# Bridging the micro- and macroevolutionary levels in phylogenomics: Hyb-Seq solves relationships from populations to species and above

**Tamara Villaverde[1]*, Lisa Pokorny[2]*, Sanna Olsson[3], Mario Rincón-Barrado[1], Matthew G. Johnson[4,5], Elliot M. Gardner[6], Norman J. Wickett[5,7], Julià Molero[8], Ricarda Riina[1†] and Isabel Sanmartín[1†]**

[1]Real Jardín Botánico (RJB-CSIC), Plaza de Murillo 2, 28014 Madrid, Spain; [2]Comparative Plant and Fungal Biology Department, Royal Botanic Gardens, Kew, Richmond, TW9 3DS, UK; [3]Department of Forest Ecology and Genetics, INIA Forest Research Centre (INIA-CIFOR), Ctra. de la Coruña km. 7.5, 28040 Madrid, Spain; [4]Department of Biological Sciences, Texas Tech University, 2901 Main St, Lubbock, TX 79409-43131, USA; [5]Department of Plant Science and Conservation, Chicago Botanical Garden, 1000 Lake Cook Road, Glencoe, IL 60022, USA; [6]The Morton Arboretum, 4100 Illinois 53, Lisle, IL 60532 USA; [7]Program in Plant Biology and Conservation, Northwestern University, 2205 Tech Drive, Evanston, IL 60208 USA; [8]Laboratori de Botànica, Departament de Biologia, Sanitat i Medi Ambient, Facultat de Farmàcia, Universitat de Barcelona, 08028 Barcelona, Spain

Authors for correspondence:
*Tamara Villaverde*
*Tel: +34 91 4203017*
*Email: tvilhid@gmail.com*

*Isabel Sanmartín*
*Tel: +34 91 4203017*
*Email: isanmartin@rjb.csic.es*

Received: *1 March 2018*
Accepted: *4 June 2018*

*New Phytologist* (2018) **220:** 636–650
**doi**: 10.1111/nph.15312

**Key words:** Euphorbiaceae, Hyb-Seq, phylogenomics, plastid skimming, Rand Flora, target enrichment.

## Summary

- Reconstructing phylogenetic relationships at the micro- and macroevoutionary levels within the same tree is problematic because of the need to use different data types and analytical frameworks. We test the power of target enrichment to provide phylogenetic resolution based on DNA sequences from above species to within populations, using a large herbarium sampling and *Euphorbia balsamifera* (Euphorbiaceae) as a case study.
- Target enrichment with custom probes was combined with genome skimming (Hyb-Seq) to sequence 431 low-copy nuclear genes and partial plastome DNA. We used supermatrix, multispecies-coalescent approaches, and Bayesian dating to estimate phylogenetic relationships and divergence times.
- *Euphorbia balsamifera*, with a disjunct Rand Flora-type distribution at opposite sides of Africa, comprises three well-supported subspecies: western Sahelian *sepium* is sister to eastern African-southern Arabian *adenensis* and Macaronesian-southwest Moroccan *balsamifera*. Lineage divergence times support Late Miocene to Pleistocene diversification and climate-driven vicariance to explain the Rand Flora pattern.
- We show that probes designed using genomic resources from taxa not directly related to the focal group are effective in providing phylogenetic resolution at deep and shallow evolutionary levels. Low capture efficiency in herbarium samples increased the proportion of missing data but did not bias estimation of phylogenetic relationships or branch lengths.

## Introduction

Evolutionary biologists, in their efforts to determine which factors govern biodiversity dynamics, have used two approaches that differ primarily in the time frame in which they operate: microevolutionary processes (genetic drift, mutation, migration) act mostly on individuals within populations (recent time), while macroevolutionary processes (speciation, extinction, dispersal) focus on diversification at and above species level in relation to environments and over longer timescales (Benton, 1995; Reznick & Ricklefs, 2009). Species experience population size changes and range shifts in response to climatic oscillations, structuring their gene pools across their geographic ranges; these processes are often addressed through population-genetic or demographic studies (Hewitt, 2000, 2004; Davis & Shaw, 2001; Parmesan, 2006). However, over longer timescales (hundreds of thousands to millions of yr), the signature of these microevolutionary processes can become saturated, in which case phylogenies addressing the order and timing of diversification events provide more information on the evolutionary fate of lineages (Svenning *et al.*, 2015). Although micro- and macroevolution operate over different geographic and temporal scales, stochastic processes have been shown to leave their imprint on deep phylogenetic histories (Oliver, 2013).

Bridging the micro- and macroevolutionary scales is difficult because the types of molecular data, sampling schemes, and phylogenetic models that are often employed differ when analysing population and species-level relationships. Microevolutionary studies typically use multiple individuals per population and rely on repeated DNA or polymorphic molecular markers (e.g. simple

*These authors contributed equally to this work.
†Joint senior authors.

© 2018 The Authors
*New Phytologist* © 2018 New Phytologist Trust
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

sequence repeat, amplified fragment length polymorphism), as they, in comparison to DNA sequence data, provide more variability at the population level and can be used to detect recent admixture (Guichoux *et al.*, 2011). Yet these types of markers are not readily analysable using the standard molecular substitution models employed in phylogenetics based on DNA sequences (e.g. Tavaré, 1986); instead, approximate models (e.g. Luo *et al.*, 2007) or still debated statistical models are employed (Hobolth *et al.*, 2008; Wu & Drummond, 2011). Conversely, macroevolutionary studies often rely on DNA sequences from only one individual per species, but the use of a single or few genetic regions is usually not enough to obtain well-resolved and supported phylogenies (e.g. Pelser *et al.*, 2007). Furthermore, these analyses are limited to parts of the genome that may contain conflicting signals, leading to spurious phylogenetic relationships (Shen *et al.*, 2017).

High-throughput sequencing (HTS) provides an avenue for bridging the micro- and macroevolutionary gap in phylogenetics by scaling up the number of loci and individuals within populations and across species that can be sequenced at a reasonable cost (Barrett *et al.*, 2016). Target sequencing has gained popularity in recent years because, unlike whole-genome sequencing (WGS), which requires fresh material (but see Dentinger *et al.*, 2016), this technique can work with both fresh and old museum material (e.g. Lemmon & Lemmon, 2013; Barrett *et al.*, 2016) and generally demands less complex bioinformatics (e.g. Fér & Schmickl, 2018). Reduced representation techniques such as restriction site-associated DNA sequencing (RADseq, Baird *et al.*, 2008) or genotype-by-sequencing (Elshire *et al.*, 2011) work well for population-level analysis and do not require a reference genome. Yet, their outputs are single nucleotide polymorphisms (SNPs) or small reads from anonymous loci, which make assessment of gene orthology challenging in the case of deep divergences (e.g. Rubin *et al.*, 2012; although see Hipp *et al.*, 2018). A modification of target sequencing, Hyb-Seq, that is, hybrid enrichment with genome skimming (Weitemier *et al.*, 2014), is promising for nonmodel organisms because it generates thousands of DNA sequences from low or single-copy nuclear genes (LCNGs), combining exon capture with genome skimming of intronic and intergenic regions, flanking the targeted exon regions. Hyb-Seq also generates highly repetitive DNA from organellar genomes as a by-product; the latter is important to detect reticulate evolution and introgression.

Here, we examine the efficiency of Hyb-Seq to provide phylogenetic resolution at deep and shallow evolutionary levels using fine-scale within-population sampling from both silica-dried tissue and old herbarium material to solve multilevel relationships. Fragmented DNA from museum samples can lead to a potential bias in phylogenomic analyses (Sayyari *et al.*, 2017); however, only few studies have tested this effect within species based on a limited number of herbarium samples (e.g. Hart *et al.*, 2016). As a case study, we use *Euphorbia balsamifera* Aiton (the sweet tabaiba, Euphorbiaceae; Fig. 1), a taxon exhibiting a deep intraspecific divergence (*c.* 3.8 million yr ago (Ma); Peirson *et al.*, 2013; Pokorny *et al.*, 2015) and a disjunct distribution spanning thousands of kilometres on opposite sides of Africa, the so-called Rand Flora pattern (Christ, 1892; Andrus *et al.*, 2004; Sanmartín *et al.*, 2010). In this biogeographic pattern, unrelated plant lineages show similar disjunct distributions with sister taxa in distantly located regions along African continental margins and adjacent islands, such as Macaronesia-northwest Africa, western Africa mountains, Horn of Africa-South Arabia, eastern Africa and southern Africa. *Euphorbia balsamifera* is a diploid species that belongs to section *Balsamis* Webb & Berthelot within subgenus *Athymalus* Neck. ex Rchb. Its current taxonomy (see Supporting Information Notes S1) recognizes two subspecies: *balsamifera*, distributed in all major islands in the Canaries, the northwestern Atlantic coast of Africa and western Sahel; and *adenensis* (Deflers) P.R.O. Bally, occurring in eastern Africa, southern Arabia and the Socotra Archipelago (Govaerts *et al.*, 2000; Peirson *et al.*, 2013). There is a third taxon, ssp. *sepium* N.E.Br. (Molero *et al.*, 2002), currently synonymized under ssp. *balsamifera*, which has been applied to all the populations of western Sahel and whose taxonomic status and phylogenetic position have never been addressed using molecular data (Bruyns *et al.*, 2011; Peirson *et al.*, 2013). Difficulties in obtaining fresh material from remote and now politically unstable countries across eastern Africa and the Middle East have restricted phylogenetic studies of *Euphorbia* from those areas to multicopy DNA regions like nuclear internal transcriber spacers (ITS) and plastid markers (Bruyns *et al.*, 2011; Peirson *et al.*, 2013; Fig. S1); these might be easily Sanger-sequenced from herbarium specimens.

In this phylogenomic study, the first within Euphorbiaceae, we designed probes to capture 431 orthologous low-copy nuclear loci (*c.* 709 kbp) using a significant amount of herbarium material. Our specific aims were to test the utility of probes based on genomic resources from taxa not directly related to the focal group in providing phylogenetic resolution from within populations and species (*E. balsamifera*) to above-species level (within section *Balsamis* and subgenus *Athymalus*); to assess the effect of extensive use of highly degraded DNA from herbarium material on capture success and potential bias in phylogenomic inference at different evolutionary scales; to test the ability of our probes to obtain off-target chloroplast sequences; and to test the monophyly of *E. balsamifera* ssp. *sepium* and its phylogenetic relationship within *E. balsamifera* and estimate lineage divergence times.

## Materials and Methods

### Sampling

We started with 165 samples, of which 121 were successfully amplified (Table S1). In the final dataset, *Euphorbia balsamifera* s.l. is represented by 40 populations (i.e. localities; Table S1, Fig. 1g,h). Seven populations of the Canarian ssp. *balsamifera* included at least eight individuals, and one (Telde, IS462) came from a single individual. Also several populations were represented by a single individual: ssp. *balsamifera* from W Africa (three), ssp. *adenensis* (19) and ssp. *sepium* (10). Our dataset also included at least one sample from other representatives within section *Balsamis* (*E. larica* Boiss., *E. masirahensis* Ghaz. and *E. noxia* Pax), and subgenus *Athymalus* (*E. antso* Denis, sect.

© 2018 The Authors
*New Phytologist* © 2018 New Phytologist Trust

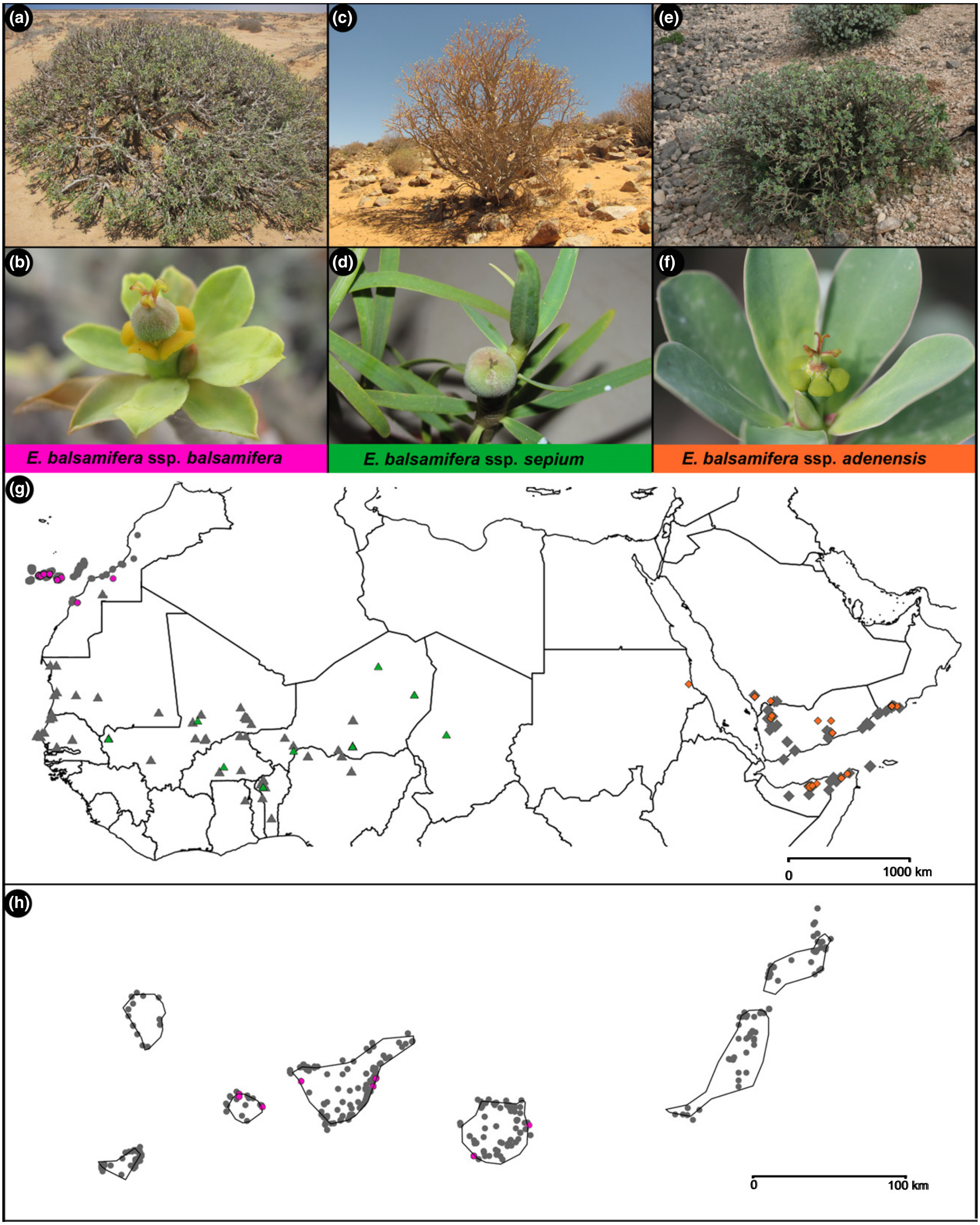
(a)
(c)
(e)
(b)
(d)
(f)
E. balsamifera ssp. balsamifera
E. balsamifera ssp. sepium
E. balsamifera ssp. adenensis
(g)
0
1000 km
(h)
0
100 km

**Fig. 1** Morphology and geographic distribution of *Euphorbia balsamifera*. (a) *Euphorbia balsamifera* ssp. *balsamifera* from the coast of Western Sahara. (b) Cyathium of ssp. *balsamifera* from Tenerife (Canary Islands). (c) *Euphorbia balsamifera* ssp. *sepium* from inland populations in Western Sahara near the border with Mauritania. (d) Fruit and long narrow leaves typical of ssp. *sepium*. (e) Specimen of *E. balsamifera* ssp. *adenensis* from Oman. (f) Cyathium and apical leaves of ssp. *adenensis*. (g, h) Geographic distribution of *Euphorbia balsamifera* in the Canary Islands, North Africa, and Arabian Peninsula based on herbarium specimens: ssp. *balsamifera* (circles), ssp. *sepium* (triangles), and ssp. *adenensis* (diamonds); (h) Detailed distribution of ssp. *balsamifera* in the Canary Islands. Fuchsia, green and orange symbols represent sampled DNA specimens of each taxon included in the phylogeny shown in Fig. 3(a), whereas grey symbols (all shapes) represent herbarium records without genomic data. Photos by: (a–d) R. Riina; (e, f) J. J. Morawetz.

© 2018 The Authors
*New Phytologist* © 2018 New Phytologist Trust

*Antso* P.E.Berry; *E. hadramautica* Baker, sect. *Pseudacalypha* Boiss. in Candolle; *E. matabelensis* Pax and *E. smithii* S.Carter, sect. *Lyciopsis* Boiss. in Candolle; *E. socotrana* Balf.f. and *E. scheffleri* Pax, sect. *Somalica* S.Carter; and *E. davyi* N.E.Br., sect. *Anthacanthae* Lem.). DNA came from silica-dried tissue and herbarium material (Table S1).

## Probe design and sequence capture

As genomic resources for the design of probes, we used the transcriptomes of two *Euphorbia* species from subgenera *Chamaesyce* and *Esula* (RHAU and PXYR, *E. mesembryanthemifolia* Jacq. and *E. pekinensis* Rupr., respectively) – available through the 1KP initiative (www.onekp.com/public_data.html) – and *Ricinus communis* L. (Euphorbiaceae) as the reference genome (v0.1, available at www.phytozome.net; Chan *et al.*, 2010) to estimate intron/exon boundaries. MarkerMiner v.1.0 (Chamala *et al.*, 2015) was employed to identify LCNGs' orthologous loci, which were used to develop the gene target probes. Probes were designed for 431 orthologous LCNGs (Table S2) ranging from 639 to 6774 bp with at least one exonic fragment per LCNGs longer than 500 bp, representing a total length of *c.* 709 kbp. Arbor Biosciences (Ann Arbor, Michigan, USA) manufactured a target enrichment kit with in-solution biotinylated probes. The 120-mer probes (20 002 total baits) were tiled at 2× density for the sequences of the two *Euphorbia* transcriptomes and 1.5× for the sequences of *R. communis.*

Genomic DNA (extraction protocols in Methods S1) from silica-dried samples and recent herbarium collections, which had fragment sizes > 550 bp, were sonicated to a target fragment size of 550 bp using a Covaris E220 Focused-ultrasonicator (Wohurn, MA, USA). The remaining samples for which the average fragment size was < 550 bp were not sonicated and thus resulted in libraries with insert size < 550 bp. Sequencing libraries were prepared using the Illumina TruSeq Nano HT DNA Kit (Illumina Inc., San Diego, CA, USA). Indexed samples were pooled in approximately equal quantities (typically 16–22 samples per equimolar 1000 ng pool, when possible). Each pool was enriched using the custom baits kit following the manufacturer's protocol, at 60–65°C for 16–24 h. Pooling of samples was phylogenetically informed; we avoided including in the same pool intra- and interspecific sampling, for example, other taxa and *E. balsamifera* specimens, to prevent a large number of baits being sequestered by ingroup samples. Enriched products were PCR-amplified for 14 cycles and purified using the QIAquick PCR purification kit (Qiagen). All sequencing took place on an Illumina MiSeq at The Field Museum of Natural History (Chicago, IL, USA). The 96 silica-dried samples (five pools) were sequenced on two 2 × 300 bp runs (600 cycle v3). The 48 herbarium samples (two pools) were sequenced on one 2 × 75 bp (150 cyle v3) run; as a result of low enrichment, these pools were re-enriched following the myBaits® protocol using previously enriched product as the input. The double-enriched products were then sequenced on another 2 × 75 bp run. To ensure recovery of chloroplast sequences, we added 10% unenriched library to all sequencing runs. Further details can be found in Methods S1.

## Data processing and phylogenetic analysis of targeted nuclear loci

Demultiplexed sequences were quality-filtered using Trimmomatic (Bolger *et al.*, 2014) to remove adapter sequences. We also removed poorly aligned regions from alignments using trimAL v.1.2 (removing all columns with gaps in > 50% of the sequences or with a similarity score < 0.001, unless this removes > 40% of the columns in the original alignment; Capella-Gutiérrez *et al.*, 2009). The HybPiper pipeline (v.1.0; Johnson *et al.*, 2016) was used to assemble loci. Summary statistics were obtained using Samtools v.1.8 (Li *et al.*, 2009). Orthologous sequences from 428 nuclear loci containing only exons were aligned using Mafft v.7.222 (Katoh & Standley, 2013). We evaluated gene capture success as the percentage of summed captured length of all target loci per individual divided by the summed mean length of all reference loci. We considered an alignment to be poor quality if the aligned pairwise identity was < 65.5% and the percentage of identical sites was < 15% (Table S3). This resulted in 132 loci removed from the final dataset and subsequent analyses. Therefore, the final dataset included exons from 296 loci (486 878 bp, 121 samples).

All 296 exon matrices were concatenated into a supermatrix and a phylogenetic tree was built by maximum likelihood (ML) after automatic model selection using ModelFinder (Kalyaanamoorthy *et al.*, 2017) in IQ-Tree v.1.4.2 (Nguyen *et al.*, 2015) (1000 ultrafast bootstraps '-bb', '-m TEST') and RAxML v.8 2.9 using the gtrcat model with 200 fast bootstraps followed by slow ML optimization (default '–fa' search; Stamatakis, 2014). Alternatively, we used methods that implement the multispecies coalescent models (MSC), where individual genes are allowed to evolve within the species tree under independent tree topologies to account for gene tree discordance. In particular, we used Astral-ii v.2.4.7.7 (Mirarab & Warnow, 2015) with default parameters to estimate a species tree from the individual gene trees by maximizing the number of quartet trees (sets of four species) shared between gene trees and the species tree (Chou *et al.*, 2015; Mirarab & Warnow, 2015). This method is not based on parameter estimation and thus is efficient with large-genomic datasets; besides it can generate species trees that are statistically consistent with the MSC model. New alignments and ML phylogenies were created for each of the 296 exon matrices with Practical Alignment using SATé and Transitivity (Pasta; Mirarab *et al.*, 2015). We ran Pasta with default parameters and different options for alignment: Mafft or Muscle (Edgar, 2004); merging: Opal (Wheeler & Kececioglu, 2007) or Muscle; and ML tree estimation: FastTree (Price *et al.*, 2010) or RAxML. The resulting ML trees were used to infer the coalescence-based species phylogeny in Astral-ii with local posterior probabilities estimated to provide support for relationships. Because of concerns with the effect of alignment trimming on phylogenetic estimation (Tan *et al.*, 2015), we ran our analyses with and without trimmed sites.

Additionally, we used SVDquartets to infer a species tree under the coalescent framework (Chifman & Kubatko, 2014, 2015). Unlike Astral-ii, SVDquartets does not require *a priori* inference of individual gene trees. Instead, quartet trees are evaluated using an algebraic approximation (i.e. the expected rank of the flattened rate matrix under MSC) and combined into a species tree using a supertree approach. Originally designed for SNPs, SVDQUARTETS has been shown to perform well on multilocus datasets even though it violates the assumption that sites are independent (Chifman & Kubatko, 2015). We used the concatenated 296-exon supermatrix as input with the option evaluating 100 000 random quartets or all possible quartets if lower than 100 000. Clade support was assessed by running 100 bootstrap replicates, using the MSC model. The analysis was performed in PAUP* 4.0a146 (Swofford, 2002). Both ASTRAL-II and SVDQUARTETS were run with samples unassigned ('blind') or assigned to the lineages identified in the 'blind' analyses: *sepium*, *adenensis*, and within *balsamifera* (Western Sahara, Gran Canaria, East Tenerife, West Tenerife, and La Gomera).

To explore the effect of more than one single copy per locus, we used HYBPIPER scripts to generate a list of potential paralogues for each sample and locus (Table S4). If a gene is identified as a paralogue in several samples, that gene is a candidate for potential duplication and paralogy; if only in one sample, it is probably a different allele; if a sample contains multiple paralogues, that sample is a potential polyploid. Thus, we ran different analyses in ASTRAL-II by sequentially removing loci with paralogues in at least one sample, two or more samples, or > 10 samples, and compared the resulting topologies with the full dataset. We also evaluated gene tree–species tree discordance by computing the level of support/conflict provided by each of the 296 gene trees used in the analyses for bipartitions shown in the ASTRAL-II species tree, as well as for other alternative bipartitions. We followed the procedure described in Smith *et al.* (2015) and used Matt Johnson's scripts (https://github.com/mossmatters/MJPythonNotebooks/blob/master/PhyParts_PieCharts.ipynb) to visualize the results. This procedure allowed us to evaluate how many genes support or conflict with individual bipartitions within the species tree, that is, if there is a dominant tree topology in the gene trees or, if there is conflict, whether this stems from an alternative tree topology or from many low-frequency alternative gene topologies or lack of support for conflicting bipartitions (Smith *et al.*, 2015).

Finally, we also analysed introns and supercontig (exon + intron) matrices generated by HYBPIPER. As before, we considered an alignment to be poor quality if the percentages of identical sites were < 40% and < 25%, respectively. This resulted in matrices of 15 out of 404 introns and 112 out of 424 supercontigs (Tables S5, S6). Phylogenetic trees using concatenated introns (97 samples, 16 817 bp) and supercontigs (117 samples, 347 878 bp) matrices were analysed with IQ-TREE, with same settings as earlier.

### Data processing and phylogenetic analysis of chloroplast (skimmed) data

To compare the nuclear and plastid phylogenetic signals, we recovered plastid DNA using the annotated plastome of *R. communis* (NC_016736) and transferring its annotations to a draft plastome of *Euphorbia esula* (Horvath *et al.*, 2018), if similarity was 95% between the two plastomes, using the transfer annotations function in GENEIOUS v.9.1.7 (http://www.geneious.com, Kearse *et al.*, 2012). Subsequently, we extracted coding sequence (CDS) regions from each gene (86 exons in total) and used HYBPIPER with the default parameters to extract exons sequences. We only included 10 samples of the three subspecies (ssp. *balsamifera*, *adenensis* and *sepium*). For ssp. *balsamifera*, samples were merged by island (i.e. Gran Canaria, Tenerife and La Gomera); for the other subspecies, samples were merged by country (Table S7). Admittedly, this strategy was not optimal but it was motivated by the low coverage recovered by sample (Table S8). Also, our main interest with this analysis was to test if the plastid genome supported the same clades, corresponding to the three subspecies, recovered in all the analyses of nuclear loci. We obtained 66 exon matrices that were aligned with MAFFT and corrected manually following a similarity criterion (Simmons, 2004). Then, they were concatenated into a supermatrix (63 133 bp) that included only seven samples because the sequence length in three samples did not reach 5% of the total length in the concatenated matrix. The latter was analysed with RAxML applying GTRCAT and 200 fast bootstraps followed by slow ML optimization (default '–fa' search).

### Divergence time estimation

Divergence times were estimated in BEAST v.1.8 (Drummond & Rambaut, 2007) using the nuclear exon supermatrix (296 loci). Analyses were run in a reduced dataset that included all the target species, and two samples for each of the three clades (subspecies) within *E. balsamifera* (Methods S1). The dataset was run unpartitioned under the best-fitting substitution model estimated in IQ-TREE (GTR+I). A birth–death process with incomplete taxon sampling (Stadler, 2009) was used as the tree-growth prior. We forced the monophyly of some clades to conform to accepted phylogenetic relationships among the *Athymalus* taxa (Peirson *et al.*, 2013; Methods S1). We did the same for ssp. *balsamifera, sepium* and *adenensis*, as these taxa were recovered as monophyletic groups in all phylogenomic analyses of the nuclear and chloroplast data (see later). We estimated divergence times under the strict clock (SC) model and two Bayesian relaxed clocks: the uncorrelated lognormal clock (UCLD; Drummond *et al.*, 2003) and a random local clock (RLC; Drummond & Suchard, 2010). Final analyses comprised Markov chain Monte Carlo (MCMC) searches run for 400 million generations, with samples logged every 40 000$^{th}$ generation. The root node and the next basal node – that is, the divergence from the rest of the tree of *E. antso* and *E. hadramautica*, respectively – were constrained using two secondary calibration points from Horn *et al.* (2014). They were assigned normal distribution priors (Ho & Phillips, 2009) spanning the mean and 95% highest posterior density (HPD) intervals in the original studies: 24.56 ± 5 Ma for *E. antso* and 18.21 ± 4 Ma for *E. hadramautica*. TRACER v.1.6 (Rambaut *et al.*, 2014) was used to verify MCMC stationarity and adequate effective sampling sizes (ESS > 200) for all parameters. TREEANNOTATOR v.1.8.0 (http://beast.bio.ed.ac.uk/treea

© 2018 The Authors
*New Phytologist* © 2018 New Phytologist Trust

nnotator) and FigTree v.1.4.2 (Rambaut, 2009) were used, respectively, to generate and visualize the resulting maximum clade credibility (MCC) tree.

To date population divergence events within Canarian *E. balsamifera* (the best sampled in our dataset), we used a reduced exon supermatrix containing one accession of ssp. *adenensis* and all (80) accessions of ssp. *balsamifera*. All priors were set as described earlier, except for constant coalescent as tree prior. The divergence of ssp. *balsamifera* from ssp. *adenensis* was constrained with a normal distribution prior (4.5 ± 1.05 Ma) from the species-level analysis described earlier.

The raw reads were deposited in GenBank under BioProject PRJNA415769. Draft assemblies, bait sequences and full-length targets are archived in Digital-csic (https://digital.csic.es); Beast xml files are provided in Methods S2.

## Results

### Gene-capture success

The baits designed based on genomic resources from *R. communis* and two species of *Euphorbia* belonging to subgenera other than *Athymalus* were effective in capturing the target genes across broad evolutionary scales: among sections of *Athymalus*, within section *Balsamis*, and among subspecies and populations of *balsamifera* (Table S9). The average number of reads per sample obtained was 1 036 822 and the percentage of mapped reads per sample was 48.61%, ranging from 7 to 69% (Table S10). In general, capture success was much higher for silica-dried material than for herbarium material (Table 1; Fig. 2). Average success was 73% and varied across taxa. Gene-capture success was almost complete for ssp. *balsamifera* (99%) with the exception of three herbarium samples (IS401, IS459, IS460) from southwest Morocco and Western Sahara (Table S9). Other taxa also had a relatively high gene-capture success (86%) in contrast to ssp. *adenensis* (43%) and *sepium* (45%) that were represented almost entirely by herbarium samples (Table S9). There was a significant correlation between age of herbarium material (which ranged between 3 and 126 yr) and percentage of success ($P < 0.01$; Fig. 2). However, the plot of residuals shows that this correlation does not hold for all age groups: it is not significant for specimens within the time interval 1950–2002 ($P > 0.5$).

**Table 1** Number of sampled individuals per taxon and type of material

| | Sample number | | | |
|---|---|---|---|---|
| Taxon | Silica | Herbarium | Total | Successful |
| *Euphorbia* ssp. *adenensis* | 1 | 43 | 44 | 19 (43%) |
| *Euphorbia* ssp. *balsamifera* | 81 | 4 | 85 | 84 (99%) |
| *Euphorbia* ssp. *sepium* | 0 | 22 | 22 | 10 (45%) |
| Other taxa | 11 | 3 | 14 | 12 (86%) |
| Total | 93 (96%*) | 72 (46%*) | 165 | 121 (73%) |

Successfully amplified samples included in the phylogenomic analyses. The two percentage values in the last row indicate the amplification success depending on the type of material used (fresh silica-dried tissue vs herbarium samples).

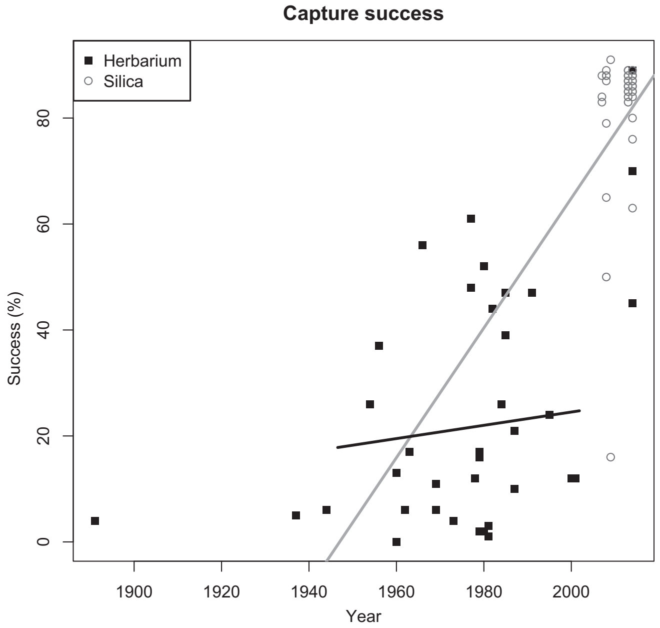

**Fig. 2** Percentage of capture success plotted against specimen age of all sampled species in *Euphorbia* subg. *Athymalus*. Black squares, herbarium samples; grey circles, silica samples. Grey line, regression line for all samples ($R^2 = 0.67$, $P < 0.001$, $F$-statistic = 255.6); black line, regression line for samples collected between 1950 and 2002 ($R^2 = 0.01$, $P = 0.66$, $F$-statistic = 0.19).

Summary statistics for the exon, intron, and supercontig matrices (Tables S11–S13) obtained using AMAS (Borowiec, 2016). The proportion of parsimony-informative sites was generally higher for introns (13%) than in supercontigs (9%) and exons (8%). The exon matrices had, on average, 23.3% of missing data (ranging from 4.9% to 80.5%) whereas the intron matrices had, on average, 59.2% (33.1–85.4%) and in the supercontigs the value was 42.9% (15.3–67.6%).

### Phylogenetic estimation of gene tree-species tree from nuclear data

All our analyses solved evolutionary relationships within the subgenus *Athymalus* as well as within *E. balsamifera*. Maximum likelihood analysis (IQ-Tree) of the 296-exon dataset recovered a well-supported phylogeny with values of bootstrap supports (BS) of 100 for backbone and main clade relationships, except for the position of *E. noxia* (BS = 94), which is included in a phylogeny here for the first time. Within *E. balsamifera*, the three subspecies were recovered in a highly supported clade. Subspecies *sepium* was sister to ssp. *adenensis* and *balsamifera* (Fig. 3a). A clade comprising *E. masirahensis*, *E. larica* and *E. noxia* was inferred as sister to *E. balsamifera*, with *E. davyi* as sister to them. Successive sister clades are two other clades: *E. socotrana*-*E. scheffleri* and *E. smithii*-*E. matabelensis*; *E. hadramautica* is sister to the aforementioned clades (Fig. 3a), with *E. antso* at the root of the tree.

At the population level, specimens of ssp. *balsamifera* were grouped into several highly supported clades (> 95 BS) that seem to follow geographic location. Specimens from Gran Canaria and eastern Tenerife were grouped into two clades that were sister to

© 2018 The Authors
*New Phytologist* © 2018 New Phytologist Trust

each other (Fig. 3a). Sequences from western Tenerife were divided into two subclades (BS = 100 and 85) embedded within a larger La Gomera clade. The three samples from southwest Morocco (IS 401) and Western Sahara (IS459, IS460) were grouped together, nested within a clade containing samples of eastern Tenerife. The clades of ssp. *adenensis* and *sepium* had substantially longer branches than the ssp. *balsamifera* clade, and in contrast to it, they lacked clear geographic structure (Fig. 3a). Similar tree topologies and relative branch lengths were recovered by the RAxML analysis (Fig. S2) and IQ-Tree, with or without trimmed low-quality sites (results not shown).

To test whether the observed differences in branch length within *E. balsamifera* were an artefact of low capture-efficiency in herbarium samples, as a result of an excess of short DNA sequences in ssp. *adenensis* and *sepium*, we repeated the ML analysis using a subset of 18 selected loci (the ones having the shortest sequences) and excluding sequences with < 25% of the target length. Each gene was aligned separately with Mafft and trimmed with trimAL to retain only sites with < 25% missing data for each gene. The exon sequences were combined in a supermatrix and analysed with RAxML. The resulting tree (Fig. S3) shows the same pattern of branch length and relationships as in the previous analyses (Figs 3a, S2). We also estimated the Astral-ii tree (Fig. S4), which shows the same topology as Fig. 3(b).

Multispecies coalescent methods recovered relationships consistent with the ML (IQ-Tree, RAxML) concatenated trees. Both Astral-ii and SVDquartets generated species trees in which a strongly supported monophyletic ssp. *sepium* was sister to ssp. *adenensis* and *balsamifera*, which were in turn reciprocally monophyletic (Figs 3b, S5, S6). Backbone relationships among *Athymalus* species were similar to those in the IQ-Tree (Fig. 3a); the only difference within sect. *Balsamis* was the position of *E. noxia*, which appeared as sister to *E. balsamifera* in Astral-ii (Fig. 3b) but formed a clade with *E. masirahensis* and *E. larica* in the IQ-Tree and SVDquartets (Figs 3a, S5, S6). Topologies obtained in Astral-ii (Fig. S7) using a diverse array of aligners (Muscle vs Mafft), mergers (Muscle vs Opal), and tree estimation algorithms (FastTree vs RAxML) all supported the same relationships among subspecies and species depicted in Fig. 3(b). Within ssp. *balsamifera*, Astral-ii trees showed similar structure to the IQ-Tree (Fig. 3b). The main differences were in the western Tenerife populations, which form a clade sister to La Gomera populations in Astral and SVDquartets (Figs 3b, S6). Similar results were found in Astral (Fig. 3b) and SVDquartets with the blind analyses (results not shown) or when samples were assigned to populations (Figs S8, S5). Another difference concerned the samples from southwest Morocco and Western Sahara. These appear as sister to the remaining ssp. *balsamifera* from the Canary Islands in the MSC trees (Figs 3b, S6), rather than nested within a Tenerife clade as in the IQ-Tree and RAxML trees (Figs 3a, S2). Also, the position of the southwest Morocco specimen (IS401) varied across the Astral-ii trees (Fig. S7e–h). The long branches subtending the northwest African samples IS401 and IS459 – which presented a < 50% capture success (Fig. 3a) – suggest the possibility of long-branch attraction, against which MSC methods are more robust (Liu *et al*., 2014). To evaluate this, we sequentially removed each of these samples and re-estimated the tree in IQ-Tree (following Bergsten, 2005). Results (Fig. S9) showed no changes in the backbone topology but the position of the problematic samples varied across trees, as expected when long-branch attraction is at play.

A total of 116 loci out of 428 (27%) were warned to contain potentially paralogue sequences (Table S4). Sensitivity analyses in Astral-ii (by sequentially removing paralogues from the final 296-loci dataset) recovered similar topologies (Fig. S10) to the full analysis (Fig. 3b), suggesting no significant bias in our results. Comparison of the individual gene tree topologies against the Astral-ii species tree (Fig. S11) reveals gene tree concordance for phylogenetic relationships between subspecies and above but gene tree discordance within subspecies of *E. balsamifera* (see also Fig. 4a). For the latter, the source of conflict in most nodes is a result of incongruence with other alternative, low-frequency bipartitions or lack of support for any bipartition, whereas conflict with an alternative dominant bipartition appears to be low (Fig. S11).

Finally, the phylogenetic tree obtained from the concatenated supercontig matrix – exons plus introns (Fig. S12) – was congruent with the topology of the concatenated exon matrix, supporting similar relationships among *Athymalus* taxa and within the *E. balsamifera* clade (BS = 100), although population structure within ssp. *balsamifera* lacked support. The phylogenetic tree obtained from the concatenated intron matrix provided little resolution (results not shown), suggesting that the level of missing data within this dataset might be responsible for the lower resolution in the supercontig vs the exon-based analyses (Figs S2, S12).

## Analysis of plastid data

We recovered a low number of plastid sequences (Table S14) for *E. balsamifera* s.l. and none for any of the other taxa. No relationship was found between type of material and capture success: that is, the average percentage of mapped reads in silica samples was 0.09%, whereas for herbarium samples it ranged between 0.05% and 0.53%. The ML trees using the exon concatenated matrix showed strong support for the monophyly of ssp. *sepium, adenensis* and *balsamifera* (Fig. 3c).

## Divergence time estimation

Divergence time estimates for major clades are shown in Table 2. MCMC runs with the SC and RLC models converged (Table S15) and gave estimates with overlapping confidence intervals (Table 2). The UCLD model provided younger age estimates (Table S15), but inspection of traces indicated poor mixing and low EES values. The mean and standard deviation parameters of the lognormal clock showed a bimodal distribution, suggesting difficulties in accommodating molecular rate variation within our mixed population/species-level large genomic dataset.

The tree topology from the SC and RLC analyses (Fig. 4a) was congruent with the one recovered by the Astral-ii MSC tree (Fig. 3b), except that ssp. *balsamifera* was not recovered as

© 2018 The Authors
*New Phytologist* © 2018 New Phytologist Trust

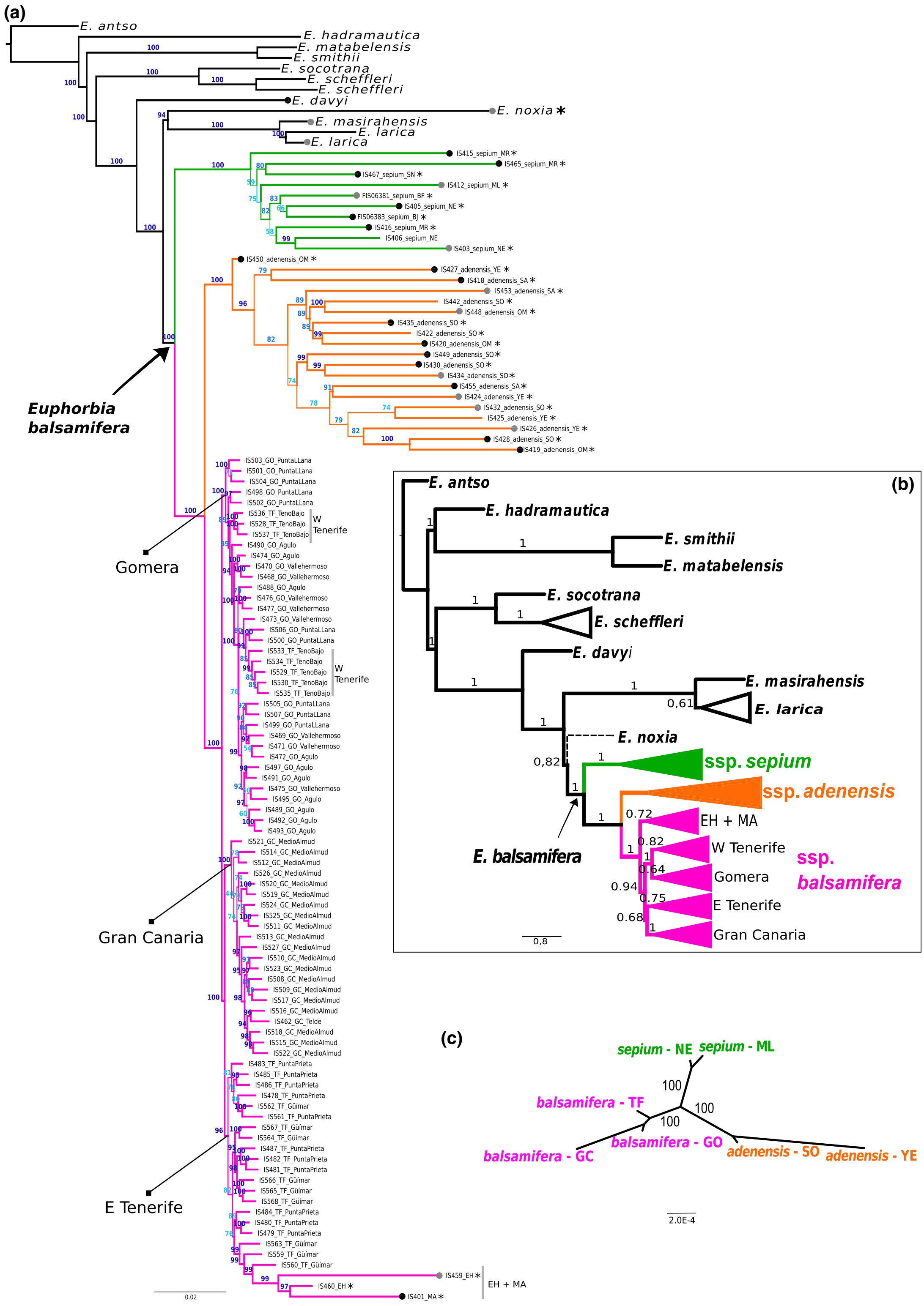

**Fig. 3** Phylogenetic relationships in *Euphorbia balsamifera* and related taxa inferred from nuclear and plastid genomic data. (a) Maximum likelihood tree (296 concatenated exon loci, 121 samples, 486 878 bp) estimated in IQ-Tree, showing phylogenetic relationships within *E. balsamifera*, section *Balsamis* and subgenus *Athymalus*. Branch width and colour indicate bootstrap support (thicker/darker for higher values). Circles at the end of branches indicate percentage of missing data: black, > 80%; grey, > 50%; < 50%, no circle; and stars indicate herbarium samples. (b) Species tree inferred with a multispecies coalescent (MSC) approach in Astral-ii using the 296-exon supermatrix; SVDquartets obtained the same tree except for the position of *E. noxia* (dashed line, see Supporting Information Fig. S5). (c) Maximum likelihood tree obtained from 66 exon loci (63 133 bp) from the chloroplast genome. BJ, Benin; BF, Burkina Faso; DJ, Djibouti; EH, Western Sahara; GC, Gran Canaria; GO, Gomera; MA, Morocco; ML, Mali; MR, Mauritania; MG, Madagascar; NE, Niger; OM, Oman; TF, Tenerife; SA, Saudi Arabia; SN, Senegal; SO, Somalia; YE, Yemen; ZA, South Africa.

© 2018 The Authors
*New Phytologist* © 2018 New Phytologist Trust

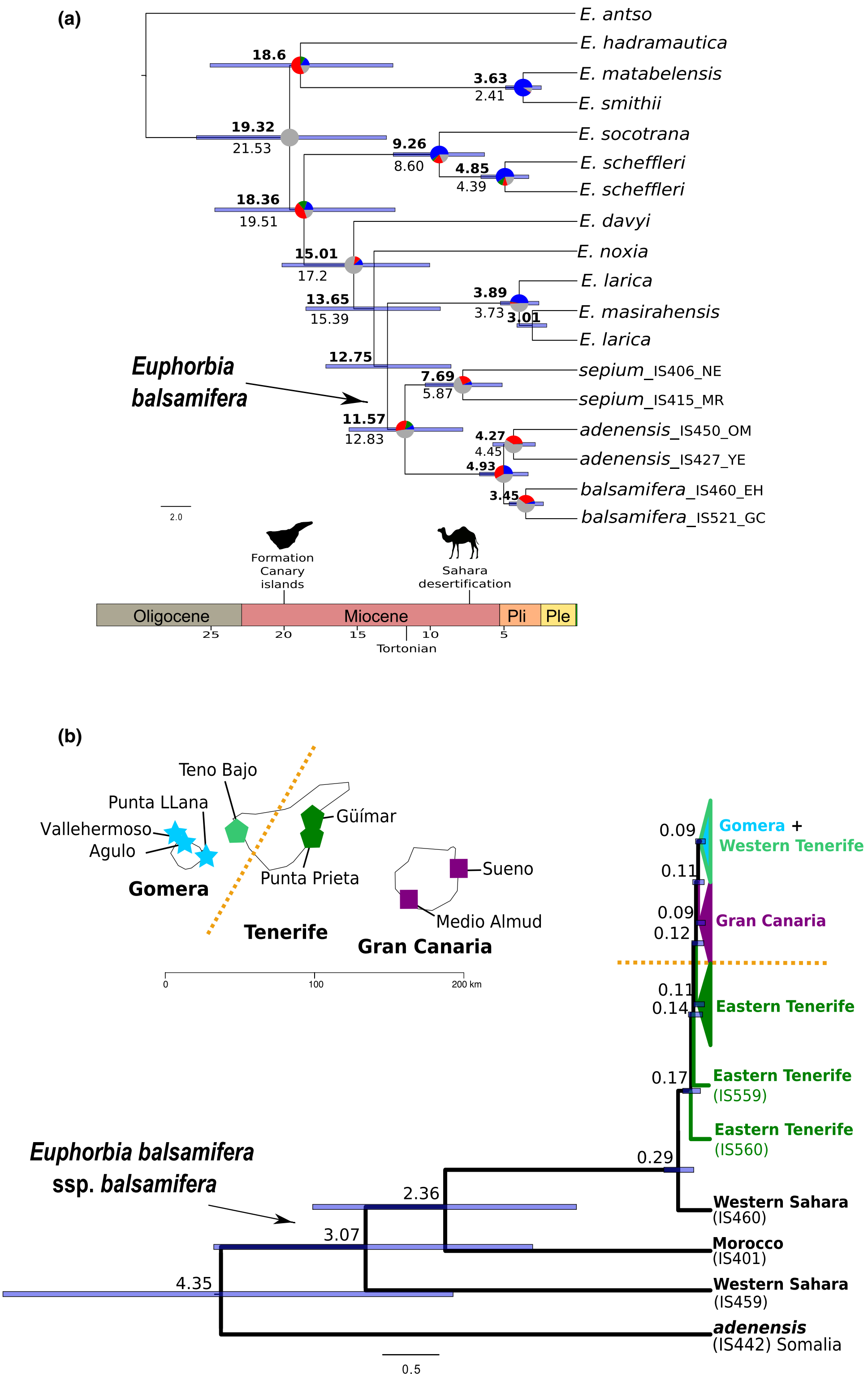

**Fig. 4** Maximum clade credibility (MCC) tree showing Bayesian estimates of divergence times (a) among species in subgenus *Athymalus*, within *Euphorbia balsamifera* (ssp. *sepium*, *adenensis*, *balsamifera*), and among populations. Values for the strict clock and random local clock analyses are given above and below branches, respectively. Pie charts show gene tree conflict at each node relative to the ASTRAL-II species tree as estimated by *phyparts* (blue, proportion concordant with the shown topology; green, proportion that support the dominant alternative topology; red, proportion that support remaining alternatives; grey: unsupported (< 50% bootstrap support, BS). (b) Expanded sampling within subspecies of *E. balsamifera* showing population-level relationships. Node bars represent the 95% highest posterior density intervals of the divergence time estimates in the strict clock analysis linked to nodes with posterior probabilities above 0.95. Mean ages inferred for clades in million yr. Map shows distribution of *E. balsamifera* sampled populations in Gran Canaria (pink squares), La Gomera (blue stars) and Tenerife (green pentagons). Yellow dashed lines indicate the split of the two major eastern/western clades. EH, Western Sahara; GC, Gran Canaria; MR, Mauritania; NE, Niger; OM, Oman; YE, Yemen.

© 2018 The Authors
*New Phytologist* © 2018 New Phytologist Trust

**Table 2** Divergence times of the main clades in *Euphorbia balsamifera* and its closest relatives in Fig. 4(a) presented as the mean time to the most recent common ancestor and the 95% highest posterior density (HPD) interval obtained from the divergence time analyses performed under the strict clock and the random local clock of the 296-exon supermatrix

| | Strict clock | | Random local clock | |
|---|---|---|---|---|
| Clades (Fig. 4a) | Mean | 95% HPD | Mean | 95% HPD |
| *Root* | 19.32 | 12.82–25.59 | 19.9 | 13.32–26.47 |
| *E. hadramautica* + *E. matabelensis* + *E. smithii* + *E. socotrana* + *E. scheffleri* + *E. davyi* + *E. noxia* + *E. larica* + *E. masirahensis* + *E. balsamifera* | 18.6 | 12.38–24.67 | 19.17 | 12.92–25.53 |
| *E. socotrana* + *E. scheffleri* | 9.26 | 6.22–12.35 | 7.64 | 5.12–10.20 |
| *E. socotrana* + *E. scheffleri* + *E. davyi* + *E. noxia* + *E. larica* + *E. masirahensis* + *E. balsamifera* | 18.36 | 12.23–24.36 | 17.18 | 11.33–22.64 |
| *E. davyi* + *E. noxia* + *E. larica* + *E. masirahensis* + *E. balsamifera* | 15.01 | 9.91–19.84 | 13.44 | 8.99–17.85 |
| *E. balsamifera* | 11.57 | 7.68–15.33 | 9.93 | 6.63–13.30 |
| Crown *E.* ssp. *sepium* | 7.69 | 5.03–10.21 | 4.55 | 3.03–6.17 |
| Crown *E.* ssp. *adenensis* | 4.27 | 2.81–5.86 | 4.16 | 2.466–5.97 |
| Crown *E.* ssp. *balsamifera* | 3.45 | 2.26–4.57 | na | na |

na, not applicable because the clade was not recovered as monophyletic.

monophyletic with RLC (Table 2). Divergence times between ssp. *sepium* and *adenensis-balsamifera* were dated in the Late Miocene (SC, 11.57 Ma, 95% HPD 7.68–15.33; RLC, 9.93 Ma; 6.63–13.30, respectively; Table 2), whereas the divergence between ssp. *balsamifera* and *adenensis* dates back to the Early Pliocene (SC, 4.93 Ma; 3.29–6.56). Divergence between *E. balsamifera* and closest relatives *masirahensis-larica-noxia-davyi* was estimated in the Mid-Miocene (15.01 Ma in SC; 13.44 RLC). Initial divergence within ssp. *sepium* was estimated as 7.69 Ma (SC) or 4.55 (RLC), while those of ssp. *adenensis* and *balsamifera* were estimated as 4.27 Ma (SC; 4.16 in RLC) and 3.45 Ma (SC), respectively (Table 2). The Beast coalescent analysis within ssp. *balsamifera* gave younger estimates for mean population divergences (Fig. 4b), but the 95% HPD credibility intervals overlap with those in Fig. 3(a). The samples from southwest Morocco/Western Sahara branched earlier (as in the Astral-ii tree; Fig. 3b), around the Early Pleistocene (2.72 (1.25–4.16); 2.08 (0.86–3.23) Ma), with a very recent split for the Canarian samples (0.26, 0.11–0.30 Ma). The population geographic structure resembled that of the IQ-Tree (Fig. 3a) except that there was no clear split into two major eastern/western clades. Instead, East Tenerife populations appeared sister to the Gran Canaria populations, while those from West Tenerife formed a grade relative to populations from La Gomera (Fig. 3b).

## Discussion

### Hyb-Seq as a tool to bridge the micro- and macroevolutionary gap in phylogenomics

The advent of HTS techniques has opened new avenues to solve phylogenetic relationships within deep and shallow evolutionary radiations (Nicholls *et al.*, 2015; Stephens *et al.*, 2015; Barrett *et al.*, 2016; Gardner *et al.*, 2016; Hamilton *et al.*, 2016; Mitchell *et al.*, 2017). Our study – the first phylogenomic analysis within Euphorbiaceae – demonstrates that our probes combined with target sequencing and genome skimming (Hyb-Seq), can be applied to solve evolutionary relationships from populations to species and above within the same tree in *Euphorbia*. Even if a proportion of an individual LCNG contained low phylogenetic signal (Table S11), our combined dataset of 296 exons alone provided enough resolution to address evolutionary relationships among species within subgenus *Athymalus* and section *Balsamis*, among subspecies within *E. balsamifera*, and even among and within populations in the latter. The phylogeny inferred from the nuclear genome dataset (Fig. 3a,b) agreed with the topologies found in previous studies with Sanger sequencing (Peirson *et al.*, 2013), but with far better support towards *E. balsamifera* and its closest relatives. The fact that the output of Hyb-Seq are DNA sequences of targeted known loci instead of SNPs as in RADs allowed us to use the same molecular evolutionary models across phylogenetic levels, from intrapopulation to populations and species, thus effectively bridging the micro- and macroevolutionary gap.

One advantage of Hyb-Seq is the recovery of plastid sequences as a by-product of nuclear enrichment (e.g. Schmickl *et al.*, 2016; Crowl *et al.*, 2017). At shallow phylogenetic levels, comparison of the nuclear and plastid signal is key to detecting any potential conflict attributed to ongoing gene flow and reticulate evolution. Here, even though < 1% of mapped reads corresponded to plastid genes, the plastid genome data recovered was enough to support the presence of three clades (subspecies) within *E. balsamifera* s.l. in agreement with the nuclear dataset (Fig. 3). Our efforts to recover introns and intergenic regions by genome skimming was met with partial success (Table S12; Fig. S12). Johnson *et al.* (2016) found a similar result: the analysis of supercontig matrices did not provide improved resolution relative to the exon-only datasets. Nonetheless, this should not be considered a failure of our approach (partly explained by our extensive use of herbarium material) as a technique to bridge across phylogenetic levels: despite the low variability of each individual loci, the combined information from all exons provided enough signal to resolve relationships at different phylogenetic depths.

Kadlec *et al.* (2017) compared target-sequencing techniques on the basis of their effectiveness for marker selection, using criteria such as gene length and variability. They concluded that approaches employing universal markers such as ultraconserved elements (Faircloth *et al.*, 2012) or anchored hybrid enrichment (Lemmon *et al.*, 2012) performed worse than those using a set of probes specific to the focal group. Among the latter, they preferred 'custom-made' scripts (Mandel *et al.*, 2015) over automated ones (MarkerMiner, Chamala *et al.*, 2015; Hyb-Seq, Weitemier *et al.*, 2014) because they tend to generate longer, more variable reads. The difference lies in how phylogenetically close to the group of interest are the genomic resources employed in the design of baits. Here, we used the proteome of *R. communis* available in MarkerMiner (Chamala *et al.*, 2015), in subfamily Acalyphoideae, against two *Euphorbia* (Euphorbioideae) transcriptomes from subgenera *Chamaesyce* and *Esula*. This probably resulted in more conserved, less variable genes (< 10%, Table S11) compared with noncoding nuclear regions such as ITS in Malpighiales (e.g. Euphorbiaceae: 45%, Horn *et al.*, 2012; Hypericaceae: 43%, Meseguer *et al.*, 2013). Indeed, inspection of gene tree discordance (Figs 4a, S11) suggests that most conflict stems from low resolution within each gene and nonoverlapping sample sets (i.e. missing data as a result of capture failure), rather than real conflict among gene trees. While we agree that the use of a fully annotated genome of the group (i.e. *Euphorbia*) is the right path to generate probes with high variability (Kadlec *et al.*, 2017), this is often out of the reach of small laboratories, both economically and time-wise. Additionally, by selecting a relatively distant proteome, we have generated a set of baits that are effective for gene capture at macro- and microevolutionary levels within *Euphorbia* (Fig. 3a) and potentially across genera within the large angiosperm family Euphorbiaceae (*c.* 6300 species; Wurdack & Farfan-Rios, 2017). This set of baits is now available for future studies in this family.

Another advantage of target sequencing such as Hyb-Seq over alternative HTS approaches is the possibility of using material from natural history collections ('museomics') which, even if degraded, is crucial when working with old, rare and endangered taxa or with species coming from under-sampled and difficult-to-access regions (e.g. Staats *et al.*, 2013; Bakker *et al.*, 2016). In our study, this is crucial because many Rand Flora taxa occur in what are currently politically unstable countries in Africa (e.g. Somalia) and the Middle East (e.g. Yemen), limiting the possibilities of doing fieldwork to obtain fresh material. Nearly all samples of ssp. *sepium* and ssp. *adenensis*, and the nonCanarian samples of ssp. *balsamifera*, were obtained from herbaria (Tables 1, S9; Figs 2, 3a). Our study demonstrates that Hyb-Seq works well with old collections. Although there is a correlation between age of specimen and capture success (Fig. 2), we show that even very old specimens (IS401, 1893) and those with very low capture success (e.g. IS435, Table S9) can be placed within their (nominal) clade in the phylogenetic trees (Figs 3a, S2, S5, S8). We also show that this correlation only holds for very old and recent samples, suggesting that for other samples, additional variables such as preservation quality might be more important in capture success.

Sayyari *et al.* (2017) showed that highly fragmentary DNA often results in a surplus of short sequences, which might lead to spurious phylogenetic reconstructions. In particular, these incomplete genes could result in long branches for the affected lineages. In our study, we used a large number of herbarium samples, which typically yield fragmented DNA below 500 bp, resulting in shorter sequences after hybridization compared with those typically obtained from fresh material. We have addressed this issue showing that even after removing these very short sequences (< 25% target gene length) the resulting phylogenetic tree recovered the same topological relationships and similar branch length differences among the three subspecies of *E. balsamifera* as in the dataset including all sequences (Fig. S4). Thus, the presence of fragmentary DNA sequences might not be necessarily misleading in a phylogenomic study. Our analyses of the influence of paralogues did not reveal a significant bias in tree estimation either (Fig. S10).

### Solving relationships within an ancient continental disjunct lineage

*Euphorbia balsamifera* represents one of the few studied intraspecific examples of the Rand Flora pattern (Pokorny *et al.*, 2015). Our results from the phylogenomic analyses strongly support the monophyly of three subspecies within *E. balsamifera*: *adenensis*, *balsamifera* and *sepium*. Subspecies *sepium*, here sequenced for the first time, was recovered as sister to the clade formed by ssp. *adenensis* and *balsamifera*, the latter being congruent with the phylogeny of Peirson *et al.* (2013). The strongly supported reciprocal monophyly by both nuclear and chloroplast genomes, and comparatively long branch lengths subtending the three subspecies (Fig. 3), suggest the need for a taxonomic revision of their taxonomic status (R. Riina *et al.*, unpublished).

The inferred divergence times between subspecies of *E. balsamifera* and with species in sect. *Balsamis* are older (ranging from Late Miocene to Early Pliocene) than those obtained in previous studies (Bruyns *et al.*, 2011; Horn *et al.*, 2014; Pokorny *et al.*, 2015), although none of them included ssp. *sepium*. Aridification of North Africa linked to the formation of the Sahara Desert (Senut *et al.*, 2009) was probably responsible for the isolation of ssp. *balsamifera* and *adenensis* during the Early to Mid-Pliocene (4.93; 3.29–6.56 Ma). The much older split of *sepium* in the Late Miocene (*c.* 11 Ma), however, predates the formation of the Sahara. This taxon has a wider distribution in North Africa and appears to be adapted to more inland xeric environments compared with ssp. *adenensis* and *balsamifera*. The Tortonian (11.6–7.2 Ma) was a period characterized by lower temperatures and wetter environments than the Pliocene, so it is possible that ecological vicariance contributed to the divergence of ssp. *sepium*

© 2018 The Authors
*New Phytologist* © 2018 New Phytologist Trust

from the ancestor of *adenensis* and *balsamifera,* as has been postulated in other Rand Flora taxa (Mairal *et al.*, 2017).

Our estimates of lineage divergence times support a recent colonization of the Canary Islands by ssp. *balsamifera* (crown age *c.* 0.26 Ma), probably from coastal Moroccan/Western Saharan populations (Fig. 4). This agrees well with the hypothesis of Mairal *et al.* (2015a) that the Macaronesian component of the Rand Flora originated from a recent dispersal event from a northwestern African population during the Pleistocene glacial cycles when geographic distances became shorter, although other taxa show exceptions to this pattern (Thiv *et al.*, 2010). The recent Canarian radiation stands in contrast with the longer branch lengths and older time estimates for population divergence within ssp. *sepium* and *adenensis* (Figs 3a, 4a). Our phylogenetic tests showed that the branch length variation observed between African and Canarian populations of ssp. *balsamifera* is not an artefact of including incomplete genes with short sequences (Fig. S4). Moreover, this pattern is not unique to *E. balsamifera.* Rand Flora genera such as *Canarina* (Campanulaceae, Mairal *et al.*, 2015a), *Camptoloma* (Scrophulariaceae) and *Plocama* (Rubiaceae; Sanmartín *et al.*, 2017) exhibit a similar pattern of shallower population divergences in the Macaronesian taxa compared with those in eastern Africa/southern Arabia.

We also found phylogeographic structure within Canarian *balsamifera.* Despite the limited amount of sequence divergence (Fig. 3a), an east/west population structure was found similar to the one described in the Rand Flora *Canarina canariensis* (Mairal *et al.*, 2015b) and in other plant lineages (*Periploca laevigata*, García-Verdugo *et al.*, 2017; *Scrophularia arguta*, Valtueña *et al.*, 2016).

## Conclusions

Our study demonstrates that the baits designed for *Euphorbia* subgenus *Athymalus* using genomic and transcriptomic resources from distantly related taxa (i.e. subfamily Acalyphoideae and *Euphorbia* subgenera *Chamaesyce* and *Esula*) are able to solve phylogenetic relationships at shallow and deep levels. We show that inclusion of fragmentary gene sequences in a dataset, a typical effect of working with degraded DNA from natural history collections, does not necessarily bias phylogenomic analyses at inter- and intraspecific levels. This might be crucial for systematic studies involving large-scale spatial sampling or fieldwork in difficult-to-access regions. Finally, we have revealed that *E. balsamifera* consists of three highly supported clades (ssp. *adenensis*, *balsamifera* and *sepium*) and that estimates of divergence times and phylogeographic structure within ssp. *balsamifera* in the Canary Islands are consistent with those found in other Rand Flora taxa.

## Acknowledgements

We thank the 1KP Initiative (Gane Ka-Shu Wong, transcriptomes), David Horvath (plastome, USDA-ARS), Alison Devault (bait design, Arbor Biosciences), Kevin Feldheim (The Field Museum), Alberto Herrero, Mario Mairal, and herbaria BCN, E, FT, K, and LPA for samples. Computing resources: CSIC cluster TRUENO, CIPRES Science Gateway and CSC-Finnish IT Center for Science. Funding grants: CGL2012-40129-C02-01 and CGL2015-67849-P (MINECO/FEDER) to I.S, CGL2015-73621-JIN (AEI/FEDER) to R.R., and EU-SYNTHESYS to L.P. (GB-TAF-5153). We thank several anonymous reviewers for insightful comments.

## Author contributions

L.P., R.R. and I.S. conceived the study; L.P., I.S., N.J.W. and M.G.J. designed the set of probes; L.P. and E.M.G. carried out the molecular work; T.V. analysed the data with help from E.M.G., S.O., L.P., M.G.J., R.R. and I.S.; J.M. and R.R. contributed samples and taxonomic knowledge and M.R-B. processed the distribution data. T.V., R.R. and I.S. wrote the manuscript with contributions from all authors. T.V., R.R. and I.S. revised the manuscript.

## ORCID

Tamara Villaverde http://orcid.org/0000-0002-9236-8616
Sanna Olsson http://orcid.org/0000-0002-1199-4499
Mario Rincón-Barrado http://orcid.org/0000-0001-5571-1473
Matthew G. Johnson http://orcid.org/0000-0002-1958-6334
Elliot M. Gardner http://orcid.org/0000-0003-1133-5167
Ricarda Riina http://orcid.org/0000-0002-7423-899X
Isabel Sanmartín http://orcid.org/0000-0001-6104-9658

## References

**Andrus N, Trusty J, Santos-Guerra A, Jansen RK, Francisco-Ortega J. 2004.** Using molecular phylogenies to test phytogeographical links between East/South Africa-Southern Arabia and the Macaronesian islands – a review, and the case of *Vierea* and *Pulicaria* section *Vieraeopsis* (Asteraceae). *Taxon* **53**: 333–346.

**Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008.** Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* **3**: 1–7.

**Bakker FT, Lei D, Yu J, Mohammadin S, Wei Z, van de Kerke S, Gravendeel B, Nieuwenhuis M, Staats M, Alquezar-Planas DE *et al.* 2016.** Herbarium genomics: plastome sequence assembly from a range of herbarium specimens using an Iterative Organelle Genome Assembly pipeline. *Biological Journal of the Linnean Society* **117**: 33–43.

**Barrett CF, Bacon CD, Antonelli A, Cano Á, Hofmann T. 2016.** An introduction to plant phylogenomics with a focus on palms. *Botanical Journal of the Linnean Society* **182**: 234–255.

**Benton M. 1995.** Diversification and extinction in the history of life. *Science* **268**: 52–58.

**Bergsten J. 2005.** A review of long-branch attraction. *Cladistics* **21**: 163–193.

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

**Borowiec ML. 2016.** AMAS: a fast tool for alignment manipulation and computing of summary statistics. *PeerJ* **4**: e1660.

**Bruyns PV, Klak C, Hanáček P. 2011.** Age and diversity in Old World succulent species of *Euphorbia* (Euphorbiaceae). *Taxon* **60**: 1717–1733.

**Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009.** trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972–1973.

© 2018 The Authors
*New Phytologist* © 2018 New Phytologist Trust

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

**Chan AP, Crabtree J, Zhao Q, Lorenzi H, Orvis J, Puiu D, Melake-Berhan A, Jones KM, Redman J, Chen G *et al.* 2010.** Draft genome sequence of the oilseed species *Ricinus communis. Nature Biotechnology* **28**: 951–956.

**Chifman J, Kubatko L. 2014.** Quartet inference from SNP data under the coalescent model. *Bioinformatics* **30**: 3317–3324.

**Chifman J, Kubatko L. 2015.** Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites. *Journal of Theoretical Biology* **374**: 35–47.

**Chou J, Gupta A, Yaduvanshi S, Davidson R, Nute M, Mirarab S, Warnow T. 2015.** A comparative study of SVDquartets and other coalescent-based species tree estimation methods. *BMC Genomics* **16**: S2.

**Christ H. 1892.** Exposé sur le rôle que joue dans le domaine de nos flores la flore dite ancienne africaine. *Archives des Sciences Physiques et Naturelles Genève* **3**: 369–374.

**Crowl AA, Myers C, Cellinese N. 2017.** Embracing discordance: phylogenomic analyses provide evidence for allopolyploidy leading to cryptic diversity in a Mediterranean *Campanula* (Campanulaceae) clade. *Evolution* **71**: 913–922.

**Davis MB, Shaw RG. 2001.** Range shifts and adaptive responses to Quaternary climate change. *Science* **292**: 673–679.

**Dentinger BTM, Gaya E, O'Brien H, Suz LM, Lachlan R, Díaz-Valderrama JR, Koch RA, Aime MC. 2016.** Tales from the crypt: genome mining from fungarium specimens improves resolution of the mushroom tree of life. *Biological Journal of the Linnean Society* **117**: 11–32.

**Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG. 2003.** Measurably evolving populations. *Trends in Ecology & Evolution* **18**: 481–488.

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

**Drummond AJ, Suchard MA. 2010.** Bayesian random local clocks, or one rate to rule them all. *BMC Biology* **8**: 114.

**Edgar RC. 2004.** MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**: 1792–1797.

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

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

**Fér T, Schmickl RE. 2018.** HybPhyloMaker: target enrichment data analysis from raw reads to species trees. *Evolutionary Bioinformatics* **14**: 1176934317742613.

**García-Verdugo C, Mairal M, Monroy P, Sajeva M, Caujapé-Castells J. 2017.** The loss of dispersal on islands hypothesis revisited: implementing phylogeography to investigate evolution of dispersal traits in *Periploca* (Apocynaceae). *Journal of Biogeography* **44**: 2595–2606.

**Gardner EM, Johnson MG, Ragone D, Wickett NJ, Zerega NJC. 2016.** Low-coverage, whole-genome sequencing of *Artocarpus camansi* (Moraceae) for phylogenetic marker development and gene discovery. *Applications in Plant Sciences* **4**: 1600017.

**Govaerts R, Frodin DG, Radcliffe-Smith A. 2000.** *World checklist and bibliography of Euphorbiaceae (and Pandaceae).* Royal Botanic Gardens, Kew, UK: Kew Publishing.

**Guichoux E, Lagache L, Wagner S, Chaumeil P, Léger P, Lepais O, Lepoittevin C, Malausa T, Revardel E, Salin F *et al.* 2011.** Current trends in microsatellite genotyping. *Molecular Ecology Resources* **11**: 591–611.

**Hamilton CA, Lemmon AR, Lemmon EM, Bond JE. 2016.** Expanding anchored hybrid enrichment to resolve both deep and shallow relationships within the spider tree of life. *BMC Evolutionary Biology* **16**: 122.

**Hart ML, Forrest LL, Nicholls JA, Kidner CA. 2016.** Retrieval of hundreds of nuclear loci from herbarium specimens. *Taxon* **65**: 1081–1092.

**Hewitt G. 2000.** The genetic legacy of the Quaternary ice ages. *Nature* **405**: 907–913.

**Hewitt G. 2004.** Genetic consequences of climatic oscillations in the Quaternary. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **359**: 183–195.

**Hipp AL, Manos PS, González-Rodríguez A, Hahn M, Kaproth M, McVay JD, Avalos SV, Cavender-Bares J. 2018.** Sympatric parallel diversification of major oak clades in the Americas and the origins of Mexican species diversity. *New Phytologist* **217**: 439–452.

**Ho SY, Phillips MJ. 2009.** Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. *Systematic Biology* **58**: 367–380.

**Hobolth A, Uyenoyama MK, Wiuf C. 2008.** Importance sampling for the infinite sites model. *Statistical Applications in Genetics and Molecular Biology* **7**: Article 32.

**Horn JW, van Ee BW, Morawetz JJ, Riina R, Steinmann VW, Berry PE, Wurdack KJ. 2012.** Phylogenetics and the evolution of major structural characters in the giant genus *Euphorbia* L. (Euphorbiaceae). *Molecular Phylogenetics and Evolution* **63**: 305–326.

**Horn JW, Xi Z, Riina R, Peirson JA, Yang Y, Dorsey BL, Berry PE, Davis CC, Wurdack KJ. 2014.** Evolutionary bursts in *Euphorbia* (Euphorbiaceae) are linked with photosynthetic pathway. *Evolution* **68**: 3485–3504.

**Horvath DP, Patel S, Doğramaci M, Chao WS, Anderson JV, Foley ME, Scheffler GL, Dorn K, Yan C, Childers A *et al.* 2018.** Gene space and transcriptome assemblies of Leafy Spurge (*Euphorbia esula*) identify promoter sequences, repetitive elements, high-quality markers, and a full-length chloroplast genome. *Weed Science* **66**: 355–367.

**Johnson MG, Gardner EM, Liu Y, Medina R, Goffinet B, Shaw AJ, Zerega NJC, Wickett NJ. 2016.** HybPiper: extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in Plant Sciences* **4**: 1600016.

**Kadlec M, Bellstedt DU, Le Maitre NC, Pirie MD. 2017.** Targeted NGS for species level phylogenomics: 'made to measure' or 'one size fits all'? *PeerJ* **5**: e3569.

**Kalyaanamoorthy S, Minh BQ, Wong TK, von Haeseler A, Jermiin LS. 2017.** ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature methods* **14**: 587.

**Katoh K, Standley DM. 2013.** MAFFT: multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**: 772–780.

**Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C *et al.* 2012.** Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**: 1647–1649.

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

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

**Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009.** The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

**Liu L, Xi Z, Davis CC. 2014.** Coalescent methods are robust to the simultaneous effects of long branches and incomplete lineage sorting. *Molecular Biology and Evolution* **32**: 791–805.

**Luo R, Hipp AL, Larget B. 2007.** A Bayesian model of AFLP marker evolution and phylogenetic inference. *Statistical Applications in Genetics and Molecular Biology* **6**: Article 11.

**Mairal M, Pokorny L, Aldasoro JJ, Alarcón M, Sanmartín I. 2015a.** Ancient vicariance and climate-driven extinction explain continental-wide disjunctions in Africa: the case of the Rand Flora genus *Canarina* (Campanulaceae). *Molecular Ecology* **24**: 1335–1354.

**Mairal M, Sanmartín I, Aldasoro JJ, Culshaw V, Manolopoulou I, Alarcón M. 2015b.** Palaeo-islands as refugia and sources of genetic diversity within volcanic archipelagos: the case of the widespread endemic *Canarina canariensis* (Campanulaceae). *Molecular Ecology* **24**: 3944–3963.

**Mairal M, Sanmartín I, Pellissier L. 2017.** Lineage-specific climatic niche drives the tempo of vicariance in the Rand Flora. *Journal of Biogeography* **44**: 911–923.

© 2018 The Authors
*New Phytologist* © 2018 New Phytologist Trust

**Mandel JR, Dikow RB, Funk VA. 2015.** Using phylogenomics to resolve mega-families: an example from Compositae. *Journal of Systematics and Evolution* **53**: 391–402.

**Meseguer AS, Aldasoro JJ, Sanmartín I. 2013.** Bayesian inference of phylogeny, morphology and range evolution reveals a complex evolutionary history in St. John's wort (*Hypericum*). *Molecular phylogenetics and evolution* **67**: 379–403.

**Mirarab S, Nguyen N, Guo S, Wang LS, Kim J, Warnow T. 2015.** PASTA: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. *Journal of Computational Biology* **22**: 377–386.

**Mirarab S, Warnow T. 2015.** ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* **31**: i44–i52.

**Mitchell N, Lewis PO, Lemmon EM, Lemmon AR, Holsinger KE. 2017.** Anchored phylogenomics improves the resolution of evolutionary relationships in the rapid radiation of *Protea* L. *American Journal of Botany* **104**: 102–115.

**Molero J, Garnatje T, Rovira A, Garcia-Jacas N, Susanna A. 2002.** Karyological evolution and molecular phylogeny in Macaronesian dendroid spurges (*Euphorbia* subsect. *Pachycladae*). *Plant Systematics and Evolution* **231**: 109–132.

**Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015.** IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* **32**: 268–274.

**Nicholls JA, Pennington RT, Koenen EJM, Hughes CE, Hearn J, Bunnefeld L, Dexter KG, Stone GN, Kidner CA. 2015.** Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the neotropical rain forest genus *Inga* (Leguminosae: Mimosoideae). *Frontiers in Plant Science* **6**: 710.

**Oliver JC. 2013.** Microevolutionary processes generate phylogenomic discordance at ancient divergences. *Evolution* **67**: 1823–1830.

**Parmesan C. 2006.** Ecological and evolutionary responses to recent climate change. *Annual Review of Ecology, Evolution and Systematics* **37**: 637–669.

**Peirson JA, Bruyns PV, Riina R, Morawetz JJ, Berry PE. 2013.** A molecular phylogeny and classification of the largely succulent and mainly African *Euphorbia* subg. *Athymalus* (Euphorbiaceae). *Taxon* **62**: 1178–1199.

**Pelser PB, Nordenstam B, Kadereit JW, Watson LE. 2007.** An ITS phylogeny of tribe Senecioneae (Asteraceae) and a new delimitation of *Senecio* L. *Taxon* **56**: 1077.

**Pokorny L, Riina R, Mairal M, Meseguer AS, Culshaw V, Cendoya J, Serrano M, Carbajal R, Ortiz S, Heuertz M *et al.* 2015.** Living on the edge: timing of Rand Flora disjunctions congruent with ongoing aridification in Africa. *Frontiers in Genetics* **6**: 154.

**Price MN, Dehal PS, Arkin AP. 2010.** FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**: e9490.

**Rambaut A. 2009.** *FigTree version 1.4. 2.* [WWW document] URL http://tree.bio.ed.ac.uk [accessed 5 April 2017].

**Rambaut A, Suchard MA, Xie W, Drummond AJ. 2014.** *Tracer. MCMC Trace analysis tool version v1.6.* [WWW document] URL http://tree.bio.ed.ac.uk/software/tracer/ [accessed 1 April 2017].

**Reznick DN, Ricklefs RE. 2009.** Darwin's bridge between microevolution and macroevolution. *Nature* **457**: 837–842.

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

**Sanmartín I, Anderson CL, Alarcon M, Ronquist F, Aldasoro JJ. 2010.** Bayesian island biogeography in a continental setting: the Rand Flora case. *Biology Letters* **6**: 703–707.

**Sanmartín I, Villaverde T, Pokorny L, Mairal M, Olsson S, Riina R. 2017. Exploring the evolutionary and ecological drivers behind intracontinental geographic disjunctions: the African Rand Flora. In:** *XIX International Botanical Congress (IBC 2017). 23–29 July 2017, Shenzen (China). Abstract Book I.* Oral Presentations, page 78: abstract T1-27-04.

**Sayyari E, Whitfield JB, Mirarab S. 2017.** Fragmentary gene sequences negatively impact gene tree and species tree reconstruction. *Molecular Biology and Evolution* **34**: 3279–3291.

**Schmickl R, Liston A, Zeisek V, Oberlander K, Weitemier K, Straub SCK, Cronn RC, Dreyer LL, Suda J. 2016.** Phylogenetic marker development for target enrichment from transcriptome and genome skim data: the pipeline and its application in southern African *Oxalis* (Oxalidaceae). *Molecular Ecology Resources* **16**: 1124–1135.

**Senut B, Pickford M, Ségalen L. 2009.** Neogene desertification of Africa. *Comptes Rendus Geoscience* **341**: 591–602.

**Shen X-X, Hittinger CT, Rokas A. 2017.** Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecology & Evolution* **1**: 126.

**Simmons MP. 2004.** Independence of alignment and tree search. *Molecular Phylogenetics and Evolution* **31**: 874–879.

**Smith SA, Moore MJ, Brown JW, Yang Y. 2015.** Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evolutionary Biology* **15**: 150.

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

**Stadler T. 2009.** On incomplete sampling under birth–death models and connections to the sampling-based coalescent. *Journal of Theoretical Biology* **261**: 58–66.

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

**Stephens JD, Rogers WL, Heyduk K, Cruse-Sanders JM, Determann RO, Glenn TC, Malmberg RL. 2015.** Resolving phylogenetic relationships of the recently radiated carnivorous plant genus *Sarracenia* using target enrichment. *Molecular Phylogenetics and Evolution* **85**: 76–87.

**Svenning J-C, Eiserhardt WL, Normand S, Ordonez A, Sandel B. 2015.** The Influence of paleoclimate on present-day patterns in biodiversity and ecosystems. *Annual Review of Ecology, Evolution, and Systematics* **46**: 551–572.

**Swofford DL. 2002.** *PAUP*: phylogenetic analysis using parsimony (* and other methods). v.4.0a147.* Sunderland, MA, USA: Sinauer.

**Tan G, Muffato M, Ledergerber C, Herrero J, Goldman N, Gil M, Dessimoz C. 2015.** Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference. *Systematic Biology* **64**: 778–791.

**Tavaré S. 1986.** Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences* **17**: 57–86.

**Thiv M, Thulin M, Hjertson M, Kropf M, Linder HP. 2010.** Evidence for a vicariant origin of Macaronesian–Eritreo/Arabian disjunctions in *Campylanthus* Roth (Plantaginaceae). *Molecular Phylogenetics and Evolution* **54**: 607–616.

**Valtueña FJ, López J, Álvarez J, Rodríguez-Riaño T, Ortega-Olivencia A. 2016.** *Scrophularia arguta*, a widespread annual plant in the Canary Islands: a single recent colonization event or a more complex phylogeographic pattern? *Ecology and Evolution* **6**: 4258–4273.

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

**Wheeler TJ, Kececioglu JD. 2007.** Multiple alignment by aligning alignments. *Bioinformatics* **23**: i559–i568.

**Wu C-H, Drummond AJ. 2011.** Joint inference of microsatellite mutation models, population history and genealogies using transdimensional Markov Chain Monte Carlo. *Genetics* **188**: 151–164.

**Wurdack KJ, Farfan-Rios W. 2017.** *Incadendron*: a new genus of Euphorbiaceae tribe Hippomaneae from the sub-Andean cordilleras of Ecuador and Peru. *PhytoKeys* **85**: 69–86.

## Supporting Information

Additional Supporting Information may be found online in the Supporting Information at the end of the article.

**Fig. S1** RAxML consensus tree based on ITS sequences of *Euphorbia balsamifera* and related taxa in subgenus *Athymalus*.

**Fig. S2** RAxML consensus tree using the same dataset as in Fig. 3 (a) (296 exon matrices concatenated, 121 samples, 486 878 bp).

© 2018 The Authors
*New Phytologist* © 2018 New Phytologist Trust

**Fig. S3** RAxML consensus tree of 18 selected exons from the ones used in previous analyses, with sequences < 25% target gene length removed.

**Fig. S4** ASTRAL-II tree obtained from individual RAxML consensus tree of the 18 shortest exons.

**Fig. S5** SVDQUARTETS analysis (gene tree) using the same dataset as in Fig. 3(a) (296 exon matrices concatenated, 121 samples, 486 878 bp).

**Fig. S6** SVDQUARTETS analysis (species tree) using the same dataset as in Fig. 3(a) (296 exon matrices concatenated, 121 samples, 486 878 bp).

**Fig. S7** Cartoon phylogenies of *Euphorbia balsamifera* and relatives obtained in RAxML with different aligners, mergers, and tree estimation algorithms.

**Fig. S8** ASTRAL-II species tree with samples assigned to the lineages identified in Fig. 3(b).

**Fig. S9** Cartoon phylogenies from different analyses in IQ-TREE, removing NW African samples of ssp. *balsamifera* (southwest Morocco and Western Sahara).

**Fig. S10** Sensitivity analysis of the effect of paralogs in ASTRAL-II tree estimation.

**Fig. S11** ASTRAL-II species tree with pie charts showing gene tree conflict at each node as estimated by *phyparts* (https://bitbucket.org/blackrim/phyparts).

**Fig. S12** RAxML phylogenetic tree obtained from the nuclear concatenated supercontig matrix (exons and introns).

**Table S1** Taxon sampling, DNA accession number and voucher information

**Table S2** List of 431 LCNGs targeted genes used in this study

**Table S3** Information about the 428 exons from the 431 targeted genes obtained in this study

**Table S4** Number of sequences found by HYBPIPER in each loci by sample

**Table S5** Information about the nuclear introns used in the analyses (Fig. S8)

**Table S6** Information about the nuclear supercontigs used in the analyses (Fig. S8)

**Table S7** Sets of merged samples of *Euphorbia balsamifera* (ssp. *adenensis*, *balsamifera*, *sepium*) in the analysis of the chloroplast data

**Table S8** Total and average number of mapped reads per sample used in the chloroplast phylogenetic analysis

**Table S9** Capture success for each exon and sample

**Table S10** Total and average number of mapped reads per sample used in the nuclear exon phylogenetic analysis

**Table S11** Summary statistics of loci used in the nuclear exon phylogenetic analysis

**Table S12** Summary statistics of loci used in the nuclear intron phylogenetic analysis

**Table S13** Summary statistics of loci used in the nuclear supercontig phylogenetic analysis

**Table S14** Information about the exons used in the chloroplast analyses

**Table S15** Sensitivity analysis exploring the effects of different molecular clock models in BEAST: strict clock, random local clock and uncorrelated lognormal clock (UCLD)

**Methods S1** Extended materials and methods.

**Methods S2** XML files used for the divergence time estimation analyses under a Bayesian framework in BEAST.

**Notes S1** Notes about the taxonomy of *Euphorbia balsamifera*.

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.

© 2018 The Authors
*New Phytologist* © 2018 New Phytologist Trust