)]; here we showed that genomic analysis of these species is not out of reach. Such data will help to reconstruct a backbone tree that will allow the integration of data from different phylogenetic studies. In turn, this phylogenomic framework will allow inferences on the worldwide colonization of the olive tribe and the evolution of some key traits (such as the breeding system transitions; Wallander, 2008). In addition, the use of low-copy genes for phylogenetic inference needs to be expanded beyond the genus or species complex level [Fraxinus (Hinsinger et al., 2013) and Olea (Besnard & El Bakkali, 2014)] to the family level in Oleaceae. Genome skimming can be used to extract reads of known single-copy genes, but with low coverage (e.g. Straub et al., 2012; Besnard et al., 2013, 2015); the development of other methods (e.g. gene bait approach; Li et al., 2013) should make the generation of such data more efficient in the near future, even from museum specimens.

ACKNOWLEDGEMENTS

We thank the Missouri Botanical Garden (MO) for permission to extract DNA. This work was funded by the fellowship ANR-12-AGRI-0002 (ARIMNET 2011 – PestOlive) and was conducted in the Laboratoire Evolution and Diversité Biologique (EDB) part of the LABEX entitled TULIP managed by Agence Nationale de la Recherche (ANR-10-LABX-0041). We also thank Sebastien Cally and Yann Bourgeois for valuable advice in data analyses, Gabriel Frey for providing a leaf sample of Chionanthus filiformis, and two anonymous reviewers for providing helpful comments on the initial version of the manuscript. LZ was supported by the University of Al Furat, Syria.

REFERENCES


Rambaut A. 2014. FigTree v1.4.1. Available at: http://tree.bio.ed.ac.uk/software/figtree/


Stamatakis A, Hoover P, Rougeme
SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher’s website:

Supplementary Material S1. Sequences of the five low-copy gene segments.

Supplementary Material S2. On the incongruent phylogenetic signals harbored by plastid and ribosomal DNAs in Oleaceae.

Figure S1. Herbarium samples of Hesperelaea palmeri A. Gray (E. Palmer 81, MO).

Figure S2. Quality control of the DNA extract and library of Hesperelaea palmeri.

Figure S3. Distribution of the size inserts of paired-end reads used to assemble the plastid DNA and the nuclear ribosomal DNA regions.

Figure S4. Maximum likelihood (ML) tree inferred from plastid DNA regions.

Figure S5. Bayesian inference (BI) tree inferred from plastid DNA regions.

Figure S6. ML tree inferred from nrDNA.

Figure S7. BI tree inferred from nrDNA.

Figure S8. ML tree inferred from a combined nrDNA and ptDNA dataset.

Figure S9. BI tree inferred from a combined nrDNA and ptDNA dataset.

Table S1. List of specimens analysed in the present study, with accession numbers of the five DNA regions used for phylogenetic reconstructions of subtribe Oleinae.

Table S2. Summary of assembly results on the five low-copy genes.