MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2016) **16**, 1124–1135 doi: 10.1111/1755-0998.12487

SPECIAL ISSUE: SEQUENCE CAPTURE

Phylogenetic marker development for target enrichment from transcriptome and genome skim data: the pipeline and its application in southern African *Oxalis* (Oxalidaceae)

ROSWITHA SCHMICKL,* AARON LISTON,† VOJTĚCH ZEISEK,*‡ KENNETH OBERLANDER,*§ KEVIN WEITEMIER,† SHANNON C. K. STRAUB,¶ RICHARD C. CRONN,** LÉANNE L. DREYER†† and JAN SUDA*‡

*Institute of Botany, The Czech Academy of Sciences, Zámek 1, 252 43 Průhonice, Czech Republic, †Department of Botany and Plant Pathology, Oregon State University, 2082 Cordley Hall, Corvallis, OR 97331, USA, ‡Department of Botany, Faculty of Science, Charles University in Prague, Benátská 2, 128 01 Prague, Czech Republic, §Department of Conservation Ecology and Entomology, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa, ¶Department of Biology, Hobart and William Smith Colleges, 213 Eaton Hall, Geneva, NY 14456, USA, **USDA Forest Service, Pacific Northwest Research Station, 3200 SW Jefferson Way, Corvallis, OR 97331, USA, ††Department of Botany and Zoology, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa*

Abstract

Phylogenetics benefits from using a large number of putatively independent nuclear loci and their combination with other sources of information, such as the plastid and mitochondrial genomes. To facilitate the selection of orthologous low-copy nuclear (LCN) loci for phylogenetics in nonmodel organisms, we created an automated and interactive script to select hundreds of LCN loci by a comparison between transcriptome and genome skim data. We used our script to obtain LCN genes for southern African *Oxalis* (Oxalidaceae), a speciose plant lineage in the Greater Cape Floristic Region. This resulted in 1164 LCN genes greater than 600 bp. Using target enrichment combined with genome skimming (Hyb-Seq), we obtained on average 1141 LCN loci, nearly the whole plastid genome and the nrDNA cistron from 23 southern African *Oxalis* species. Despite a wide range of gene trees, the phylogeny based on the LCN genes was very robust, as retrieved through various gene and species tree reconstruction methods as well as concatenation. Cytonuclear discordance was strong. This indicates that organellar phylogenies alone are unlikely to represent the species tree and stresses the utility of Hyb-Seq in phylogenetics.

Keywords: cytonuclear discordance, genome skimming, low-copy nuclear genes, *Oxalis*, species tree, target enrichment

Received 24 July 2015; revision received 6 October 2015; accepted 5 November 2015

Introduction

High-throughput sequencing (HTS) has the potential to greatly increase the amount of phylogenetically informative signal in molecular data sets (Parks *et al.* 2009, 2012) and overcome difficulties in phylogenetic reconstructions, such as polytomies and low support values, which are often the result of using only a small fraction of the genome. However, HTS also 'opens the era of real incongruence' (Jeffroy *et al.* 2006), and even massive amounts of sequence data do not always result in strongly resolved phylogenies (Pyron 2015). When HTS was introduced to plant phylogenetics, sequencing of the plastid genome was its first focus (e.g. Parks *et al.* 2009; Givnish *et al.* 2010). Later approaches of genome skimming, the sequencing of the high-copy fractions of the nuclear, plastid and mitochondrial genome (Straub *et al.* 2012), resulted in the assembly of the rDNA cistron and nearly the complete plastid and mitochondrial genome. Currently, target enrichment (sequence capture) of hundreds of loci is becoming increasingly popular in phylogenetics. In animal phylogenomics, nonexonic or partly exonic ultraconserved elements and their quite variable flanking regions are often utilized (e.g. Faircloth *et al.* 2012; Hedtke *et al.* 2013; Smith *et al.* 2013). For plant phylogenetics, low-copy nuclear (LCN) genes are targeted (Mandel *et al.* 2014, 2015; Weitemier *et al.* 2014; Grover *et al.* 2015; Heyduk *et al.* 2015; Nicholls *et al.* 2015; Stephens *et al.* 2015a,b) due to the paucity of ultraconserved nuclear sequences (Reneker *et al.* 2012).

Correspondence: Roswitha Schmickl, Fax: +420-221-951-645; E-mail: roswitha.schmickl@ibot.cas.cz

Target sequencing strategies for plant nuclear genomes are largely lineage-specific, requiring the *de novo* design of target enrichment probes. Chamala *et al.* (2015) recently introduced a pipeline for phylogenetic marker development in angiosperms using transcriptomes, and they obtained several hundred putative LCN genes that can be utilized at three phylogenetic levels (genus, family, order); however, empirical evidence for the phylogenetic utility of these loci was not demonstrated. Alternative phylogenetic marker developments, also utilizing transcriptomes (Rothfels *et al.* 2013; Pillon *et al.* 2014; Tonnabel *et al.* 2014), resulted in a much smaller number (up to 20) of mainly LCN loci, but these loci were evaluated with PCR in the empirical data sets, not target enrichment. In recently published phylogenies based on target enrichment of several hundred LCN genes, these loci were selected from transcriptomes, gene expression studies, the literature, or a combination of these sources (Mandel *et al*. 2014; Grover *et al*. 2015; Heyduk *et al.* 2015; Stephens *et al.* 2015a; b, Mandel *et al.* 2015; Nicholls *et al*. 2015). Weitemier *et al.* (2014) designed LCN probes for target enrichment based on a combination of transcriptome and genome data and demonstrated their phylogenetic utility in *Asclepias* L. The limitation of this probe design pipeline is that (draft) genomes are still infrequent, especially for nonmodel species, and are costly to generate. This limitation also applies to the approach of De Sousa *et al.* (2014), who selected 50 LCN loci from a genomic source and amplified them using target enrichment. Except for Chamala *et al.* (2015), who offer a user-friendly but empirically untested probe design pipeline, and Weitemier *et al.* (2014), whose Hyb-Seq pipeline is designed for more advanced users, no automated probe design pipeline for LCN genes is currently available.

In this study, we developed a novel probe design pipeline for targeting orthologous LCN loci for phylogenetic reconstruction by using genome skim and transcriptome data. In particular, genome skim data of one accession of the studied plant group were combined with a congeneric transcriptome from the 1000 Plants (1KP) initiative (http://www.onekp.com/). We implemented our software workflow in the user-friendly, automated and interactive BASH script Sondovač, which allows a straightforward design of LCN probes also for users with limited bioinformatics skills. The utility of this approach is demonstrated by the design of probes for southern African *Oxalis*, and over 1000 candidate LCN loci were obtained. Use of the probes for targeted sequencing of these loci in 23 southern African *Oxalis* species resulted in sufficient sequencing depth of the LCN loci, as well as the plastid and high-copy nuclear genome (e.g. nuclear ribosomal DNA (nrDNA) cistron). Considering their different evolutionary rates and modes of inheritance (Small *et al.* 2004), the combination of all three data sets can substantially contribute to understanding speciation from a phylogenetic perspective. The observed, strong cytonuclear discordance (nuclear tree topology deviating from organellar tree topology) suggests that organellar phylogenies alone do not resemble the species tree.

Materials and methods

Taxonomic focus

Oxalis L. (*c.* 500 species) is common in the flora of the New World and a major component of the Greater Cape Floristic Region (GCFR) (Born *et al.* 2006). Southern African taxa comprise *c.* 46% of the genus (*c.* 230 species) and represent the seventh-largest genus and the largest geophytic genus in the GCFR (Proches *et al.* 2006); they bear bulbs with above-ground parts emanating from seasonal stems. There is some evidence of rapid, possibly adaptive radiation of southern African *Oxalis*, as the base of the clade is poorly resolved (Oberlander *et al.* 2011). Results of dating the southern African crown *Oxalis* radiation varied between an age of 9.9 and 32.2 Myr (Oberlander *et al.* 2014).

Oberlander *et al.* (2011) published a phylogeny of the southern African species, based on a combined data set of plastid *trn*L intron, *trn*L-*trn*F intergenic spacer, and *trn*S-*trn*G, as well as the internal transcribed spacers of nuclear ribosomal DNA (ITS), and found that these species are monophyletic. However, the phylogeny lacked good resolution and high support values for many clades, and strong cytonuclear discordance was observed. For Hyb-Seq, we selected 23 *Oxalis* species from the core southern African *Oxalis* clade (clade four of Oberlander *et al.* 2011). Twelve of those species were from the '*O. hirta* and relatives clade' (clade 11 of Oberlander *et al.* 2011; hereafter the Hirta clade), which showed strong cytonuclear discordance. *Oxalis hirta* L. and *Oxalis obtusa* Jacq. had two accessions each, as they are among the most morphologically variable taxa within the core southern African *Oxalis* clade. Approximately one-third of the accessions were polyploid. Silica-dried leaf material was used in all cases. Sampling information is available in Table S1 (Supporting information).

Target enrichment probe design

The transcriptome draft assembly of the cosmopolitan weed *Oxalis corniculata* L. from the 1KP initiative (accession JHCN) and genome skim raw data of an accession of southern African *O. obtusa* (accession J12; see 'Illumina library preparation and Hyb-Seq') were combined to get hundreds of orthologous LCN loci. Enrichment of multicopy loci was minimized by using unique

transcripts only, which were obtained by comparing all transcripts and removing those sharing ≥90% sequence similarity using BLAT v.32x1 (Kent 2002). Before matching the *O. obtusa* genome skim data against those unique transcripts, reads of plastid and mitochondrial origin were removed with BOWTIE 2 (Langmead & Salzberg 2012), SAMTOOLS (Li *et al.* 2009) and BAM2FASTQ (http://gsl.hudsonalpha.org/information/software/bam2fastq) utilizing *Ricinus communis* L. GenBank accessions NC_016736 and NC_015141 as references. Paired-end reads were subsequently combined with FLASH (Magoč & Salzberg 2011). These processed reads were matched against the unique *O. corniculata* transcripts sharing ≥85% sequence similarity with BLAT. Transcripts with >1000 BLAT hits, indicating repetitive elements, and *O. obtusa* BLAT hits containing masked nucleotides were removed before *de novo* assembly of the *O. obtusa* BLAT hits to larger contigs with GENEIOUS v.6.1.7 (Kearse *et al.* 2012), using the medium sensitivity/fast setting. After assembly, only those contigs that comprised exons ≥120 bp and had a total locus length ≥600 bp were retained. To ensure that probes did not target multiple similar loci, any probe sequences sharing ≥90% sequence similarity were removed using CD-HIT-EST v.4.5.4 (Li & Godzik 2006), followed by a second filtering step for contigs containing exons ≥120 bp and totalling loci length ≥600 bp. To ensure that plastid sequences were absent from the probes, the probe sequences were matched against the *Ricinus* plastome reference sharing ≥90% sequence similarity with BLAT, and the hits were removed from the probe set. Repetitive elements were then masked with REPEATMASKER (http://www.repeatmasker.org/). Ambiguous sites in the probe sequences, which were generated during assembly of the genome skim reads, were randomly replaced by one of the relevant nucleotides and stretches of up to 5N replaced by T. Tiling density 2× was used.

The workflow described above up to the removal of remaining plastid sequences is summarized in Fig. 1. It was implemented in an automated and interactive BASH script named Sondovač, which is deposited in GITHUB (https://github.com/V-Z/sondovac/wiki/) and licensed under open-source licence GPL v.3 allowing further modifications. The script runs on major Linux distributions and Mac OS X; it runs on a standard desktop computer equipped with modern CPU like Intel i5 or i7. Strong bioinformatics skills and access to high-performance computer clusters are not required.

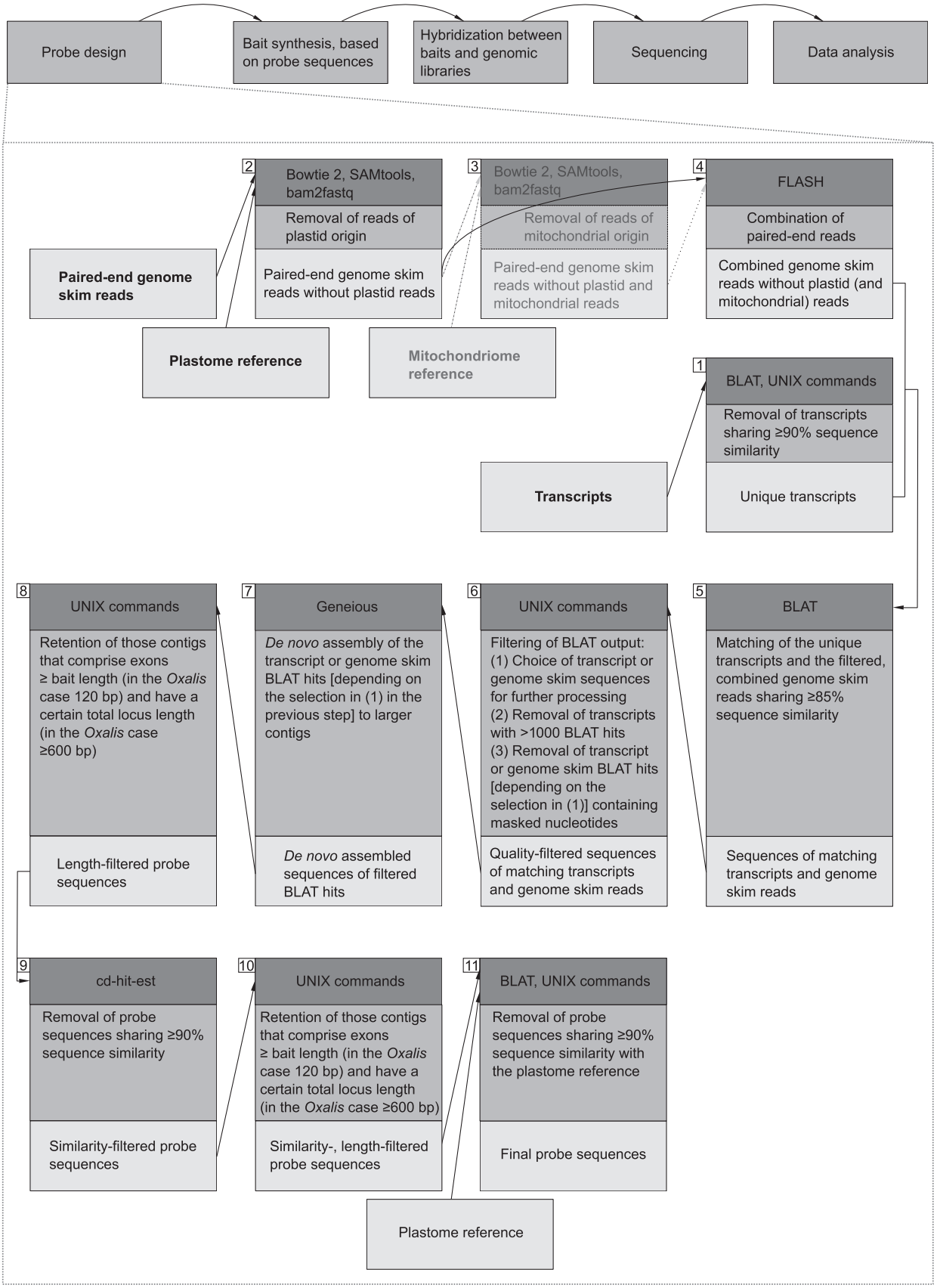
Illumina library preparation and Hyb-Seq

Genomic DNA was extracted according to the CTAB protocol (Doyle & Doyle 1987). The genome skim data of *O. obtusa* accession J12 was obtained with 250 bp paired-end reads from a partial lane of an Illumina MiSeq (San Diego, CA, USA) performed by StarSEQ (Mainz, Germany). For the other 23 species of *Oxalis*, used for Hyb-Seq, 200 ng to 1 μg extracted DNA was sheared with a Covaris (Woburn, MA, USA) S220 sonicator using the program for fragmentation to 1000 bp for 45 s (200 cycles, 4 °C). Library preparation followed the NEBNext Ultra DNA Library Prep (New England Biolabs, Ipswich, MA, USA) protocol for Illumina with a few modifications: (i) Size selection (~600–650 bp) was performed on a 1% agarose gel and (ii) two additional cleanup steps were implemented, one after adapter ligation with the QIAquick PCR Purification kit (Qiagen, Venlo, Netherlands), and a second after gel extraction with the QIAquick Gel Extraction kit (Qiagen). (iii) Enriched PCR products were cleaned up with the QIAquick PCR Purification Kit and subsequently with Agencourt AMPure XP beads (Beckman Coulter Genomics, Danvers, MA, USA). Amplification of ligated, size-selected fragments was performed with 10 cycles of PCR, using NEBNext Multiplex Oligos for Illumina Index Primers Set 1 and 2 (New England Biolabs). Libraries were subsequently pooled in approximately equimolar ratios in a 24-plex reaction. Solution hybridization with MyBaits biotinylated RNA baits (MYcroarray, Ann Arbor, MI, USA), which were synthesized from our custom-designed probes, and enrichment followed the MYBAITS manual v.1.3.8 with approximately 200 ng of input DNA (9 ng per accession) and 12 cycles of PCR enrichment. Target-enriched libraries were sequenced on an Illumina MiSeq at Oregon State University (v.3 chemistry) to obtain 150 bp paired-end reads. All DNA concentration measurements were performed with the QUBIT 2.0 fluorometer (Invitrogen/Life Technologies, Carlsbad, CA, USA).

Data analysis pipelines for quality filtering, assembly, alignment, and quality assessment of the Hyb-Seq data

Adapter sequences and low quality reads were removed with TRIMMOMATIC v.0.30 (Bolger *et al.* 2014). In case of quality <Q20 of read ends, these bases were discarded. The remaining part of the read was trimmed, if average

Fig. 1 Workflow of the probe design script Sondovač. An overview on the main steps of Hyb-Seq is given in the top part of the figure; probe design is the first one. Each step of Sondovač is numbered and illustrated by three boxes each: Software is highlighted in dark grey, a summary of each step is given in medium grey, and input/output of each step is depicted in light grey. Optional removal of reads of mitochondrial origin from the genome skim data is marked by decoloration of the text. The required input files of Sondovač are highlighted in bold. The direction of the workflow is indicated by arrows.



quality in a 5 bp window was <Q20, and removed, if read length fell below 36 bp after trimming. The FASTX-TOOLKIT (Gordon & Hannon 2010) was used to remove duplicate reads. Reference-guided assembly of the targeted loci was performed with ALIGNREADS v.2.25 (Straub *et al.* 2011), using the probe sequences separated by a string of 200 Ns each as reference and the parameter settings of Weitemier *et al.* (2014). Steps up to the final alignments of the LCN loci were performed following Weitemier *et al.* (2014). Although the LCN loci were created by randomly concatenating the respective exons, they will be called genes. All analyses based on the LCN genes were conducted on 727 loci that contained sequence information for at least part of these genes for all accessions; the remaining 437 genes out of the total 1164 targeted had completely missing sequences for certain accessions.

The plastid genome and the nrDNA cistron (18S-ITS1-5.8S-ITS2-26S) were also assembled with ALIGNREADS, utilizing *Oxalis* reference sequences that were built according to Straub *et al.* (2011, 2012). Genome skim reads of *O. obtusa* were quality trimmed, using the same settings as for the Hyb-Seq data, and duplicate reads were removed. ALIGNREADS was run with the following masking parameters, utilizing the plastid genome sequence of *Ricinus communis* (GenBank accession JF937588) as reference: Any base with sequencing depth <5 was masked, and single nucleotide polymorphisms (SNPs) were only called in *O. obtusa*, if 80% of reads supported that SNP with sequencing depth ≥25 at that site. Read type 454, accounting for longer read length, and linear setting were chosen in YASRA, the assembler within ALIGNREADS, and medium percentage sequence identity between genome skim data and reference sequence was chosen, as this resulted in the highest quality assembly (data not shown). Plastome regions present in the *Ricinus* reference, but absent in *O. obtusa*, were masked. Hyb-Seq reads were then assembled with guidance of the draft *Oxalis* plastome reference with the same settings used for read assembly of the targeted nuclear loci. The resulting contigs of all accessions were aligned with MULAN (Ovcharenko *et al.* 2005), utilizing the draft *Oxalis* plastome reference sequence. Positions in consensus sequences were masked in the alignment editor MEGA v.6 (Tamura *et al.* 2013) in case of (i) regions with many SNPs compared to the reference due to wrongly assembled indels or SNPs between overlapping contig ends, or (ii) insertions not found in other *Oxalis* accessions present at contig ends. Visual inspection of the assemblies was performed with TABLET v.1.14.04.10 (Milne *et al.* 2013).

The nrDNA cistron (without the external transcribed spacer due to repetitive elements) was assembled with ALIGNREADS based on an *Oxalis* reference that was built from a 774 bp partial sequence of *O. obtusa* (GenBank accession EU436922). The same masking parameters as for building the *Oxalis* draft plastome reference were chosen. Hyb-Seq reads were assembled to this reference with ALIGNREADS without masking parameters. In cases of more than one contig per accession and divergent SNPs between them, such sites were masked. The consensus sequences were aligned using MAFFT v.6864b (Katoh & Toh 2008) with default settings. Data completeness of all final alignments was calculated considering missing data per base pair, and in case of the LCN gene matrix also considering missing number of targeted exons and missing number of targeted genes, which were regarded missing if all targeted exons failed enrichment. Sequence similarity between probes and assembled reads of each *Oxalis* accession was estimated using BLAT, based on a minimum 85% sequence similarity.

Plastome annotation

Annotation of the draft plastid genome was performed with DOGMA (Wyman *et al.* 2004) on the *Oxalis* accession with the longest and one of the most complete plastome sequence (*Oxalis hirsuta* Sond.). The annotated plastome was visualized with GENOMEVX (Conant & Wolfe 2008).

Phylogenetic network analysis

Putative conflicting signals within the LCN gene, plastome and nrDNA cistron data sets were visualized with NEIGHBORNET (Bryant & Moulton 2004) implemented in SPLITSTREE v.4.13.1 (Huson & Bryant 2006) using concatenated sequence alignment. Networks were constructed based on uncorrected pairwise matrices. Bootstrapping was performed with 100 replicates.

Phylogenetic tree reconstructions and gene–gene/gene–species tree comparisons

Maximum-likelihood (ML) and Bayesian Markov chain Monte Carlo (MCMC)/Bayesian inference (BI) methods were used for phylogenetic reconstruction of gene trees. ML gene trees were run with RAXML v.7.3.0 (Stamatakis 2006) with the GTR + Γ nucleotide substitution model and rapid bootstrap with 100 replicates each. Bayesian MCMC analysis was performed with MRBAYES v.3.2 (Ronquist & Huelsenbeck 2003), utilizing sampling across substitution models and Γ correction. Two simultaneous runs with four chains each were performed for 5 million generations, in each run 1001 trees were sampled, of which the first 25% were discarded as burn-in. In order to diagnose convergence between the individual Bayesian MCMC runs in such a large data set, three measures

of convergence were chosen: (i) average effective sample size (avgESS) >100, (ii) potential scale reduction factor (PSRF) ~1.0, and (iii) average standard deviation of split frequencies (ASDSF) <0.05.

Phylogenetic hypotheses based on the targeted LCN genes were inferred in three different ways, which are intensely debated (e.g. von Haeseler 2012; DeGiorgio *et al.* 2014; Liu *et al.* 2015; Tonini *et al.* 2015): (i) species tree reconstruction under the multispecies coalescent model utilizing three different programs: (a) ASTRAL (Mirarab *et al.* 2014), which finds the species tree that agrees with the largest number of quartet trees induced by the set of gene trees, (b) MP-EST (Liu *et al.* 2010), which uses maximum pseudolikelihood for the estimation of species trees, and (c) STAR (Liu *et al.* 2009a), which uses average ranks of gene coalescence times to build species trees; bootstrapping of the STAR species tree was performed according to the multilocus bootstrap method of Seo (2008), (ii) a supertree approach using matrix representation with parsimony (MRP) (Baum 1992; Ragan 1992), and (iii) a supermatrix approach using concatenation of the data set. The first two methods were performed both on the set of MRBAYES maximum posterior probability trees and RAXML best trees. The concatenated data set was run with RAXML, applying the GTR + Γ nucleotide substitution model and rapid bootstrap with 100 replicates.

Topological differences between gene trees were assessed by calculating the Robinson-Foulds (RF) distance (Robinson & Foulds 1981) with the R function RF.dist of the phangorn package (Schliep 2011) and visualized as principal coordinate (PCoA) plot. Topological differences between the gene trees and the STAR species tree were calculated as modified RF distance (RFD) using STRAW (Shaw *et al.* 2013) and visualized as a histogram.

Phylogenetic reconstruction based on the draft plastid genome and nrDNA cistron was performed using MRBAYES on partitioned data sets. In the case of the plastome data, the alignment was partitioned into protein-, tRNA-, and rRNA-coding as well as intron/spacer (i.e. noncoding) regions. The nrDNA alignment was partitioned into 18S, ITS1, 5.8S, ITS2 and 26S. All partitions were sampled across substitution models, and an initial partition-specific among-site rate variation correction, estimated by PARTITIONFINDER v.1.1.1 (Lanfear *et al.* 2012), was employed, which was updated based on output from initial MRBAYES runs. All parameters were unlinked across partitions, and each partition had its own, independent evolutionary rate. Two simultaneous runs with four chains each were performed for 100 million generations, in each run 1001 trees were sampled, of which the first 25% were discarded as burn-in. For plastome analysis, the complete protein-coding plastid complement (Moore *et al.* 2010) and inverted repeat of *Oxalis latifolia* Kunth (GenBank accession HQ664602) was used as outgroup, whereas *O. corniculata* served as outgroup both in the LCN and nrDNA cistron data set. However, as *O. corniculata* and *O. latifolia* are phylogenetically very distantly related, and the branch leading to *O. corniculata* was very long in all analyses, these outgroups were trimmed from the final trees, and trees were rooted using *Oxalis imbricata* Eckl. & Zeyh. in the comparison of phylogenetic hypotheses based on all three data sets.

Results

Probe design, sequencing, quality trimming, assembly, alignment, and quality assessment

Design of *Oxalis* LCN probes resulted in 4926 exons ≥120 bp length and 1164 genes (Table S2, Supporting information) of total 1 127 209 bp. Gene length ranged from 600 to 4125 bp with a mean of 968 bp. Adapter trimming and quality filtering resulted in an average loss of 7% of the reads (Table S3, Supporting information). In *O. palmifrons* T.M. Salter 22% of the reads were dropped, indicating its poor initial DNA quality; DNA of this accession was nearly degraded. After duplicate read removal on average 47% of quality-filtered reads remained (Table S3, Supporting information). This was a high number of duplicate reads, likely the result of PCR duplicates from too many PCR cycles during genomic library preparation and enrichment. Of the quality-filtered reads after duplicate removal on average 59% mapped to the LCN genes (on-target), 27% to the draft plastome and 3% to the nrDNA cistron reference (Table S3, Supporting information). The mean number of targeted loci with complete or partial sequence information was 4124 for exons (from total of 4926) and 1141 for genes (from total of 1164) (Table S2, Supporting information). Sequences diverged by 3–4% between reads and LCN probes (Table S2, Supporting information), and sequence divergence was 11% between probes and the *O. corniculata* transcriptome used to define exon boundaries. Completeness of the data sets was as follows: 90% for the LCN exons, 96% for the plastome, and 96% for the nrDNA cistron (Table S2, Supporting information). These mean values are an underestimate due to two accessions with exceptionally low read numbers (*O. hirta* J62, *O. palmifrons*). Mean sequencing depth for the data sets was 16 (LCN exons), 173 (plastome) and 290 (nrDNA cistron) (Table S2, Supporting information).

Plastome annotation

The plastid genome was annotated to enable partitioning of the alignment for phylogenetic reconstruction, and the

annotation can be summarized as follows (Fig. S1, Supporting information): 79 protein-coding genes were found, including three genes that were absent in closely related *Ricinus* (*infA*, *ycf15*, *ycf68*). Two genes (*rps16*, *rpl32*) were missing in *Oxalis* compared to *Ricinus*. All 30 tRNA-coding genes were present, as were all four rRNA-coding genes.

Phylogenetic reconstructions

Phylogenetic networks. In the phylogenetic network based on the concatenated LCN gene matrix (Fig. S2a, Supporting information) splits were numerous but short, and in the plastome network (Fig. S2b, Supporting information) splits were largely absent, thereby justifying the use of phylogenetic tree reconstruction methods for those two data sets. In contrast, the nrDNA cistron network contained numerous splits (Fig. S2c, Supporting information).

Robustness of phylogenetic hypothesis based on the LCN genes. Although all avgESS estimates were satisfactory, a small minority of PSRF and ASDSF values (<20) strongly differed from expectation (Fig. S3, Supporting information). Removing these loci from species tree reconstruction idiosyncratically affected the uncertain nodes discussed in the following (data not shown), and the loci were kept. Species tree topology was relatively robust to both the method of species and gene tree reconstruction, as the major clades were obtained and relationships within clades were nearly identical in all cases (Fig. S4, Supporting information): (i) Topology within the Hirta clade was identical between the different methods except for the weakly supported position of *Oxalis primuloides* R. Knuth in the concatenated tree with 45% bootstrap support (BS). (ii) In all cases *Oxalis amblyosepala* Schltr. and *Oxalis polyphylla* Jacq. formed a clade, which was in sister relationship to the Hirta clade. (iii) *Oxalis hirsuta*, *Oxalis inconspicua* T.M. Salter, *O. obtusa*, and *Oxalis pulchella* Jacq. formed a clade, and within-clade relationships were identical with all methods except for the MRP supertree based on BI of gene trees. Clustering of *O. inconspicua* and the outgroup *O. corniculata* was apparently due to long-branch attraction: These two species exhibited the longest branches in the tree, which is likely to result in clustering together, independent on the relationships of the underlying sequences (Felsenstein 1978). (iv) Phylogenetic placement of *O. imbricata*, *Oxalis orthopoda* T.M. Salter, *Oxalis truncatula* Jacq., *O. palmifrons* and *Oxalis smithiana* Eckl. & Zeyh. partly deviated between the different methods. In all cases the latter two were in a successive sister relationship to the clade comprising the Hirta clade and *O. amblyosepala* and *O. polyphylla*.

The number of incongruent nodes between species trees based on BI of gene trees vs. ML gene trees was two to three, and in a comparison of all utilized species tree methods with concatenation there were six incongruent nodes (Fig. S4, Supporting information). All nodes of the STAR tree had 100% BS. The concatenated data set showed weak support for all those nodes, which were discussed above in (i) and (iv) as ambiguous in phylogenetic placement; otherwise the BS values were also 100%.

Incongruences between gene–gene and gene–species trees. The extensive and largely homogeneous cluster of RF distances between gene trees shown in the PCoA demonstrated the great variability of gene tree topologies (Fig. S5, Supporting information). RF distances between the gene trees and the STAR species tree, displayed in a histogram (Fig. S6, Supporting information), was on average RFD = 34, which is a relatively high value, considering that the maximum RF value for this number of accessions is RFD = 46.

Comparison between phylogenetic trees based on the LCN genes, the plastome, and the nrDNA cistron. The STAR species tree reconstruction method was chosen to compare data sets (Fig. 2). The plastome and nrDNA cistron trees resulted in the same major clades as the LCN gene data set: the Hirta clade, *O. amblyosepala* and *O. polyphylla* as sister, and *O. hirsuta*, *O. inconspicua*, *O. obtusa*, and *O. pulchella* as a clade. However, relationships within the Hirta clade were incongruently resolved, especially in the plastome tree (Fig. 2a). In the nrDNA cistron tree (Fig. 2b) there were numerous polytomies and partly low posterior probability (PP), which is possibly the result of a lack of parsimony informative sites in the ribosomal RNAs combined with a high rate of evolution in the spacers, hindering direct comparison. Similar to the comparison of reconstruction methods of trees based on the LCN loci, phylogenetic placement of *O. imbricata*, *O. orthopoda*, *O. truncatula*, *O. palmifrons*, and *O. smithiana* deviated between the trees based on the three data sets; the differences were slightly stronger, as *O. palmifrons* and *O. smithiana* were not in sister relationship to the clade comprising the Hirta clade and *O. amblyosepala* and *O. polyphylla* in the plastome and nrDNA cistron trees. All nodes of the plastome tree were supported with 1 PP except for the *Oxalis ciliaris* Jacq./*O. hirta* and *Oxalis tenella* Jacq./*Oxalis callosa* R. Knuth – *O. primuloides* splits (Fig. 2a). The 18 incongruent nodes between the species tree based on the LCN loci and the plastome tree revealed the strong cytonuclear discordance, which was underlined by the high support values in both the plastome tree and species tree based on the LCN genes (Fig. 2a).

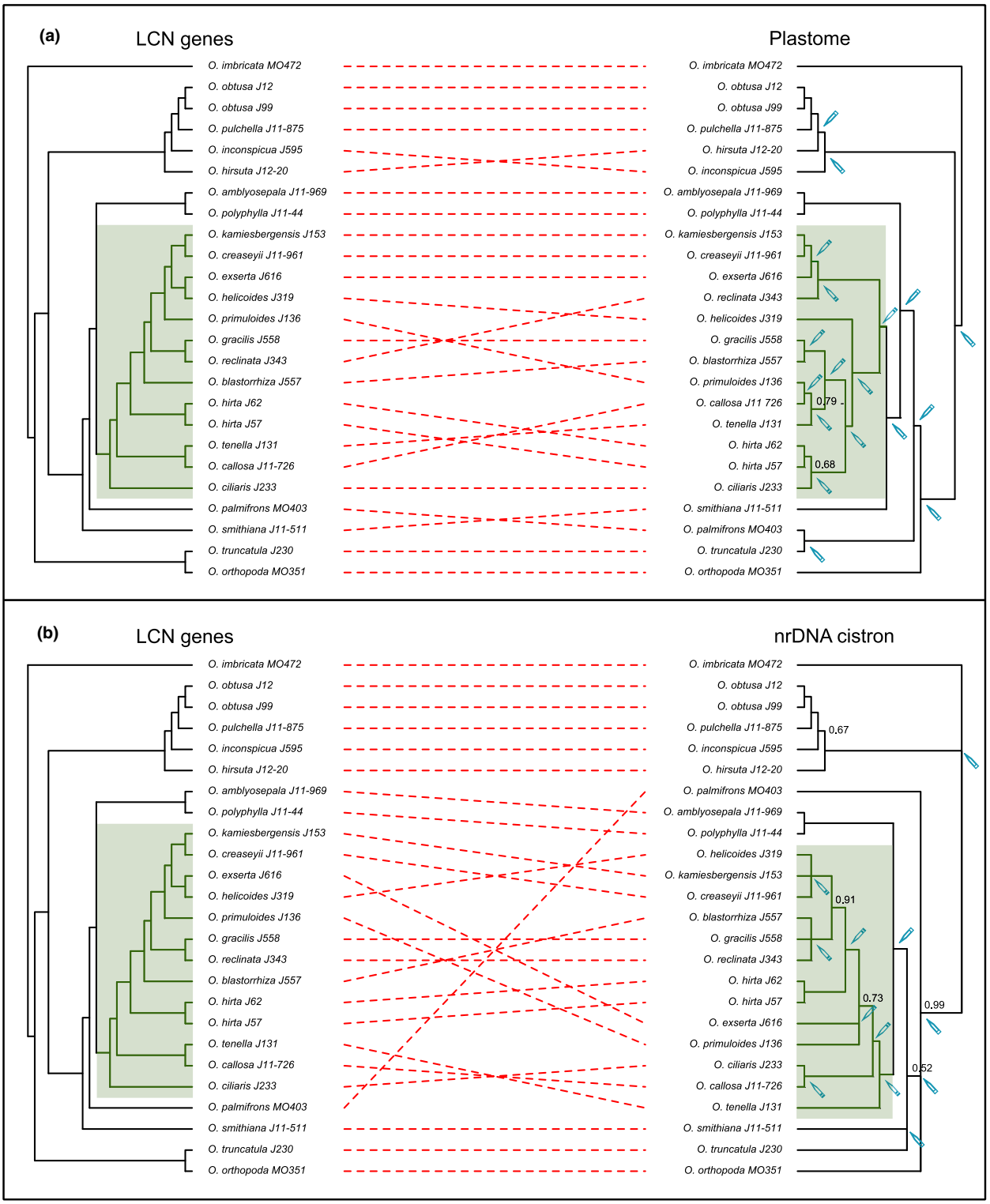


Fig. 2 Comparison of phylogenetic hypotheses based on 727 LCN genes, the plastid genome and the nrDNA cistron of southern African *Oxalis*. (a) Comparison between the STAR species tree, based on maximum-likelihood gene trees, and the plastome tree. (b) Comparison between the STAR species tree and the nrDNA cistron tree. Dashed lines connect each accession to its placement in the contrasting tree. Values close to the nodes are bootstrap support (* = 100%) in the STAR species tree or posterior probability (* = 1) in the plastome and nrDNA cistron tree. The Hirta clade (Oberlander *et al.* 2011) is coloured in light grey (respectively, green in the online colour figure). Nodes that are incongruent between the contrasting phylogenetic trees of the different data sets are marked with arrows.

Discussion

The value of our target enrichment probe design for plant phylogenetics and its application in Hyb-Seq

Our LCN probe design pipeline, implemented in the BASH script Sondovač, generates LCN loci from a combination of transcriptome and genome skim data. Many transcriptomes can be taken from the already existing HTS resource 1KP initiative, in which approximately 70% of APG III families are represented (APG III 2009). Depending on the phylogenetic level of the group under study and the number of LCN genes users want to obtain, a transcriptome of a more or less closer relative of the study group must be utilized. Users need to provide paired-end genome skim data of one of the taxa of their study group, as this accounts for the specificity of the probe design. The reference sequences of the plastid and possibly mitochondrial genome, which are used in the pipeline to remove organellar reads from the genome skim read pool, are also part of existing HTS resources (http://www.ncbi.nlm.nih.gov/genome/organelle/); plastome reference sequences from taxa up to the same order of the studied plant group are suitable (Straub *et al.* 2012).

By running Sondovač for southern African *Oxalis*, over 1000 orthologous LCN genes of adequate length were obtained. Target enrichment of these loci resulted in a nearly complete data matrix. Efficiency of target enrichment of the LCN loci was similar to that reported for the *Asclepias* data set by Weitemier *et al.* (2014), if considering only their accessions with external barcodes; there were approximately 60% on-target, quality-filtered reads after duplicate removal, although sequence divergence to the LCN probes was larger for *Oxalis* (average 4% compared to 1.5% for *Asclepias*). The average sequencing depth 16× of LCN exons should facilitate unambiguous SNP calling in orthologous genes, but also enable the identification of paralogous genes in polyploid accessions, which will be an essential step towards using target enrichment on polyploids. Plastome assembly and annotation suggest that it is possible to obtain (nearly) the whole plastid genome with Hyb-Seq. Only a few quickly evolving regions, such as the full-length copy of *ycf1*, were partial in this analysis.

Towards a refined phylogeny of southern African Oxalis

Although the selected 23 *Oxalis* species comprised only a small subset of species from the southern African *Oxalis* clade, preliminary phylogenetic conclusions and evolutionary implications can be drawn. Major phylogenetic relationships based on the LCN loci did not strongly differ from the published phylogeny of Oberlander *et al.* (2011), but node resolution and support improved dramatically. Topologies of the trees based on the LCN loci were relatively robust to the methods by which they were obtained (multispecies coalescent, MRP supertree, concatenation) and also to the methods through which the gene trees were estimated (BI and ML). Only few highly supported, conflicting relationships were found between the concatenated tree and the coalescent trees, indicating that the coalescent model likely reduces to the concatenation model in this case (Liu *et al.* 2009b, 2015).

The phylogenetic tree topology based on the LCN genes showed widespread and strong incongruences with the plastome tree topology. Cytonuclear discordance is considered as evidence for either incomplete lineage sorting (ILS) of the chloroplast or chloroplast introgression (chloroplast capture). Studies usually interpreted this discordance in terms of hybridization (e.g. *Helianthus* L.: Dorado *et al*. 1992; *Mitella* L.: Okuyama *et al*. 2005; *Ficus* L.: Renoult *et al*. 2009; Senecioneae Cass.: Pelser *et al.* 2010), although the presence of ILS was rarely tested. In addition to cytonuclear discordance, both gene–gene and gene–species trees showed quite strong topological incongruences in *Oxalis*. There was an immense variety of gene tree topologies in general, and gene tree topologies strongly differed from the species tree topology. The underlying reasons for this topological variation need to be found by testing for ILS, hybridization, paralogy, and a combination of those under coalescent models (e.g. Rasmussen & Kellis 2012; Yu *et al.* 2014). Given the very short branch lengths at the base of the southern African *Oxalis* clade (Oberlander *et al.* 2011) as well as generally large population sizes of most species, strong ILS effects are perhaps to be expected, particularly at the base of the putative southern African radiation. The extent of hybridization in *Oxalis* is poorly known. The only confirmed hybrid is South American *Oxalis tuberosa* Molina (Emshwiller & Doyle 2002). Both heterostyly and pollinator specificity may reduce the potential for hybridization: In *Oxalis* tristyly is predominant and distyly rare (Weller *et al.*, 2007, Gardner *et al.*, 2012), meaning that three (in distyly two) floral morphs with reciprocal placement of stigma and anthers persist in a population. Heterostyly promotes precise pollen transfer, which could thus promote reproductive isolation and prevent hybridization, if the position of sexual organs is not similar between the mating individuals (Keller *et al.* 2012). Dense taxon sampling in the southern African lineage is required to address the extent of hybridization. Finally, gene duplication and loss could well be a major cause of the observed incongruences, especially in the polyploid accessions, but testing that requires the identification of paralogous genes. A bioinformatic pipeline for paralog identification of

LCN genes from target enrichment data is currently under development by AL.

The observed, strong cytonuclear discordance suggests that organellar phylogenies alone are unlikely to represent the species tree (Davis *et al*. 2014), and the strong incongruences between gene–gene and gene–species trees imply that a large number of LCN loci is needed to robustly resolve phylogenies in the presence of ILS and hybridization, especially of putatively radiating lineages such as southern African *Oxalis*.

Acknowledgements

The authors thank Gane Ka-Shu Wong (University of Alberta) and Douglas E. Soltis (University of Florida) for access to the 1KP transcriptome data, Tomáš Fér (Charles University in Prague) for access to the genome skim data, Lenka Flašková (Charles University in Prague) for DNA extraction, Filip Pardy (Central European Institute of Technology, Brno) for help with sonication, Mark Dasenko (Oregon State University Center for Genome Research and Biocomputing/CGRB) for Illumina sequencing support, Sanjuro Jodgeo (Oregon State University) for data analysis support, and Daisie Huang, Jennifer Mandel and an anonymous reviewer for constructive comments. Data were analysed on the CGRB computer cluster, on MetaCentrum under the program LM2010005, and on CERIT-SC under the program Centre CERIT Scientific Cloud. MetaCentrum and CERIT-SC are part of the Operational Program Research and Development for Innovations, Reg. no. CZ.1.05/3.2.00/08.0144, Czech Republic. This work, including three stays in the Liston Laboratory, was financially supported to R.S. by project Reg. no. CZ.1.07/2.3.00/30.0048 of the European Social Fund in the Czech Republic through the Operational Program Education for Competitiveness. Additional funding came from the long-term research development projects RVO 67985939 (The Czech Academy of Sciences) and institutional resources of the Ministry of Education, Youth and Sports of the Czech Republic for the support of science and research.

References

Baum BR (1992) Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon*, **41**, 3–10.

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

Born J, Linder HP, Desmet P (2006) The greater cape floristic region. *Journal of Biogeography*, **34**, 147–162.

Bryant D, Moulton V (2004) Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution*, **21**, 255–265.

Chamala S, García N, Godden GT *et al*. (2015) MarkerMiner 1.0: a new application for phylogenetic marker development using angiosperm transcriptomes. *Applications in Plant Sciences*, **3**, 1400115.

Conant GC, Wolfe KH (2008) GenomeVx: simple web-based creation of editable circular chromosome maps. *Bioinformatics*, **24**, 861–862.

Davis CC, Xi Z, Mathews S (2014) Plastid phylogenomics and green plant phylogeny: almost full circle but not quite there. *BMC Biology*, **12**, 11.

De Sousa F, Bertrand YJK, Nylinder S *et al*. (2014) Phylogenetic properties of 50 nuclear loci in *Medicago* (Leguminosae) generated using multiplexed sequence capture and next-generation sequencing. *PLoS ONE*, **9**, e109704.

DeGiorgio M, Syring J, Eckert AJ *et al*. (2014) An empirical evaluation of two-stage species tree inference strategies using a multilocus dataset from North American pines. *BMC Evolutionary Biology*, **14**, 67.

Dorado O, Rieseberg LH, Arias DM (1992) Chloroplast DNA introgression in southern California sunflowers. *Evolution*, **46**, 566–572.

Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin*, **19**, 11–15.

Emshwiller E, Doyle JJ (2002) Origins of domestication and polyploidy in oca (*Oxalis tuberosa*; Oxalidaceae). 2. Chloroplast-expressed glutamine synthetase data. *American Journal of Botany*, **89**, 1042–1056.

Faircloth BC, McCormack JE, Crawford NG *et al*. (2012) Ultraconserved elements anchor thousands of genetic markers for target enrichment spanning multiple evolutionary timescales. *Systematic Biology*, **61**, 717–726.

Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, **27**, 401–410.

Gardner AG, Vaio M, Guerra M, Emshwiller E (2012) Diversification of the American bulb-bearing *Oxalis* (Oxalidaceae): dispersal to North America and modification of the tristylous breeding system. *American Journal of Botany*, **99**, 152–164.

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

Gordon A, Hannon GJ (2010) FASTX-Toolkit. FASTQ/A short-reads preprocessing tools. http://hannonlab.cshl.edu/fastx_toolkit/ [accessed 29 May 2015].

Grover CE, Gallagher JP, Jareczek JJ *et al*. (2015) Re-evaluating the phylogeny of allopolyploid *Gossypium* L.. *Molecular Phylogenetics and Evolution*, **92**, 45–52.

von Haeseler A (2012) Do we still need supertrees? *BMC Biology*, **10**, 13.

Hedtke SM, Morgan MJ, Cannatella DC, Hillis DM (2013) Targeted enrichment: maximizing orthologous gene comparisons across deep evolutionary time. *PLoS ONE*, **8**, e67908.

Heyduk K, Trapnell DW, Barrett CF, Leebens-Mack J (2015) Phylogenomic analyses of species relationships in the genus *Sabal* (Arecaceae) using targeted sequence capture. *Biological Journal of the Linnean Society* doi: 10.1111/bij.12551. [Epub ahead of print]

Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, **23**, 254–267.

Jeffroy O, Brinkmann H, Delsuc F, Philippe H (2006) Phylogenomics: the beginning of incongruence. *Trends in Genetics*, **22**, 225–231.

Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics*, **9**, 286–298.

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

Keller B, De Vos JM, Conti E (2012) Decrease of sexual organ reciprocity between heterostylous primrose species, with possible functional and evolutionary implications. *Annals of Botany*, **110**, 1233–1244.

Kent WJ (2002) BLAT – the BLAST-like alignment tool. *Genome Research*, **12**, 656–664.

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

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**, 357–359.

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Li H, Handsaker B, Wysoker A *et al*. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Liu L, Yu L, Kubatko L *et al*. (2009a) Coalescent methods for estimating phylogenetic trees. *Molecular Phylogenetics and Evolution*, **53**, 320–328.

Liu L, Yu L, Pearl DK, Edwards SV (2009b) Estimating species phylogenies using coalescence times among sequences. *Systematic Biology*, **58**, 468–477.

Liu L, Yu L, Edwards SV (2010) A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology*, **10**, 302.

Liu L, Xi Z, Wu S, Davis C, Edwards SV (2015) Estimating phylogenetic trees from genome-scale data. *Annals of the New York Academy of Sciences*, **1360**, 36–53.

Magoč T, Salzberg SL (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, **27**, 2957–2963.

Mandel JR, Dikow RB, Funk VA *et al.* (2014) A target enrichment method for gathering phylogenetic information from hundreds of loci: an example from the Compositae. *Applications in Plant Sciences*, **2**, 1300085.

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.

Milne I, Stephen G, Bayer M *et al.* (2013) Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics*, **14**, 193–202.

Mirarab S, Reaz R, Bayzid MDS *et al.* (2014) ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, **30**, i541–i548.

Moore MJ, Soltis PS, Bell CD *et al.* (2010) Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 4623–4628.

Nicholls JA, Pennington RT, Koenen EJM *et al.* (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.

Oberlander KC, Dreyer LL, Bellstedt DU (2011) Molecular phylogenetics and origins of southern African *Oxalis*. *Taxon*, **60**, 1667–1677.

Oberlander KC, Roets F, Dreyer LL (2014) Pre-Pleistocene origin of an endangered habitat: links between vernal pools and aquatic *Oxalis* in the Greater Cape Floristic Region of South Africa. *Journal of Biogeography*, **41**, 1572–1582.

Okuyama Y, Fujii N, Wakabayashi M *et al.* (2005) Nonuniform concerted evolution and chloroplast capture: heterogeneity of observed introgression patterns in three molecular data partition phylogenies of Asian *Mitella* (Saxifragaceae). *Molecular Biology and Evolution*, **22**, 285–296.

Ovcharenko I, Loots GG, Giardine BM *et al.* (2005) Mulan: multiple-sequence local alignment and visualization for studying function and evolution. *Genome Research*, **15**, 184–194.

Parks M, Cronn R, Liston A (2009) Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC Biology*, **7**, 84.

Parks M, Cronn R, Liston A (2012) Separating the wheat from the chaff: mitigating the effects of noise in a plastome phylogenomic dataset from *Pinus* L. (Pinaceae). *BMC Evolutionary Biology*, **12**, 100.

Pelser PB, Kennedy AH, Tepe EJ *et al.* (2010) Patterns and causes of incongruence between plastid and nuclear Senecioneae (Asteraceae) phylogenies. *American Journal of Botany*, **97**, 856–873.

Pillon Y, Johansen J, Sakishima T *et al.* (2014) Primers for low-copy nuclear genes in *Metrosideros* and cross-amplification in Myrtaceae. *Applications in Plant Sciences*, **2**, 1400049.

Proches S, Cowling RM, Goldblatt P *et al.* (2006) An overview of the Cape geophytes. *Biological Journal of the Linnean Society*, **87**, 27–43.

Pyron RA (2015) Post-molecular systematics and the future of phylogenetics. *Trends in Ecology and Evolution*, **30**, 384–389.

Ragan MA (1992) Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution*, **1**, 53–58.

Rasmussen MD, Kellis M (2012) Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Research*, **22**, 755–765.

Reneker J, Lyons E, Conant GC *et al.* (2012) Long identical multispecies elements in plant and animal genomes. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, E1183–E1191.

Renoult JP, Kjellberg F, Grout C *et al.* (2009) Cyto-nuclear discordance in the phylogeny of *Ficus* section *Galoglychia* and host shifts in plant-pollinator associations. *BMC Evolutionary Biology*, **9**, 248.

Robinson DF, Foulds LR (1981) Comparison of phylogenetic trees. *Mathematical Biosciences*, **53**, 131–147.

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

Rothfels CJ, Larsson A, Li F-W *et al.* (2013) Transcriptome-mining for single-copy nuclear markers in ferns. *PLoS ONE*, **8**, e76957.

Schliep KP (2011) Phangorn: phylogenetic analysis in R. *Bioinformatics*, **27**, 592–593.

Seo TK (2008) Calculating bootstrap probabilities of phylogeny using multilocus sequence data. *Molecular Biology and Evolution*, **25**, 960–971.

Shaw TI, Ruan Z, Glenn TC, Liu L (2013) STRAW: species TRee analysis web server. *Nucleic Acids Research*, **41**, W238–W241.

Small RL, Cronn RC, Wendel JF (2004) Use of nuclear genes for phylogeny reconstruction in plants. *Australian Systematic Botany*, **17**, 145–170.

Smith BT, Harvey MG, Faircloth BC *et al.* (2013) Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. *Systematic Biology*, 1–13.

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

Stephens JD, Rogers WL, Heyduk K *et al.* (2015a) Resolving phylogenetic relationships of the recently radiated carnivorous plant genus *Sarracenia* using target enrichment. *Molecular Phylogenetics and Evolution*, **85**, 76–87.

Stephens JD, Rogers WL, Mason CM *et al.* (2015b) Species tree estimation of diploid *Helianthus* (Asteraceae) using target enrichment. *American Journal of Botany*, **102**, 910–920.

Straub SCK, Fishbein M, Livshultz T *et al.* (2011) Building a model: developing genomic resources for common milkweed (*Asclepias syriaca*) with low coverage genome sequencing. *BMC Genomics*, **12**, 211.

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

Tamura K, Stecher G, Peterson D *et al.* (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution*, **30**, 2725–2729.

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

Tonini J, Moore A, Stern D *et al.* (2015) Concatenation and species tree methods exhibit statistically indistinguishable accuracy under a range of simulated conditions. *PLoS Currents Tree of Life*, doi: 10.1371/currents.

Tonnabel J, Olivieri I, Mignot A *et al.* (2014) Developing nuclear DNA phylogenetic markers in the angiosperm genus *Leucadendron* (Proteaceae): a next-generation sequencing transcriptomic approach. *Molecular Phylogenetics and Evolution*, **70**, 37–46.

Weitemier K, Straub SCK, Cronn RC *et al.* (2014) Hyb-Seq: combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences*, **2**, 1400042.

Weller SG, Domínguez CA, Molina-Freaner FE *et al.* (2007) The evolution of distyly from tristyly in populations of Oxalis alpina (Oxalidaceae) in the Sky Islands of the Sonoran Desert. *American Journal of Botany*, **94**, 972–985.

Wyman SK, Jansen RK, Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. *Bioinformatics*, **20**, 3252–3255.

Yu Y, Dong J, Liu KJ, Nakhleh L (2014) Maximum likelihood inference of reticulate evolutionary histories. *Proceedings of the National Academy of Sciences of the United States of America*, **111**, 16448–16453.

A.L., R.S. and J.S. designed the study; R.S. and A.L. developed the probe design pipeline; V.Z. converted this pipeline into the BASH script Sondovač; J.S., K.O. and L.D. provided samples; R.S. conducted laboratory work; K.W., S.C.K.S. and R.C.C. advised on data collection and

computational analyses, R.S., V.Z. and K.O. analysed the data; and R.S. wrote the initial draft of the manuscript. All authors contributed to the final version of the manuscript.

Data accessibility

Sondovač is deposited in GITHUB (https://github.com/V-Z/sondovac/wiki/) together with the input *Oxalis* data (genome skim paired-end reads, plastid and mitochondrial reference). There we provided a link to the transcriptome data, which were obtained in the framework of the 1KP initiative. Raw reads, assemblies, alignments, phylogenetic trees, the GENEIOUS *de novo* assembly output and the final probe sequences are available under Dryad doi: 10.5061/dryad.dn08t.

Supporting Information

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

Fig. S1 Map of the draft plastid genome of *Oxalis hirsuta*.

Fig. S2 Splits network of the LCN gene matrix, the plastome and the nrDNA cistron datasets.

Fig. S3 Convergence measures of Bayesian MCMC runs.

Fig. S4 Comparison of phylogenetic trees based on 727 LCN genes.

Fig. S5 Principal coordinate analysis of Robinson-Foulds (RF) distances between gene trees.

Fig. S6 Robinson-Foulds (RF) distances between gene trees and the species tree.

Table S1 Voucher information of southern African *Oxalis* accessions used in this study.

Table S2 Success of Hyb-Seq in terms of number of reads after duplicate read removal, number of assembled LCN exons and genes, sequence divergence between the LCN matrix and the LCN probes, and completion and sequencing depth of all three datasets of southern African *Oxalis*.

Table S3 Success of Hyb-Seq in terms of number of raw reads after removal of PhiX reads, reads after quality-filtering, reads after duplicate read removal, number and proportion of reads mapped to the LCN probes (on-target), the plastome reference and the nrDNA cistron reference in southern African *Oxalis*.