This is the peer reviewed version of the following article:


**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.

which has been published in final form at

[https://doi.org/10.1111/1755-0998.12487](https://doi.org/10.1111/1755-0998.12487).

This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.
Phylogenetic marker development for target enrichment from transcriptome and genome skim data: the pipeline and its application in southern African Oxalis (Oxalidaceae)

Roswitha Schmickl¹, Aaron Liston², Vojtěch Zeisek¹,³, Kenneth Oberlander¹,⁴, Kevin Weitemier², Shannon C.K. Straub⁵, Richard C. Cronn⁶, Léanne L. Dreyer⁷, Jan Suda¹,³

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

*Correspondence: Roswitha Schmickl; Institute of Botany, The Czech Academy of Sciences, Zámek 1, 252 43 Průhonice, Czech Republic; Fax: +420 221 951 645; E-mail: roswitha.schmickl@ibot.cas.cz

Keywords: cytonuclear discordance, genome skimming, low-copy nuclear genes, Oxalis, species tree, target enrichment
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 non-model 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 1,164 LCN genes greater than 600 bp. Using target enrichment combined with genome skimming (Hyb-Seq) we obtained on average 1,141 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.

Introduction
High-throughput sequencing (HTS) has the potential to greatly increase the amount of phylogenetically informative signal in molecular datasets (Parks et al., 2009, 2012) and overcome difficulties in phylogenetic reconstructions, such as polytomies and low support values, that 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 non-exonic 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, Weitemier et al., 2014, Grover et al., 2015, Heyduk et al., 2015, Stephens et al., 2015a, b, Mandel et al., 2015, Nicholls et al., 2015) due to the paucity of ultraconserved nuclear sequences (Reneker et al., 2012).

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 datasets, 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 non-model 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 1,000 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 datasets 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. (ca. 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 ca. 46% of the genus (ca. 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 mil. years (Oberlander et al., 2014).

Oberlander et al. (2011) published a phylogeny of the southern African species, based on a combined dataset of plastid trnL intron, trnL-trnF intergenic spacer, and trnS-trnG, 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 multi-copy 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 and 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 and 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 and Godzik, 2006), followed by a second filtering step for contigs containing exons ≥120 bp and totaling 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 Figure 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 license 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 and 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, California, 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, Massachusetts, USA) S220 sonicator using the program for fragmentation to 1,000 bp for 45 sec (200 cycles, 4°C). Library preparation followed the NEBNext Ultra DNA Library Prep (New England Biolabs, Ipswich, Massachusetts, USA) protocol for Illumina with a few modifications: (1) Size-selection (~600–650 bp) was performed on a 1% agarose gel and (2) 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). (3) Enriched PCR products were cleaned up with the QIAquick PCR Purification Kit and subsequently with Agencourt AMPure XP beads (Beckman Coulter Genomics, Danvers, Massachusetts, 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, Michigan, 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, California, 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 quality in a 5 bp window was <Q20, and removed, if read length fell below 36 bp after trimming. The FASTX-Toolkit (Gordon and 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 1,164 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 (1) regions with many SNPs compared to the reference due to wrongly assembled indels or SNPs between overlapping contig ends, or (2) 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 and 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 and Wolfe, 2008).

**Phylogenetic network analysis**

Putative conflicting signals within the LCN gene, plastome and nrDNA cistron datasets were visualized with NeighborNet (Bryant and Moulton, 2004) implemented in SplitsTree v.4.13.1 (Huson and Bryant, 2006) using concatenated sequence alignments, and networks were constructed based on uncorrected pairwise matrices. Bootstrapping was performed with 100 replicates.

**Phylogenetic tree reconstructions and 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 + Ѕ nucleotide substitution model and rapid bootstrap with 100 replicates each. Bayesian MCMC analysis was performed with MrBayes v.3.2 (Ronquist and Huelsenbeck, 2003), utilizing sampling across substitution models and Ѕ correction. Two simultaneous runs with four chains each were performed for 5 mil. generations,
in each run 1,001 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 dataset, three measures of convergence were chosen: (1) average effective sample size (avgESS) >100, (2) potential scale reduction factor (PSRF) ~1.0, and (3) 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): (1) species tree reconstruction under the multispecies coalescent model utilizing three different programs: (i) 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, (ii) MP-EST (Liu et al., 2010), which uses maximum pseudo-likelihood for the estimation of species trees, and (iii) STAR (Liu et al., 2009), 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), (2) a supertree approach using matrix representation parsimony (MRP) (Baum, 1992, Ragan, 1992), and (3) a supermatrix approach using concatenation of the dataset. The first two methods were performed both on the set of MrBayes maximum posterior probability trees and RAxML best trees. The concatenated dataset 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 and 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 datasets. In the case of the plastome data the alignment was partitioned into protein-, tRNA-, and rRNA-coding as well as intron/spacer (i.e., non-coding) 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 mil. generations, in each run 1,001 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 dataset. 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 datasets.

Results

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

Design of Oxalis LCN probes resulted in 4,926 exons ≥120 bp length and 1,164 genes (Table S2, Supporting information) of total 1,127,209 bp. Gene length ranged from 600 to 4,125 bp with a mean of 968 bp. Adapter 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 4,124 for exons (from total of 4,926) and 1,141 for genes (from total of 1,164) (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 datasets were 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 datasets 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**

1) 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 datasets. In contrast, the nrDNA cistron network contained numerous splits (Fig. S2C, Supporting information).

2) 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): (1) 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). (2) In all cases *Oxalis amblyosepala* Schltr. and *Oxalis polyphylla* Jacq. formed a clade, which was in sister relationship to the Hirta clade. (3) *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). (4) 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 versus 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 dataset showed weak support for all those nodes, which were
discussed above in (1) and (4) as ambiguous in phylogenetic placement; otherwise
the BS values were also 100%.

3) 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.

4) 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 datasets
(Figure 2). The plastome and nrDNA cistron trees resulted in the same major clades
as the LCN gene dataset: 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 (Figure 2A). In the nrDNA cistron tree (Figure 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 datasets; 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 (Figure 2A). The 18 incongruent nodes between the species tree based on the LCN loci and the plastome tree revealed the strong cytonuclear discordance, which is supported by the high support values in both the plastome tree and species tree based on the LCN genes (Figure 2A).

**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 organelar 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 1,000 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* dataset 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 $16 \times$ 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., 2009, 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 statistically 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 and 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 and 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., 2011), 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 paralogue 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.

Author contributions
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.

References


Tonnabel J, Olivieri I, Mignot A et al. (2014) Developing nuclear DNA phylogenetic markers in the angiosperm genus *Leucadendron* (Proteaceae): a next-generation


**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.
Figures and tables

Figure 1
Workflow of the probe design script Sondovač. An overview on the main steps of Hyb-Seq are 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 gray, a summary of each step is given in medium gray, and input/output of each step is depicted in light gray. 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.

Figure 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 colored in light gray (respectively green in the on-line color figure). Nodes that are incongruent between the contrasting phylogenetic trees of the different datasets are marked with arrows.
UNIX commands

Length-filtered probe sequences

Sequence analysis

De novo assembly of the transcript or genome skim BLAT hits [depending on the selection in (1) in the previous step] to larger contigs

Quality-filtered sequences of matching transcripts and genome skim reads

Retention of those contigs that comprise exons ≥ bait length (in the Oxalis case 120 bp) and have a certain total locus length (in the Oxalis case ≥ 600 bp)

Retention of those contigs that comprise exons ≥ bait length (in the Oxalis case 120 bp) and have a certain total locus length (in the Oxalis case ≥ 600 bp)

Retention of probe sequences sharing ≥ 90% sequence similarity

Final probe sequences

Removal of probe sequences sharing ≥ 90% sequence similarity with the plastome reference

Probe design

Bait synthesis, based on probe sequences

Hybridization between baits and genomic libraries

Sequencing

Data analysis

Paired-end genome skim reads

Plastome reference

Mitochondriome reference

Transcripts

Paired-end genome skimreads without plastid reads

Paired-end genome skim reads

Bowtie 2, SAMtools, bam2fastq

Removal of reads of plastid origin

Flash

Combination of paired-end reads

BLAT, UNIX commands

Removal of transcripts sharing ≥ 90% sequence similarity

Unique transcripts

Sequences of matching transcripts and genome skim reads

BLAT

Matching of the unique transcripts and the filtered, combined genome skim reads sharing ≥ 85% sequence similarity

BLAT, UNIX commands

Retention of those contigs that comprise exons ≥ bait length (in the Oxalis case 120 bp) and have a certain total locus length (in the Oxalis case ≥ 600 bp)

Final probe sequences

Paired-end genome skim reads without plastid and mitochondrial reads

Removal of reads of plastid origin

Removal of reads of mitochondrial origin

Paired-end genome skim reads without plastid and mitochondrial reads

Combined genome skim reads without plastid (and mitochondrial) reads
# Supporting information

## Table S1

Voucher information of southern African *Oxalis* accessions used in this study. Herbarium specimens are either deposited in the Herbarium Collections at Charles University in Prague (PRC) or the Stellenbosch University Herbarium (STEU). Ploidy level estimates are unpublished data of JS.

<table>
<thead>
<tr>
<th>Species</th>
<th>Herbarium/ Voucher no.</th>
<th>Locality</th>
<th>Latitude/ Longitude</th>
<th>Collector/ Date</th>
<th>Ploidy</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Oxalis amblyosepala</em> Schltr.</td>
<td>PRC J11-969</td>
<td>Vanrhynsdorp, Western Cape, South Africa</td>
<td>S31° 46' 52.5'' E18° 46' 02.7''</td>
<td>J. Suda &amp; R. Sudová 24.08.2011</td>
<td>2x</td>
</tr>
<tr>
<td><em>Oxalis blastorrhiza</em> T.M. Salter</td>
<td>PRC J557</td>
<td>Vanrhynsdorp, Western Cape, South Africa</td>
<td>S31° 36' 07.9'' E18° 43' 47.9''</td>
<td>J. Suda &amp; R. Sudová 31.05.2010</td>
<td>2x</td>
</tr>
<tr>
<td><em>Oxalis callosa</em> R. Knuth</td>
<td>PRC J11-726</td>
<td>Calvinia, Northern Cape, South Africa</td>
<td>S31° 58' 22.4'' E20° 08' 50.3''</td>
<td>J. Suda &amp; R. Sudová 19.08.2011</td>
<td>6x</td>
</tr>
<tr>
<td><em>Oxalis ciliaris</em> Jacq.</td>
<td>PRC J233</td>
<td>Swellendam, Western Cape, South Africa</td>
<td>S34° 01' 28.3'' E20° 25' 40.3''</td>
<td>J. Suda &amp; P. Trávniček 01.09.2009</td>
<td>4x</td>
</tr>
<tr>
<td><em>Oxalis creaseyii</em> T.M. Salter</td>
<td>PRC J319</td>
<td>Garies, Northern Cape, South Africa</td>
<td>S30° 28' 09.9'' E18° 08' 08.2''</td>
<td>J. Suda &amp; R. Sudová 24.08.2011</td>
<td>2x</td>
</tr>
<tr>
<td><em>Oxalis exserta</em> T.M. Salter</td>
<td>PRC J616</td>
<td>Springbok, Northern Cape, South Africa</td>
<td>S29° 26' 14.4'' E17° 57' 10.1''</td>
<td>J. Suda &amp; R. Sudová 08.06.2010</td>
<td>8x</td>
</tr>
<tr>
<td><em>Oxalis gracilis</em> Jacq.</td>
<td>PRC J558</td>
<td>Vanrhynsdorp, Western Cape, South Africa</td>
<td>S31° 43' 04.4'' E18° 46' 07.6''</td>
<td>J. Suda &amp; R. Sudová 31.05.2010</td>
<td>2x</td>
</tr>
<tr>
<td><em>Oxalis helicoides</em> T.M. Salter</td>
<td>PRC J319</td>
<td>Springbok, Northern Cape, South Africa</td>
<td>S29° 46' 22.9'' E17° 49' 15.5''</td>
<td>J. Suda &amp; P. Trávniček 06.09.2009</td>
<td>2x</td>
</tr>
<tr>
<td><em>Oxalis hirsuta</em> Sond.</td>
<td>PRC J12-20</td>
<td>Williston, Northern Cape, South Africa</td>
<td>S31° 22' 59.6'' E20° 35' 24.6''</td>
<td>J. Suda &amp; R. Sudová 14.04.2012</td>
<td>2x</td>
</tr>
<tr>
<td><em>Oxalis hirta</em> L.</td>
<td>PRC J57</td>
<td>Yzerfontein, Western Cape, South Africa</td>
<td>S33° 19' 42.1'' E18° 11' 15.9''</td>
<td>J. Suda &amp; R. Sudová 29.08.2007</td>
<td>4x</td>
</tr>
<tr>
<td><em>Oxalis hirta</em> L.</td>
<td>PRC J62</td>
<td>Yzerfontein, Western Cape, South Africa</td>
<td>S33° 08' 10.2'' E17° 59' 54.5''</td>
<td>J. Suda &amp; R. Sudová 30.08.2007</td>
<td>4x</td>
</tr>
<tr>
<td><em>Oxalis imbricata</em> Eckl. &amp; Zeyh.</td>
<td>STEU MO472</td>
<td>Worcester, Western Cape, South Africa</td>
<td>S33°53.624' E19°26.284'</td>
<td>K. Oberlander 05.05.2004</td>
<td>2x</td>
</tr>
<tr>
<td><em>Oxalis inconspicua</em> T.M. Salter</td>
<td>PRC J595</td>
<td>Garies, Northern Cape, South Africa</td>
<td>S30° 32' 03.9&quot; E18° 08' 14.9&quot;</td>
<td>J. Suda &amp; R. Sudová 02.06.2010</td>
<td>2x</td>
</tr>
<tr>
<td><em>Oxalis kamiesbergensis</em> T.M. Salter</td>
<td>PRC J153</td>
<td>Kamieskroon, Northern Cape, South Africa</td>
<td>S30° 10' 43.6&quot; E18° 00' 45.0&quot;</td>
<td>J. Suda &amp; R. Sudová 14.09.2007</td>
<td>2x</td>
</tr>
<tr>
<td><em>Oxalis obtusa</em> Jacq.</td>
<td>PRC J12</td>
<td>Kommetjie, Western Cape, South Africa</td>
<td>S34° 08' 00.5&quot; E18° 20' 13.6&quot;</td>
<td>J. Suda &amp; R. Sudová 26.08.2007</td>
<td>2x</td>
</tr>
<tr>
<td><em>Oxalis obtusa</em> Jacq.</td>
<td>PRC J99</td>
<td>Calvinia, Northern Cape, South Africa</td>
<td>S31° 19' 51.1&quot; E19° 52' 09.4&quot;</td>
<td>J. Suda &amp; R. Sudová 02.09.2007</td>
<td>2x</td>
</tr>
<tr>
<td><em>Oxalis orthopoda</em> T.M. Salter</td>
<td>STEU MO351</td>
<td>Mossel Bay, Western Cape, South Africa</td>
<td>S34°10.723' E21°56.650'</td>
<td>K. Oberlander 24.04.2003</td>
<td>4x</td>
</tr>
<tr>
<td><em>Oxalis palmifrons</em> T.M. Salter</td>
<td>STEU MO403</td>
<td>NA</td>
<td>NA</td>
<td>Ex Hort./ NA</td>
<td>2x</td>
</tr>
<tr>
<td><em>Oxalis polyphylla</em> Jacq.</td>
<td>PRC J11-44</td>
<td>De Hoop, Western Cape, South Africa</td>
<td>S34° 25' 51.7&quot; E20° 25' 06.5&quot;</td>
<td>J. Suda &amp; J. Rauchová 21.05.2011</td>
<td>2x</td>
</tr>
<tr>
<td><em>Oxalis primuloideas</em> R. Knuth</td>
<td>PRC J136</td>
<td>Calvinia, Northern Cape, South Africa</td>
<td>S31° 21' 03.1&quot; E19° 31' 29.0&quot;</td>
<td>J. Suda &amp; R. Sudová 12.09.2007</td>
<td>2x</td>
</tr>
</tbody>
</table>
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. The table is complemented by the genome skim accession O. obtusa J12 and the transcriptome accession O. corniculata JHCN, which did not form part of the calculation of the means.

<table>
<thead>
<tr>
<th>Species</th>
<th>Accession number</th>
<th># quality-filtered reads after duplicate removal</th>
<th># LCN exons assembled from total 4,926</th>
<th># LCN genes assembled from total 1,164</th>
<th>Mean sequence divergence from LCN probes</th>
<th>Completion of LCN matrix</th>
<th>Completion of plastome</th>
<th>Completion of nrDNA cistron</th>
<th>Mean sequencing depth of LCN loci</th>
<th>Mean sequencing depth of plastome</th>
<th>Mean sequencing depth of nrDNA cistron</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oxalis amblyosepala</td>
<td>J11-869</td>
<td>497,402</td>
<td>4,246 (86%)</td>
<td>1,161 (100%)</td>
<td>4%</td>
<td>94%</td>
<td>98%</td>
<td>100%</td>
<td>13</td>
<td>166</td>
<td>209</td>
</tr>
<tr>
<td>Oxalis bulbosa</td>
<td>J557</td>
<td>715,360</td>
<td>4,339 (88%)</td>
<td>1,160 (100%)</td>
<td>4%</td>
<td>95%</td>
<td>99%</td>
<td>100%</td>
<td>16</td>
<td>220</td>
<td>427</td>
</tr>
<tr>
<td>Oxalis caffosa</td>
<td>J11-726</td>
<td>788,393</td>
<td>4,490 (91%)</td>
<td>1,160 (100%)</td>
<td>4%</td>
<td>97%</td>
<td>98%</td>
<td>100%</td>
<td>17</td>
<td>139</td>
<td>405</td>
</tr>
<tr>
<td>Oxalis ciliaris</td>
<td>J233</td>
<td>542,186</td>
<td>4,345 (88%)</td>
<td>1,160 (100%)</td>
<td>4%</td>
<td>95%</td>
<td>99%</td>
<td>100%</td>
<td>13</td>
<td>99</td>
<td>223</td>
</tr>
<tr>
<td>Oxalis creaseyi</td>
<td>J11-961</td>
<td>703,091</td>
<td>4,454 (90%)</td>
<td>1,162 (100%)</td>
<td>4%</td>
<td>96%</td>
<td>99%</td>
<td>99%</td>
<td>15</td>
<td>155</td>
<td>293</td>
</tr>
<tr>
<td>Oxalis exserta</td>
<td>J916</td>
<td>496,193</td>
<td>4,170 (85%)</td>
<td>1,160 (100%)</td>
<td>4%</td>
<td>92%</td>
<td>97%</td>
<td>100%</td>
<td>12</td>
<td>103</td>
<td>225</td>
</tr>
<tr>
<td>Oxalis gracilis</td>
<td>J508</td>
<td>749,774</td>
<td>4,432 (90%)</td>
<td>1,162 (100%)</td>
<td>4%</td>
<td>96%</td>
<td>94%</td>
<td>100%</td>
<td>17</td>
<td>118</td>
<td>418</td>
</tr>
<tr>
<td>Oxalis helicoides</td>
<td>J319</td>
<td>798,432</td>
<td>4,450 (90%)</td>
<td>1,162 (100%)</td>
<td>4%</td>
<td>97%</td>
<td>99%</td>
<td>100%</td>
<td>16</td>
<td>184</td>
<td>379</td>
</tr>
<tr>
<td>Oxalis hirsuta</td>
<td>J12-29</td>
<td>933,290</td>
<td>4,514 (92%)</td>
<td>1,163 (100%)</td>
<td>3%</td>
<td>98%</td>
<td>99%</td>
<td>100%</td>
<td>21</td>
<td>393</td>
<td>306</td>
</tr>
<tr>
<td>Oxalis hirta</td>
<td>J57</td>
<td>472,230</td>
<td>4,137 (84%)</td>
<td>1,162 (100%)</td>
<td>4%</td>
<td>91%</td>
<td>98%</td>
<td>93%</td>
<td>11</td>
<td>111</td>
<td>273</td>
</tr>
<tr>
<td>Oxalis hirta</td>
<td>J62</td>
<td>156,655</td>
<td>2,611 (53%)</td>
<td>1,093 (94%)</td>
<td>4%</td>
<td>63%</td>
<td>79%</td>
<td>98%</td>
<td>8</td>
<td>55</td>
<td>69</td>
</tr>
<tr>
<td>Oxalis imbricata</td>
<td>MO472</td>
<td>3,280,494</td>
<td>4,520 (92%)</td>
<td>1,163 (100%)</td>
<td>4%</td>
<td>99%</td>
<td>98%</td>
<td>83%</td>
<td>45</td>
<td>361</td>
<td>1,583</td>
</tr>
<tr>
<td>Oxalis inopinata</td>
<td>J955</td>
<td>655,265</td>
<td>4,408 (89%)</td>
<td>1,161 (100%)</td>
<td>4%</td>
<td>95%</td>
<td>99%</td>
<td>98%</td>
<td>16</td>
<td>217</td>
<td>172</td>
</tr>
<tr>
<td>Oxalis kamesbergenensis</td>
<td>J153</td>
<td>401,417</td>
<td>3,913 (79%)</td>
<td>1,158 (99%)</td>
<td>4%</td>
<td>88%</td>
<td>99%</td>
<td>79%</td>
<td>12</td>
<td>138</td>
<td>228</td>
</tr>
<tr>
<td>Oxalis obtusa</td>
<td>J99</td>
<td>534,889</td>
<td>4,287 (87%)</td>
<td>1,163 (100%)</td>
<td>3%</td>
<td>95%</td>
<td>99%</td>
<td>81%</td>
<td>14</td>
<td>166</td>
<td>233</td>
</tr>
<tr>
<td>Oxalis orthopoda</td>
<td>MO351</td>
<td>939,291</td>
<td>4,516 (92%)</td>
<td>1,163 (100%)</td>
<td>4%</td>
<td>98%</td>
<td>96%</td>
<td>100%</td>
<td>21</td>
<td>168</td>
<td>660</td>
</tr>
<tr>
<td>Oxalis palmarum</td>
<td>MO403</td>
<td>168,543</td>
<td>1,253 (25%)</td>
<td>753 (65%)</td>
<td>4%</td>
<td>22%</td>
<td>64%</td>
<td>100%</td>
<td>16</td>
<td>83</td>
<td>58</td>
</tr>
<tr>
<td>Oxalis polypylla</td>
<td>J11-44</td>
<td>599,495</td>
<td>4,298 (87%)</td>
<td>1,159 (100%)</td>
<td>4%</td>
<td>94%</td>
<td>99%</td>
<td>100%</td>
<td>16</td>
<td>234</td>
<td>190</td>
</tr>
<tr>
<td>Oxalis primuloides</td>
<td>J136</td>
<td>500,786</td>
<td>4,093 (83%)</td>
<td>1,161 (100%)</td>
<td>4%</td>
<td>91%</td>
<td>99%</td>
<td>100%</td>
<td>13</td>
<td>153</td>
<td>317</td>
</tr>
<tr>
<td>Oxalis pulchella</td>
<td>J11-875</td>
<td>549,374</td>
<td>4,225 (86%)</td>
<td>1,162 (100%)</td>
<td>4%</td>
<td>93%</td>
<td>99%</td>
<td>100%</td>
<td>15</td>
<td>226</td>
<td>248</td>
</tr>
<tr>
<td>Oxalis reclinata</td>
<td>J343</td>
<td>375,386</td>
<td>4,003 (81%)</td>
<td>1,158 (99%)</td>
<td>4%</td>
<td>90%</td>
<td>99%</td>
<td>100%</td>
<td>11</td>
<td>111</td>
<td>158</td>
</tr>
<tr>
<td>Oxalis smithiana</td>
<td>J11-511</td>
<td>1,074,130</td>
<td>4,451 (90%)</td>
<td>1,162 (100%)</td>
<td>3%</td>
<td>97%</td>
<td>96%</td>
<td>100%</td>
<td>26</td>
<td>299</td>
<td>981</td>
</tr>
<tr>
<td>Oxalis tenella</td>
<td>J311</td>
<td>511,049</td>
<td>4,343 (88%)</td>
<td>1,160 (100%)</td>
<td>4%</td>
<td>94%</td>
<td>99%</td>
<td>100%</td>
<td>14</td>
<td>137</td>
<td>246</td>
</tr>
<tr>
<td>Oxalis truncatula</td>
<td>J230</td>
<td>627,899</td>
<td>4,488 (91%)</td>
<td>1,164 (100%)</td>
<td>3%</td>
<td>98%</td>
<td>98%</td>
<td>79%</td>
<td>15</td>
<td>126</td>
<td>231</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td>711,293</td>
<td>4,124 (84%)</td>
<td>1,141 (98%)</td>
<td>4%</td>
<td>90%</td>
<td>96%</td>
<td>96%</td>
<td>16</td>
<td>173</td>
<td>290</td>
</tr>
<tr>
<td>Oxalis obtusa</td>
<td>J12</td>
<td>7,938,349</td>
<td>3,972 (81%)</td>
<td>1,160 (100%)</td>
<td>1%</td>
<td>88%</td>
<td>99%</td>
<td>100%</td>
<td>11</td>
<td>825</td>
<td>4,490</td>
</tr>
<tr>
<td>Oxalis corniculata</td>
<td>JHCN</td>
<td>NA</td>
<td>NA</td>
<td>1,163 (100%)</td>
<td>11%</td>
<td>93%</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>
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*. In cases where the sum of on-target, plastid and nrDNA reads after assembly exceeded the initial, total number of quality-filtered reads after duplicate removal, which was obviously due to read mapping across datasets, the three read sources were marked with an asterisk.

The table is complemented by the genome skim accession *O. obtusa* J12, which did not form part of the calculation of the means. The effect of target enrichment is evident from a comparison of the proportion of reads mapped on-target between the genome skim accession J12 and the Hyb-Seq accession J99 of *O. obtusa*: Instead of on average 2% genome skim reads 59% Hyb-Seq reads mapped on-target.

<table>
<thead>
<tr>
<th>Species</th>
<th>Accession number</th>
<th># raw reads</th>
<th># quality-filtered reads</th>
<th># quality-filtered reads after duplicate removal (as % of quality-filtered reads)</th>
<th># on-target reads (as % of quality-filtered reads after duplicate removal)</th>
<th># plastid reads (as % of quality-filtered reads after duplicate removal)</th>
<th># nrDNA reads (as % of quality-filtered reads after duplicate removal)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Oxalis amblyosepala</em></td>
<td>J11-969</td>
<td>1,220,088</td>
<td>1,101,853</td>
<td>497,402 (45%)</td>
<td>313,410 (63%)</td>
<td>162,330 (33%)</td>
<td>11,079 (2%)</td>
</tr>
<tr>
<td><em>Oxalis blastorrhiza</em></td>
<td>J557</td>
<td>1,895,744</td>
<td>1,724,917</td>
<td>715,360 (41%)</td>
<td>402,111 (56%)</td>
<td>222,377 (31%)</td>
<td>26,254 (4%)</td>
</tr>
<tr>
<td><em>Oxalis callosa</em></td>
<td>J11-726</td>
<td>1,451,514</td>
<td>1,311,491</td>
<td>788,393 (60%)</td>
<td>472,837 (60%)</td>
<td>142,150 (18%)</td>
<td>20,045 (3%)</td>
</tr>
<tr>
<td><em>Oxalis ciliaris</em></td>
<td>J233</td>
<td>1,012,810</td>
<td>931,389</td>
<td>542,186 (58%)</td>
<td>317,293 (59%)</td>
<td>96,490 (18%)</td>
<td>10,788 (2%)</td>
</tr>
<tr>
<td><em>Oxalis creaseyi</em></td>
<td>J11-961</td>
<td>1,543,348</td>
<td>1,423,598</td>
<td>703,091 (49%)</td>
<td>424,600 (60%)</td>
<td>154,369 (22%)</td>
<td>16,048 (2%)</td>
</tr>
<tr>
<td><em>Oxalis exserta</em></td>
<td>J616</td>
<td>935,348</td>
<td>841,451</td>
<td>496,193 (59%)</td>
<td>268,508 (54%)</td>
<td>97,016 (20%)</td>
<td>13,602 (3%)</td>
</tr>
<tr>
<td><em>Oxalis gracilis</em></td>
<td>J558</td>
<td>1,296,462</td>
<td>1,188,885</td>
<td>749,774 (63%)</td>
<td>447,138 (60%)</td>
<td>109,747 (15%)</td>
<td>28,477 (4%)</td>
</tr>
<tr>
<td><em>Oxalis helicoides</em></td>
<td>J319</td>
<td>1,838,158</td>
<td>1,704,468</td>
<td>798,432 (47%)</td>
<td>458,664 (57%)</td>
<td>186,043 (23%)</td>
<td>19,722 (2%)</td>
</tr>
<tr>
<td><em>Oxalis hirsuta</em></td>
<td>J12-20</td>
<td>3,207,132</td>
<td>2,876,738</td>
<td>933,290 (32%)</td>
<td>625,456 (67%)*</td>
<td>407,215 (44%)*</td>
<td>20,964 (2%)*</td>
</tr>
<tr>
<td><em>Oxalis hirta</em></td>
<td>J57</td>
<td>1,039,590</td>
<td>942,496</td>
<td>472,230 (50%)</td>
<td>236,745 (50%)</td>
<td>111,585 (24%)</td>
<td>12,304 (3%)</td>
</tr>
<tr>
<td><em>Oxalis hirta</em></td>
<td>J62</td>
<td>276,908</td>
<td>246,727</td>
<td>156,655 (63%)</td>
<td>77,330 (49%)</td>
<td>39,072 (25%)</td>
<td>3,420 (2%)</td>
</tr>
<tr>
<td><em>Oxalis imbricata</em></td>
<td>MO472</td>
<td>6,867,908</td>
<td>6,243,085</td>
<td>3,280,494 (53%)</td>
<td>1,775,687 (54%)</td>
<td>372,088 (11%)</td>
<td>69,811 (2%)</td>
</tr>
<tr>
<td><em>Oxalis inconspicua</em></td>
<td>J595</td>
<td>1,773,838</td>
<td>1,635,614</td>
<td>655,265 (40%)</td>
<td>419,477 (64%)*</td>
<td>224,056 (34%)*</td>
<td>18,540 (3%)*</td>
</tr>
<tr>
<td><em>Oxalis kamiesbergensis</em></td>
<td>J153</td>
<td>1,029,982</td>
<td>927,395</td>
<td>401,417 (43%)</td>
<td>232,487 (58%)</td>
<td>134,519 (34%)</td>
<td>9,216 (2%)</td>
</tr>
<tr>
<td><em>Oxalis obtusa</em></td>
<td>J99</td>
<td>1,378,380</td>
<td>1,260,241</td>
<td>534,889 (42%)</td>
<td>331,024 (62%)</td>
<td>162,427 (30%)</td>
<td>9,398 (2%)</td>
</tr>
<tr>
<td><em>Oxalis orthopoda</em></td>
<td>MO351</td>
<td>1,849,216</td>
<td>1,711,627</td>
<td>939,291 (55%)</td>
<td>620,990 (66%)</td>
<td>167,454 (18%)</td>
<td>87,142 (9%)</td>
</tr>
<tr>
<td><em>Oxalis palmifrons</em></td>
<td>MO403</td>
<td>507,374</td>
<td>395,164</td>
<td>168,543 (43%)</td>
<td>66,170 (39%)</td>
<td>64,900 (39%)</td>
<td>3,728 (2%)</td>
</tr>
<tr>
<td><em>Oxalis polyphylla</em></td>
<td>J11-44</td>
<td>1,806,672</td>
<td>1,631,230</td>
<td>599,495 (37%)</td>
<td>383,590 (64%)*</td>
<td>235,690 (39%)*</td>
<td>11,903 (2%)*</td>
</tr>
<tr>
<td><em>Oxalis primuloides</em></td>
<td>J136</td>
<td>1,197,900</td>
<td>1,100,614</td>
<td>500,786 (46%)</td>
<td>264,869 (53%)</td>
<td>147,514 (29%)</td>
<td>19,336 (4%)</td>
</tr>
<tr>
<td><em>Oxalis pulchella</em></td>
<td>J11-875</td>
<td>1,719,890</td>
<td>1,540,537</td>
<td>549,374 (36%)</td>
<td>339,571 (62%)*</td>
<td>237,816 (43%)*</td>
<td>17,299 (3%)*</td>
</tr>
<tr>
<td><em>Oxalis reclinata</em></td>
<td>J343</td>
<td>826,650</td>
<td>742,405</td>
<td>375,386 (51%)</td>
<td>224,180 (60%)</td>
<td>109,011 (29%)</td>
<td>7,297 (2%)</td>
</tr>
</tbody>
</table>
**Fig. S1**
Map of the draft plastid genome of *Oxalis hirsuta*. The thin black lines of the inner circle indicate the location of the large single copy (LSC) and small single copy (SSC) regions, the thick black lines those of the inverted repeats (IR). Transcription is clockwise for genes on the outside of the circle and counterclockwise for genes on the inside of the circle. Genes with multiple exons are marked with an asterisk. Quasi full-length (lacking start and stop codon) and partial genes are marked with an arrow.

**Fig. S2**
Splits network of the LCN gene matrix, the plastome and the nrDNA cistron datasets.
(A) 727 concatenated LCN loci (least squares fit = 99.5%), (B) plastid genome (least squares fit = 98.0%), (C) nrDNA cistron (least squares fit = 96.7%). For better visualization only bootstrap support values above 90% are shown.

**Fig. S3**
Convergence measures of Bayesian MCMC runs. (A) Average effective sample size (avgESS) and (B) potential scale reduction factor (PSRF). Parameters are TL: total tree length; r(\(A\langle -\rangle C\))−r(\(G\langle -\rangle T\)): reversible substitution rates; k_revmat: substitution rates of the GTR model; pi(\(A\))−pi(\(T\)): stationary state frequencies; alpha: shape of gamma distribution of rate variation across sites. Convergence is indicated by avgESS >100 and PSRF ~1.0. (C) Average standard deviation of split frequencies (ASDSF), shown for each of the 727 LCN genes. Convergence is indicated by ASDSF <0.05.

**Fig. S4**
Comparison of phylogenetic trees based on 727 LCN genes. Species trees were reconstructed under the assumption of the multispecies coalescent model implemented in ASTRAL (A), MP-EST (B), and STAR (C), as MRP supertree (D) and concatenated supermatrix (E). Except for concatenation both Bayesian inference (BI) and maximum likelihood (ML) were used for building the species tree. Dashed lines connect each accession to its placement in the contrasting tree. Values close to the nodes are bootstrap support (* = 100%). The Hirta clade (Oberlander et al., 2011) is colored in green. Nodes which are incongruent between BI of gene trees and ML gene trees within the same species tree method are marked with a blue arrow. Nodes which are incongruent between the different species tree methods and concatenation are marked with orange arrows on the concatenated tree.

**Fig. S5**
Principal coordinate analysis of Robinson-Foulds (RF) distances between gene trees. (A) RF distances between gene trees obtained with Bayesian inference (BI), (B) RF distances between maximum likelihood (ML) gene trees. Eigenvalues were highest for the first three axes, and the first two are displayed. Each dot represents a gene tree. Kernel density is indicated with blue lines.

**Fig. S6**
Robinson-Foulds (RF) distances between gene trees and the species tree. (A) RF
distances between gene trees obtained with Bayesian inference (BI) and the STAR
species tree, (B) RF distances between maximum likelihood (ML) gene trees and the
STAR species tree.
Oxalis hirsuta draft plastome

153,678 bp
A: BI of gene trees
B: ML gene trees
BI of gene trees

ML gene trees