Retrieval of hundreds of nuclear loci from herbarium specimens

METHODS AND TECHNIQUES

Retrieval of hundreds of nuclear loci from herbarium specimens

Michelle L. Hart,[1] Laura L. Forrest,[1] James A. Nicholls[1,2] & Catherine A. Kidner[1,3]

1 *Royal Botanic Garden Edinburgh, 20A Inverleith Row, Edinburgh EH3 5LR, U.K.*
2 *Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3FL, U.K.*
3 *Institute of Molecular Plant Sciences, University of Edinburgh, Edinburgh EH9 3BF, U.K.*
Author for correspondence: *Catherine Kidner, c.kidner@rbge.ac.uk*

ORCID MLH, http://orcid.org/0000-0001-9503-7786; LLF, http://orcid.org/0000-0002-0235-9506; JAN, http://orcid.org/0000-0002-9325-563X; CAK, http://orcid.org/0000-0001-6426-3000

DOI http://dx.doi.org/10.12705/655.9

Abstract Herbaria are unparalleled collections of biodiversity information representing the world's flora. However, this treasure has remained largely inaccessible to genetic studies, frequently limited by the low yields of poor-quality DNA. Next-generation sequencing (NGS) has transformed every field of biological research. The different strategies for accessing genetic data using NGS are changing the direction of biodiversity research—we are no longer constrained by a relatively small number of markers for non-model organisms, by time and cost limited sample sizes, or by incomplete datasets due to recalcitrant DNA extractions or PCR amplification failure. Here we show that targeted enrichment through hybrid capture can be used to generate hundreds of kilobases of nuclear sequence data of the Neotropical genus *Inga*, from herbarium specimens as old as 180 years and using as little as 16 ng of degraded DNA.

Keywords degraded DNA; herbarium specimen; hybrid baits; hybrid capture; *Inga*

Supplementary Material Electronic Supplement (Figs. S1–S3; Table S1) and phylogenies for individual bait loci are available in the Supplementary Data section of the online version of this article at http://www.ingentaconnect.com/content/iapt/tax

■ INTRODUCTION

Herbaria have been likened to "treasure chests" or "Genomic Treasure Troves" (Särkinen & al., 2012; Staats & al., 2013). They contain a wealth of information about plant diversity and distribution, including many type specimens, some of which are the only recorded accession of their species. Assessing the genetic component of this diversity is difficult due to the degraded nature of DNA in specimens after many years of storage, yet accessing the genetic information in herbaria and museums would hugely expand the utility of such collections (Rowe & al., 2011; Bi & al., 2013; Nachman, 2013; Staats & al., 2013; Jones & Good, 2015). Several papers describe optimization of DNA extraction from herbarium specimens for use in PCR-based methods such as Sanger sequencing (e.g., Drábková & al., 2002; Telle & Thines, 2008; Särkinen & al., 2012), but such PCR methods are typically limited to short sequences present in high copy number, which do not give phylogenetic resolution for many clades.

Next-generation sequencing (NGS) has the potential to open up these collections to the genomic era. The major stumbling blocks associated with working with low quantities of degraded DNA extracted from herbarium specimens of non-model plant species can be overcome or reduced by using NGS techniques. Many NGS approaches employ library preparation that requires shearing genomic DNA into short fragments (40–500 bp). DNA library construction can be tailored to work with low DNA starting quantities (Rowe & al., 2011). Different strategies of genome partitioning or targeted sequencing can be employed to overcome or assess issues of contaminant DNA (Rowe & al., 2011; Enk & al., 2014; Linderholm, 2016) and interrogation of the resulting data can be used to assess levels of variability due to post mortem DNA damage (Rowe & al., 2011; Bi & al., 2013; Staats & al., 2013). Via experimental design, key elements such as target specificity and mean depth of coverage can be managed, thus allowing the researcher to tailor the technique to their biological question (Grover & al., 2012).

The application of NGS to herbarium and museum specimens has typically been restricted to reconstruction of plastid or mitochondrial genomes present at high copy number, or low-coverage genome skims to re-sequence species with existing reference genomes (Rowe & al., 2011; Staats & al., 2013). We focus here on the potential of targeted enrichment to obtain sequence data for hundreds of nuclear genes from herbarium material. Targeted enrichment is an example of a genome partitioning approach developed to allow the sequencing of a selected subset of the genome (Cronn & al., 2012; Jones & Good, 2015). Other methods of genome partitioning NGS include PCR based enrichment (multiplex PCR), restriction site associated DNA sequencing (RAD-Seq) and whole-transcriptome shotgun sequencing (RNA-seq; see Wang & al., 2009; Davey & Blaxter, 2010; Cronn & al., 2012; Jones & Good, 2015). Targeted enrichment was developed as a more cost effective and

Received: 4 Feb 2016 | returned for (first) revision: 18 Mar 2016 | (last) revision received: 19 May 2016 | accepted: 19 May 2019 || publication date(s): online fast track, 12 Oct 2016; in print and online issues, 27 Oct 2016

high-throughput alternative to whole genome sequencing and multiplex PCR (Gnirke & al., 2009; Mamanova & al., 2010) and was first applied to studies of the human genome. The targeted enrichment, or "hybrid capture", technique used in this study uses a hybridisation reaction involving custom-designed short RNA probes ("baits") in solution, to capture hundreds of target loci from fragmented genomic DNA libraries.

The advantages of moderate cost, low input amounts of genomic DNA and the ability to target large numbers of informative markers make targeted enrichment highly applicable to phylogenomic and population genomic studies in non-model organisms (Lemmon & al., 2012; McCormack & al., 2013). The scale of targeted enrichment can range from several targeted loci to over a million targeted regions (Grover & al., 2012) and has been applied to intraspecific population studies (Zhou & Holliday, 2012), resolving intra- and interfamily phylogenetic relationships (Sass & al., 2016), species-level phylogenetics (Uribe-Convers & al., 2016) and recent radiations (Nicholls & al., 2015).

This approach has been used successfully for museum specimens by Bi & al. (2013), who used targeted exon capture to enrich for ca. 4 MB of DNA for single-nucleotide polymorphism (SNP) analysis, obtained from skins of alpine chipmunks (*Tamias alpinus*) collected in 1915. We show here that a similar approach works very well for plant herbarium specimens; by using hybrid capture, depth of coverage and assembly issues are solved, resulting in retrieval of hundreds of kilobases of conservatively called nuclear sequences from specimens as old as 180 years, and from as little as 16 ng of degraded starting DNA.

We used the genus *Inga* Mill. to develop protocols for targeted enrichment of DNA from herbarium samples as this method has been employed successfully on silica gel dried *Inga* specimens using an existing set of capture baits (Nicholls & al., 2015). This legume bait set comprised 276 loci designed from *Inga* transcriptomes, and a further 1124 loci designed from other taxa across the Mimosoideae. *Inga umbellifera* (Vahl) Steud. ex D.C., a widespread species, was selected as our focal taxon because population-level targeted enrichment sequence data already exist (from 20 libraries representing 19 accessions; Nicholls & al., 2015) to provide the context in which our data could be analysed. As part of developing targeted enrichment methods for herbarium material, we explore the impacts of a number of different parameters on hybrid capture and sequencing success including: method of initial genomic DNA extraction, DNA quality and yield, DNA repair before library preparation, and different DNA library size selection strategies, designed to deal with the degraded DNA typically obtained from herbarium material. We also assess the use of two commercially available library preparation kits.

■ MATERIALS AND METHODS

Taxon sampling. — We sampled from six herbarium sheets of *Inga umbellifera*. These included isotypes of *Inga lawranceana* Britton & Killip (*Lawrance 260* (E), 1932) and *Inga sciadion* Steud. (*Hostmann 170* (K), 1841), species synonymised under *I. umbellifera* in Pennington's 1997 monograph of the genus and in the International Legume Database & Information Service (http://www.ildis.org/; accessed 6 Jan 2016).

The samples ranged from 6 to 180 years old, and were collected in Peru, Suriname, French Guiana, Guyana and Colombia (Table 1). Silica gel preserved material from the same accession as the herbarium specimen *16L 145* (E) was included for the 2009 collection, to determine whether there was a notable difference in quality between the two types of tissue preservation. In total, we generated 13 DNA extractions from eight different starting materials (Table 2). Data from a silica gel preserved sample for one of the herbarium collections (*K. Dexter 401* (E), 2004) were included in a previous study (Nicholls & al., 2015), so it was not sequenced here.

DNA extraction and repair. — Approximately 2 cm² sections of leaf or 10–20 mg of flower parts (Electr. Suppl.: Fig. S1; Table 2) were used for each DNA extraction. Multiple extractions (2–7) were made from each accession from the same starting material, to allow for subsequent pooling and concentration of low yield DNA as well as to provide experimental duplicates. Plant tissue was ground in 2 ml tubes, using a TissueLyser II with flattened tungsten beads. DNA was extracted using the manufacturer's protocol for Qiagen DNeasy Plant mini-kits up to the DNA binding stage, at which point several DNA extractions from the same starting material were combined by eluting them through single DNeasy Mini spin columns or Qiagen QiaQuick PCR purification columns to concentrate the DNA (see Table 2). The two types of columns were used to test the impact of recovering degraded DNA using kits designed either for whole genomic DNA (DNeasy mini kit) or smaller fragments (QiaQuick columns; elution of DNA

Table 1. Herbarium voucher details for *Inga umbellifera* samples used for DNA extractions.

Year	Voucher	Herbarium barcode	Country	Locality
1835	*Matthews 1593* (E)	E00705600	Peru	Departamento San Martín: Provincia San Martín, Tarapato
1841	*Hostmann 170* (K)	unbarcoded	Suriname	*In sylvia humidis*
1932	*Lawrance 260* (E)	E00326264	Colombia	NW Chapor, Boyaca, 100 m NW of Bogata
1948	*For. Dept. Brit. Guiana 5682* (K)	unbarcoded	Guyana	
2004	*K. Dexter 401* (E)	E00757815	Peru	Madre de Dios, Los Amigos Biological Station, floodplain
2009	*K. Dexter 16L 145* (E)	E00757816	French Guiana	Nouragues Ecological Research Station, Inselberg, Grand Plateau

fragments between 100 and 10,000 bases). Yield and integrity (size distributions) of genomic DNA extracts were quantified by Qubit (Invitrogen, Carlsbad, California, U.S.A.) using the dsDNA HS kit, and TapeStation (Agilent, Edinburgh, Scotland, U.K.) Genomic DNA ScreenTape, respectively. The TapeStation provides a DNA integrity number (DIN), which is a metric of DNA integrity on a scale of 1 (strongly degraded) to 10 (highly intact; Gassmann & McHoull, 2015).

Ancient DNA is routinely treated to repair damage, such as nicks in sequence strands (e.g., Mouttham & al., 2015). We repaired most of our DNA extractions using NEBNext FFPE DNA Repair Mix (New England Biolabs, Ipswich, Massachusetts, U.S.A.), following the manufacturer's protocol. As a control, we left aliquots of the 2009 (*16L 145* (E)) silica and herbarium DNA extractions unrepaired (denoted by "-" in the Library names, Table 3), to test whether the use of a proprietary DNA repair kit increased final library amounts and sequencing quality, or introduced errors into the sequences.

Library preparation. — We trialled two commercial library kits: Illumina's TruSeq Nano library preparation kit (Illumina, FC-121-4001, San Diego, California, U.S.A.), which recommends 100 ng of starting DNA, and NEBNext Ultra library preparation kit for Illumina (New England Biolabs, E7370S), which has been optimized for as little as 5 ng starting DNA. Most library preparations followed the respective manufacturer's protocols; however, for some samples modifications were made to the fragmentation, size selection and final amplification steps (as outlined below) to accommodate small quantities of or degraded starting DNA (Table 2).

DNA was fragmented using a Bioruptor Plus (Diagenode, Liège, Belgium), with fragmentation tailored to the size distribution of extracted DNA. Samples with higher molecular weight DNA were sonicated for 8 cycles of 30 seconds on/90 seconds off (low power). For samples with more degraded genomic DNA the number of sonication cycles was reduced, with no sonication for the most degraded samples (Table 3). After sonication the samples were cleaned-up following the TruSeq Nano protocol; during this bead clean-up process 20%–30% of the starting DNA is lost (data not shown), therefore post-sonication clean-up was omitted for libraries where starting DNA amounts were low. The amount of input genomic DNA for samples that were not sonicated was reduced to a

Table 2. DNA extraction details for *Inga umbellifera* samples.

Year	Voucher	Material type	Tissue	Extraction amount	DNA extraction method	DNA extraction number	DNA quality (description)	DNA integrity number	DNA in extraction (ng)
1835	*Matthews 1593* (E)	Herb.	Leaf	2 replicates, total ca. 2 cm²	DNeasy	13	degraded, mostly ca. 100 bp fragments	not assigned	110
1841	*Hostmann 170* (K)	Herb.	Leaf	4 replicates each ca. 2 cm²	DNeasy + QiaQuick (pool)	7	too little to visualize	not assigned	5.8
1841	*Hostmann 170* (K)	Herb.	Leaf	4 replicates, each ca. 2 cm²	DNeasy	8	too little to visualize	not assigned	5.4
1932	*Lawrance 260* (E)	Herb.	Inflorescence	10–20 mg of fragments	DNeasy	11	degraded, mostly ca. 100 bp fragments	not assigned	2305
1932	*Lawrance 260* (E)	Herb.	Leaf	10–20 mg of fragments	DNeasy	12	mostly small fragments (ca. 100 bp), some longer fragments to 1000 bp	1.7	809
1948	*For. Dept. Brit. Guiana 5682* (K)	Herb.	Leaf	4 replicates, each ca. 2 cm²	DNeasy	5	mostly small fragments (ca. 100 bp), some longer fragments to 1000 bp	not assigned	240
1948	*For. Dept. Brit. Guiana 5682* (K)	Herb.	Leaf	4 replicates, each ca. 2 cm²	DNeasy	6	mostly small fragments (ca. 100 bp), some longer fragments to 700 bp	not assigned	200
2004	*K. Dexter 401* (E)	Herb.	Leaf	ca. 2 cm²	DNeasy	9	smear up to ca. 2000 bp	not assigned	325
2004	*K. Dexter 401* (E)	Herb.	Leaf	ca. 2 cm²	QiaQuick	10	smear up to ca. 2000 bp	2.7	220
2009	*K. Dexter 16L 145* (E)	Herb.	Leaf	ca. 2 cm²	DNeasy	1	high molecular weight smear	5.3	690
2009	*K. Dexter 16L 145* (E)	Herb.	Leaf	ca. 2 cm²	QiaQuick	2	high molecular weight smear	5.3	575
2009	*K. Dexter 16L 145* (E)	Silica	Leaf	ca. 2 cm²	DNeasy	3	high molecular weight band	6.3	540
2009	*K. Dexter 16L 145* (E)	Silica	Leaf	ca. 2 cm²	DNeasy	4	high molecular weight band	6.2	430

maximum of 70–80 ng (Table 3), consistent with the estimated DNA loss after sonication.

We followed different size selection protocols for the NEBNext Ultra and TruSeq Nano Kits to allow for the different average fragment sizes in different DNA extractions. We followed two size selection protocols for the NEBNext libraries. For samples with higher molecular weight genomic DNA, we followed the 400–500 bp insert protocol; for more degraded genomic DNA, we followed the 250–300 bp insert protocol (Table 3). Three NEBNext libraries were generated without any size selection, due either to low starting DNA quantities (libraries H1841_NEB7+, H1841_NEB8+), or as a comparison of making libraries from highly degraded starting DNA both with and without size selection (library H1932_NEB11b+v2 without size selection, cf H1932_NEB11b+ with size selection).

Size selection for most of the TruSeq libraries proceeded as recommended for a 350 bp average insert, using 30 µl of undiluted beads at the second step in the size selection process (Table 3). We tested modifications to this protocol using larger volumes of beads as a way of increasing the recovery of

Table 3. Library preparation metrics. Sample S2004 (in italics) is from Nicholls & al. (2015), for comparison to H2004 libraries generated from herbarium collections of the same material.

Library name	DNA extraction number	Library type	DNA repair	Starting DNA (ng)	Sonication cycles	Post sonication cleanup	Size selection protocol	No. of pre-capture PCR cycles	Final library mean fragment size (bp)	Capture/ sequencing pool
H1835_NEB13+	13	NEB	yes	16	1	no	250–300 bp insert	10	341	pool2
H1841_NEB7+	7	NEB	yes	<5	none	—	none	12	419	pool2
H1841_NEB8+	8	NEB	yes	5	none	—	none	12	405	pool2
H1932_NEB11a+	11	NEB	yes	80	none	—	250–300 bp insert	8	348	pool2
H1932_NEB11b+	11	NEB	yes	80	none	—	250–300 bp insert	8	341	pool2
H1932_NEB11b+v2	11	NEB	yes	40	none	—	none	10	330	pool2
H1932_NEB12+	12	NEB	yes	100	3	yes	400–500 bp insert	7	456	pool3
H1948_NEB5+	5	NEB	yes	23	none	—	250–300 bp insert	10	279	pool2
H1948_NEB6+	6	NEB	yes	59	none	—	250–300 bp insert	7	326	pool2
H2004_NEB9+	9	NEB	yes	100	3	yes	400–500 bp insert	7	618	pool3
H2004_NEB10+	10	NEB	yes	39	3	yes	400–500 bp insert	10	565	pool3
H2009_NEB1-	1	NEB	no	100	8	yes	400–500 bp insert	7	619	pool3
H2009_NEB1+	1	NEB	yes	102	8	yes	400–500 bp insert	7	576	pool1
H2009_NEB2-	2	NEB	no	100	8	yes	400–500 bp insert	7	604	pool1
H2009_NEB2+	2	NEB	yes	28	8	yes	400–500 bp insert	10	633	pool1
S2009_NEB3-	3	NEB	no	100	8	yes	400–500 bp insert	7	649	pool1
S2009_NEB3+	3	NEB	yes	94	8	yes	400–500 bp insert	7	620	pool1
H1835_Tru13+	13	TruSeq	yes	40	1	no	80 µl beads	10	352	pool2
H1932_Tru11a+	11	TruSeq	yes	80	none	—	80 µl beads	8	397	pool2
H1932_Tru11b+	11	TruSeq	yes	80	none	—	80 µl beads	8	377	pool2
H1932_Tru11b+v2	11	TruSeq	yes	80	none	—	50 µl beads	8	347	pool3
H1932_Tru12+	12	TruSeq	yes	101	3	yes	30 µl beads	8	449	pool3
H1948_Tru5+	5	TruSeq	yes	70	none	—	80 µl beads	8	356	pool2
H1948_Tru6+	6	TruSeq	yes	80	none	—	80 µl beads	8	351	pool2
H2004_Tru9+	9	TruSeq	yes	101	3	yes	30 µl beads	8	509	pool3
H2004_Tru10+	10	TruSeq	yes	71	3	yes	30 µl beads	9	523	pool3
S2004_TruKD401	*n/a*	*TruSeq*	*no*	*100*	*8*	*yes*	*30 µl beads*	*8*	*532*	*n/a*
H2009_Tru1-	1	TruSeq	no	101	8	yes	30 µl beads	8	539	pool3
H2009_Tru1+	1	TruSeq	yes	101	8	yes	30 µl beads	8	541	pool1
H2009_Tru2-	2	TruSeq	no	101	8	yes	30 µl beads	8	537	pool1
H2009_Tru2+	2	TruSeq	yes	71	8	yes	30 µl beads	9	549	pool1
S2009_Tru3-	3	TruSeq	no	101	8	yes	30 µl beads	8	553	pool1
S2009_Tru3+	3	TruSeq	yes	101	8	yes	30 µl beads	8	573	pool1

fragments <300 bp (data not shown), resulting in a modified protocol using 80 µl of undiluted beads that was then used for libraries made from degraded genomic DNA. An intermediate bead volume of 50 µl was used for one sample (library H1932_Tru11b+v2) to test whether a final library containing larger fragments could be made from highly degraded starting DNA.

The final step in both the TruSeq Nano and NEBNext Ultra protocols, prior to the capture stage, involves a PCR amplification. The number of PCR cycles performed was varied as a function of the amount of starting DNA (range 7–12 cycles; see Table 3). Post-PCR libraries were run on a Bioanalyser (Agilent) to determine fragment size distributions (see Electr. Suppl.: Fig. S2), and all were diluted to 10 nM.

Library pooling. — Equimolar amounts of the 32 TruSeq and NEB post-amplification libraries were combined into three pools based on their size distributions (Table 3; Electr. Suppl.: Fig. S2). Two pools contained either 9 or 10 libraries with average fragment sizes typical of libraries produced using the standard TruSeq Nano or NEBNext Ultra protocols (pool 1: average fragment sizes 537–649 bp; pool 3: average fragment sizes 347–619 bp), and one pool contained 13 libraries from samples with much more degraded starting DNA (pool 2: average fragment sizes 279–419 bp).

Target enrichment. — Targeted enrichment was carried out using the same bait set as Nicholls & al. (2015), manufactured by MYcroarray (Ann Arbor, Michigan, U.S.A.) and following the MYbaits protocol v.2.3.1. Hybridisation was carried out for 19 hours, a high-stringency wash was used, and the post-hybridisation PCR involved 14 cycles. Enriched pools were quantified by Qubit and their fragment size distributions assessed on a Bioanalyser. Each pool was sequenced on a separate 250 bp paired-end run of an Illumina MiSeq machine at the Edinburgh Genomics facility.

Analyses. — Analysis of the sequences followed the procedure in Nicholls & al. (2015), using scripts available at https://github.com/ckidner/Targeted_enrichment.git. Reads were quality trimmed, then aligned using default parameters to the bait sequences using Bowtie v.2 (Langmead & Salzberg, 2012) to determine the success of capture (Fig. 1A; Table 4). The bait set was designed using transcriptome sequence from three *Inga* species, choosing sequences showing a phylogenetically useful level of variation, coding for key secondary synthesis enzymes or differentially expressed between species (see Nicholls & al., 2015). The bait set can capture multiple paralogs per target locus so we followed the procedures in Nicholls & al. (2015) to minimise the effects of this. A conservative set of parameters for more stringent Bowtie mapping was derived empirically in order to derive data for just a single paralog per target locus. Bowtie uses a formula including read length and an intercept constant to determine the alignment score. We ran Bowtie

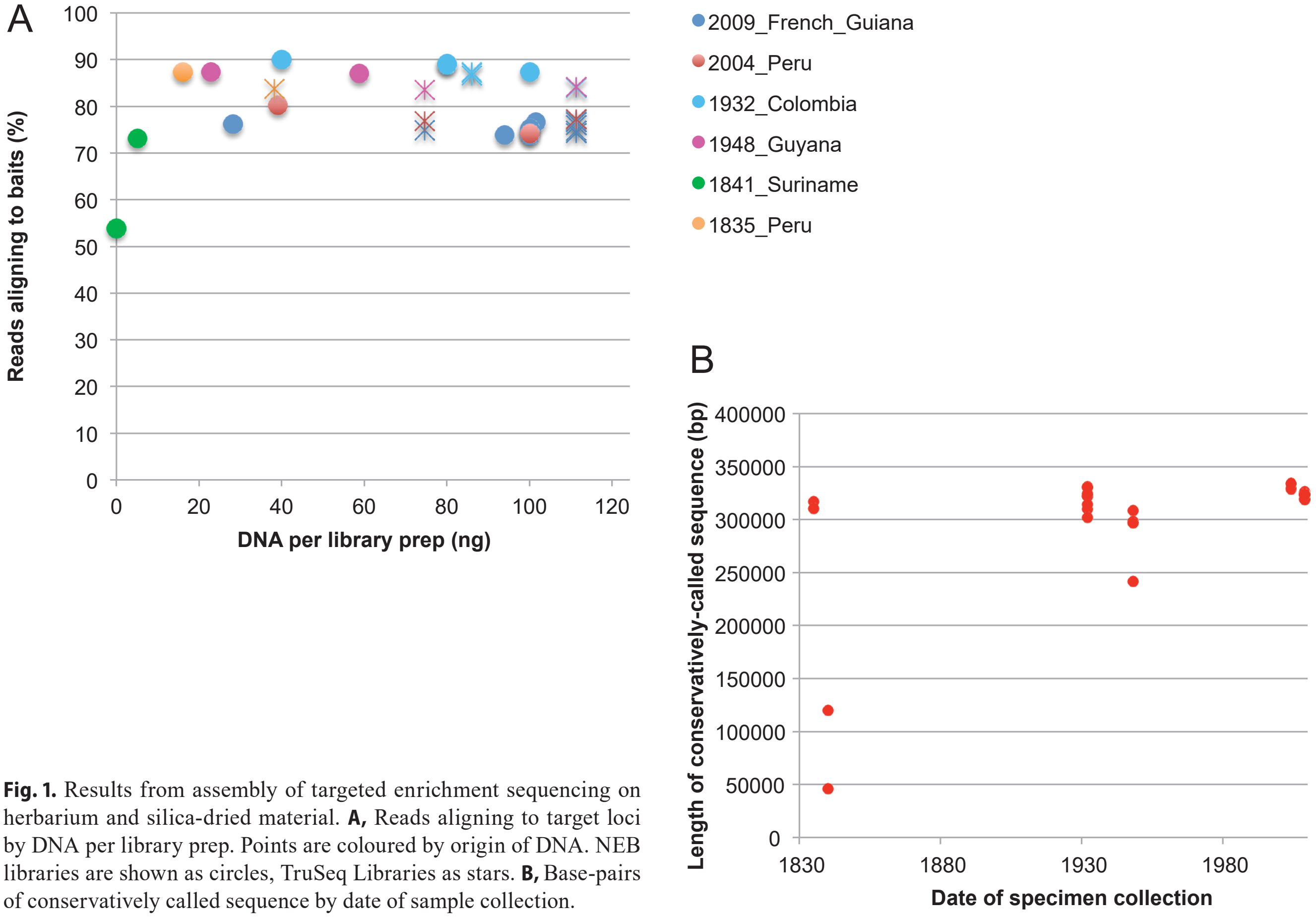

Fig. 1. Results from assembly of targeted enrichment sequencing on herbarium and silica-dried material. **A,** Reads aligning to target loci by DNA per library prep. Points are coloured by origin of DNA. NEB libraries are shown as circles, TruSeq Libraries as stars. **B,** Base-pairs of conservatively called sequence by date of sample collection.

alignments on four libraries (H2009_Tru1-, H2009_NEB1-, S2009_NEB3-, H1932_NEB11b+v2) using intercept constants from 20 to 420. As in Nicholls & al. (2015) the change in standardised quality of variant calls levelled out from above 260, reflecting fewer paralogs mapping at this higher stringency. We used an intercept constant of 320 as a conservative threshold. vcf files were derived from the conservatively mapped BAM files using SAMtools v.0.1.18 (Li & al., 2009). These files were then edited to remove low-quality base calls and to remove indels, and a consensus sequence for each target locus was derived using bcftools (part of SAMtools; Li & al., 2009). We subsequently limited our sequence quality and phylogenetic analyses to reads that mapped to the subset of 194 target loci designed specifically for *Inga* which did not show evidence of paralogy by either anomalous coverage levels or anomalous placement of an outgroup (Nicholls & al., 2015).

We combined the consensus sequence for each locus from each herbarium sample with data derived from silica samples

Table 4. Results of MiSeq runs by library, with the two poorest performing libraries (from Suriname) in italics.

Library	No. of trimmed reads	% reads aligned to baits	% reads aligned to *Inga* plastid	Average quality score of variant positions (AQV)	Number of variant bases	Loci recovered (max 276)	Conservatively called sequence (CCS), bp
H1835_NEB13+	1013414	87.4%	4.3%	139.18	7186	249	317244
H1841_NEB7+	*214315*	*53.9%*	*0.7%*	*101.80*	*883*	*137*	*46045*
H1841_NEB8+	*365550*	*73.2%*	*0.8%*	*73.44*	*2773*	*226*	*120148*
H1932_NEB11+a	1226043	89.0%	4.7%	157.83	6377	248	322337
H1932_NEB11+b	862599	89.1%	4.1%	141.89	6253	246	310470
H1932_NEB11+bv2	1152606	90.0%	2.3%	133.15	5930	248	301994
H1932_NEB12+	1919229	87.4%	6.4%	173.56	6463	249	331326
H1948_NEB5+	583010	87.4%	1.6%	94.94	5028	239	241758
H1948_NEB6+	704977	87.1%	3.7%	136.32	6132	247	298809
H2004_NEB9+	1787314	74.3%	9.2%	168.53	7018	248	328618
H2004_NEB10+	1595602	80.3%	10.4%	174.46	7135	250	334242
H2009_NEB1-	1711918	75.0%	8.6%	169.24	6482	248	326187
H2009_NEB1+	1658799	76.6%	8.2%	169.21	6484	250	324340
H2009_NEB2-	1355984	75.2%	8.3%	163.79	6525	247	322957
H2009_NEB2+	1668026	76.2%	8.5%	171.90	6516	250	326466
H2009_NEB3-	1513515	73.8%	8.3%	162.85	6463	246	319683
H2009_NEB3+	1504758	74.0%	8.4%	161.80	6419	245	320273
H1835_Tru13+	659161	84.2%	5.2%	132.97	7045	247	310949
H1932_Tru11+a	1584437	87.7%	3.8%	155.89	6246	248	322199
H1932_Tru11+b	1015706	87.5%	3.8%	144.88	6194	248	314862
H1932_Tru11+b2	1416246	87.0%	4.6%	159.42	6448	249	324910
H1932_Tru12+	1774508	84.4%	6.3%	169.72	6503	248	330462
H1948_Tru5+	1042441	83.9%	2.6%	136.01	5941	248	296844
H1948_Tru6+	892927	84.6%	3.9%	145.22	6211	247	308853
H2004_Tru9+	1958838	77.9%	9.2%	173.90	7041	249	333904
H2004_Tru10+	1576572	77.4%	9.5%	170.05	7066	248	330278
H2009_Tru1-	1338317	77.6%	9.1%	167.51	6601	249	324201
H2009_Tru1+	1536759	77.2%	8.4%	167.03	6594	248	325184
H2009_Tru2-	1476338	76.6%	8.4%	166.63	6569	249	323881
H2009_Tru2+	1226123	75.6%	8.5%	161.46	6572	249	319045
H2009_Tru3-	1630041	75.4%	8.6%	168.09	6509	250	324451
H2009_Tru3+	1753019	75.0%	8.4%	167.90	6512	249	323951
S2004_TruKD401	689439	74.4%	9.2%	156.29	5809	245	330396

The final row is the silica-dried material from *Dexter 401* (E) sequenced by Nicholls & al. (2015), for comparison with libraries from H2004.

obtained by Nicholls & al. (2015) representing multiple individuals from multiple *I. umbellifera* populations as well as a few closely related *Inga* species. The consensus multi-fasta files for each accession were converted into multi-fastas of loci by accession and ambiguous nucleotides and Ns (IUPAC code for aNy base) were removed. Each single-locus matrix was aligned using MAFFT v.7.130b with linsi settings (Katoh & al., 2009) and trimmed using trimAl v.1.2 with strict settings to remove poorly aligned regions (Capella-Gutiérrez & al., 2009). Metrics for each of these alignments are shown in Electr. Suppl.: Table S1. Key metrics are AVQ (average quality score for a consensus call), and CCS (the number of bp of conservatively called sequence produced). The maximum length of a single locus alignment was 3390 bp, with an average of 1321 bp. The maximum number of phylogenetically informative characters per alignment was 130, with an average of 36.4 (Electr. Suppl.: Table S1). RAxML analysis of the 113 loci with least missing data was performed but the individual locus trees provided no support at this level (see set of phylogenies for individual loci in the supplementary data). All loci were then concatenated, producing an alignment of 229,995 bp and 5975 phylogenetically informative sites. Phylogenetic analysis of the final alignment was performed using RaxML (Stamatakis & al., 2008) employing a GTR+G model of sequence evolution with 1000 bootstrap replicates to estimate node support.

■ RESULTS

We successfully enriched and sequenced DNA libraries constructed from herbarium material of *Inga umbellifera* collected in 1835, 1841, 1932, 1948, 2004 and 2009 (Table 4; Fig. 1). Despite wide variation in the quantity and quality of DNA extracted, we produced good quality libraries from most of our samples (Electr. Suppl.: Fig. S2). The exception was the herbarium specimen collected in 1841 (H*ostmann 170* (E), 1841), which yielded very little genomic DNA for library construction (5 ng for one replicate, <5 ng for the other replicate). Between 54% and 90% of reads for each library were on target, mapping to a set of 1400 target loci. Within this, our focal set of loci (276 loci designed specifically for *Inga*) produced >300,000 base pairs of high-quality sequence in nearly all libraries. Below we summarise results relating to different aspects of the DNA and library preparations.

Starting DNA quality and quantity. — DNA quality varied widely amongst samples (Table 2; Fig. 2B). Unsurprisingly, the highest DNA integrity number (DIN) values were given to extractions from the most recent silica-dried material (2009; DIN 6.2–6.3), and the second-highest to those from herbarium material from the same collection (DIN 5.3). The herbarium material from 2004 had the third-highest DIN (2.7), but this represented a substantial drop from the 2009 collection and the DNA was degraded compared to that from the 2009 herbarium sample, as shown by the absence of a high molecular weight band. None of the herbarium samples that were more than 11 years old contained high molecular weight DNA and, with the exception of the H1932 extraction 12 (DIN 1.7), all failed to fulfil the recommendations for DIN assignment; however, these DNA extractions were still adequate for constructing high-quality libraries.

Contrasting amounts of starting DNA were used in five replicate pairs of libraries (Table 3): H2009_NEB1+ (102 ng) versus H2009_NEB2+ (28 ng); H2009_Tru1+ (101 ng) versus H2009_Tru2+ (71 ng); H1948_NEB5+ (23 ng) versus H1948_NEB6+ (59 ng); H2004_NEB9+ (100 ng) versus H2004_NEB10+ (39 ng); H2004_Tru9+ (101 ng) versus H2004_Tru10+ (71 ng). In three of the five pairs, the library generated from more input DNA had a higher average quality score of variant

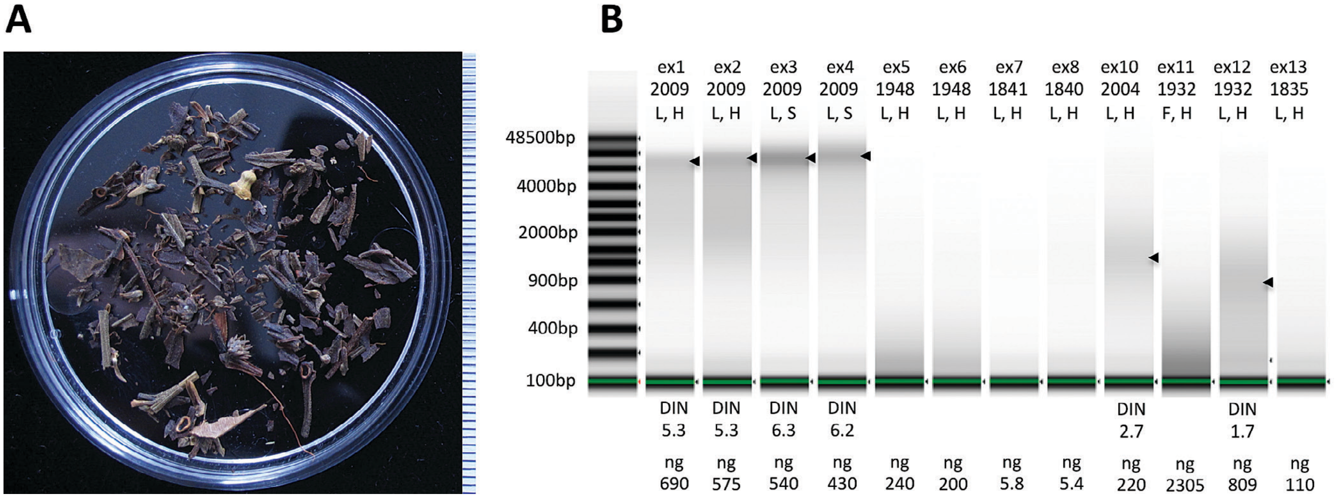

Fig. 2. Herbarium DNA used for targeted enrichment. **A,** Fragments from herbarium voucher *Lawrance 260* (E) (syntype for *Inga lawranceana*) sampled for DNA extraction, scale bar 1 mm increments. **B,** Agilent TapeStation Genomic DNA ScreenTape gel image for extracted DNA. Samples are labeled with their extraction number, year of collection, tissue type (F, floral parts; L, leaf parts), preservation method (H, herbarium; S, silica), DNA integrity number (DIN) and the total amount of DNA extracted; the bar at the bottom of each image is the 100 bp lane standard. Six of the 12 samples lacked higher molecular weight bands (black arrows) and could not be assigned DINs.

positions (AQV) and a greater amount of conservatively called sequence (CCS, Table 4). This is no more than expected by chance, and differences in AQV and CCS are of the same magnitude as those seen between replicate libraries constructed using the same amounts of starting DNA.

There was no strong link between quantity of starting DNA and capture success, except for library H1841_NEB7+, from *Hostmann 170* (K), which was made with <5 ng input DNA and showed the lowest reads on target (Fig. 1A; Table 4). The three libraries with the shortest amount of CCS all had very low amounts of input DNA—the two *Hostmann 170* (K) libraries (H1841_NEB7+: 46 kb, H1841_NEB8+: 120 kb) and a NEBNext library from the 1948 accession (*FDBG 5682* (K)) generated from 22.6 ng of DNA (H1948_NEB5+: 242 kb; Fig. 1B). However, the relationship between input DNA quantity and data output was not absolute: a NEBNext library (H1835_NEB13+) generated from 16 ng of DNA from the 1835 collection (*Matthews 1593* (E)) gave as much high-quality CCS (317 kb) as libraries made with 100 ng of input DNA.

Tissue preservation: silica-dried versus herbarium. — There was no significant difference in AQV or CCS between libraries derived from silica-dried material and herbarium material (Table 5), with the same magnitude of variation between replicates made from the same starting material as seen between libraries made from silica-dried versus herbarium sheet leaves. This implies that the targeted enrichment method employed in this study can be used to generate high-quality sequence data regardless of tissue preservation method.

DNA elution kits. — Of two comparisons between DNA extraction using DNeasy and QiaQuick elution columns, using the same starting DNA quantity (H2009_Tru1- versus H2009_Tru2-; H2009_NEB1- versus H2009_NEB2-), the AQV (163.79–169.24) and CCS (322,957–326,187) were high for all 4 libraries, although values were marginally lower for the two libraries generated from DNA that had been eluted through QiaQuick columns (Table 4).

DNA repair. — The DNA repair process led to a substantial loss of DNA (data not shown). Comparing repaired and unrepaired DNA for Libraries S2009 and H2009, DNA repair did not increase AQV or CCS per library (Table 5). Libraries with and without repair gave consistently large amounts (319–326 kb CCS) of high-quality (AQV 161–172) sequence. More importantly, no sequence errors were introduced by the repair process. Within the final alignment of 229,995 bases used for the phylogenetic analysis, there were eight variable sites (0.0035%) across the six replicate libraries made from unrepaired DNA of the 2009 (*16L 145* (E)) accession, and nine variable sites (0.0039%) across the six replicates constructed using repaired DNA.

Library-construction kit comparisons. — We performed a t-test on AQV and CCS values of TruSeq and NEB libraries generated from both silica gel preserved and herbarium material of sample *16L 145 (E)*. Values were not significantly different (Table 5). However, the direct comparison of TruSeq and NEBNext libraries is partially confounded by different quantities of starting DNA used for both kits. Equal amounts of nine DNA extractions were used as starting material for the two different kits: H2009_1-, H2009_1+, H2009_2-, S2009_3-, S2009_3+, H2004_9+, H1932_11a+, H1932_11b+, and H1932_12+ (Table 4). In five of these, the TruSeq libraries had higher AQV, and in six the TruSeq libraries had more CCS. In contrast, for three of the five samples where significantly more DNA went into the TruSeq library, the NEBNext libraries, despite starting with lower input amounts of DNA, had higher AQV and more CCS. This high performance of the NEB kit at lower input amounts is consistent with the published recommendations for starting amounts for the respective kits. There is a slight additional time cost to generate TruSeq libraries, but sufficient sequence to generate a fully resolved and statistically supported phylogram was recovered from both preparation methods.

Size selection. — Of the two NEB libraries generated from the same DNA extraction but performed with and without size selection, the AQV and CCS were both higher for the library generated with size selection (H1932_NEB11b+: AQV 142, CCS 310 kb) rather than that generated without size selection (H1932_NEB11b+v2: AQV 133, CCS 302 kb), although capture success was marginally lower (89% versus 90%). The same patterns were seen in the TruSeq libraries for the same DNA extraction but generated using selection strategies for larger (50 µl beads; H1932_Tru11b+v2: AQV 159, CCS 325 kb, 87% on target) and smaller (80 µl beads; H1932_Tru11b+: AQV 144, CCS 315 kb, 88% on target) sized fragments. Reads appear to be more reliably mapped (and hence coverage increased, with resultant higher quality base calls) when they originate from fragments coming from a library with a narrower distribution of insert sizes. This could be related to more reliable sequencing in more uniform libraries

Table 5. t-test (Welch two sample) for effect of sample type, library type and DNA repair.

		N	Average quality of variants	*P*-value	Average bp of sequence	*P*-value
Source	Herbarium	8	167.0961	0.3737	324032.6	0.2388
	Silica	4	165.1587		322089.5	
Library	TruSeq	6	166.435	0.988	323452.2	0.9298
	NEB	6	166.4657		323317.7	
Repair	Repaired	6	166.548	0.9239	323209.8	0.8184
	Unrepaired	6	166.3527		323560.0	

Phylogenetic analysis. — The trimmed alignment combining the consensus sequences from the herbarium accessions sampled herein with those derived from silica-dried accessions of *I. umbellifera* sampled by Nicholls & al. (2015) was 229,995 bp long. In RAxML analyses, all the herbarium-derived libraries nested within the *I. umbellifera* samples from Nicholls & al. (2015) (Fig. 3). Replicate libraries from the same accession formed monophyletic clusters, with the exception of the two low-quality libraries from the sample *Hostmann 170* (K) 1841 that were generated from very low amounts of input DNA and had most missing data. Despite this, they did have sufficient signal to allow placement as *I. umbellifera*. The branch lengths between libraries from the same accessions were very short. This was regardless of whether the libraries were generated from silica-dried (black on Fig. 3) or herbarium (coloured on Fig. 3) material, the type of library preparation kit or the DNA extraction method. This confirms the robustness of the targeted enrichment procedure for producing reliable genome-scale sequence data, and its insensitivity to library type, DNA repair or input genomic DNA quantity or quality (above quite low thresholds).

The intraspecific sequence variation within *I. umbellifera* resolves samples geographically. For example, the five accessions from Peru are monophyletic, with libraries from the 1835 collection from San Martín in northern Peru (H1835, in yellow) sister to the rest (all from southern Peru). A clade otherwise from French Guiana also includes the sample from nearby Guyana (H1948, in pink). The isotype of *I. lawranceana* (*Lawrance 260* (E), H1932, in blue), from the western side of the Andes mountains in Colombia, is nested within *I. umbellifera*, and

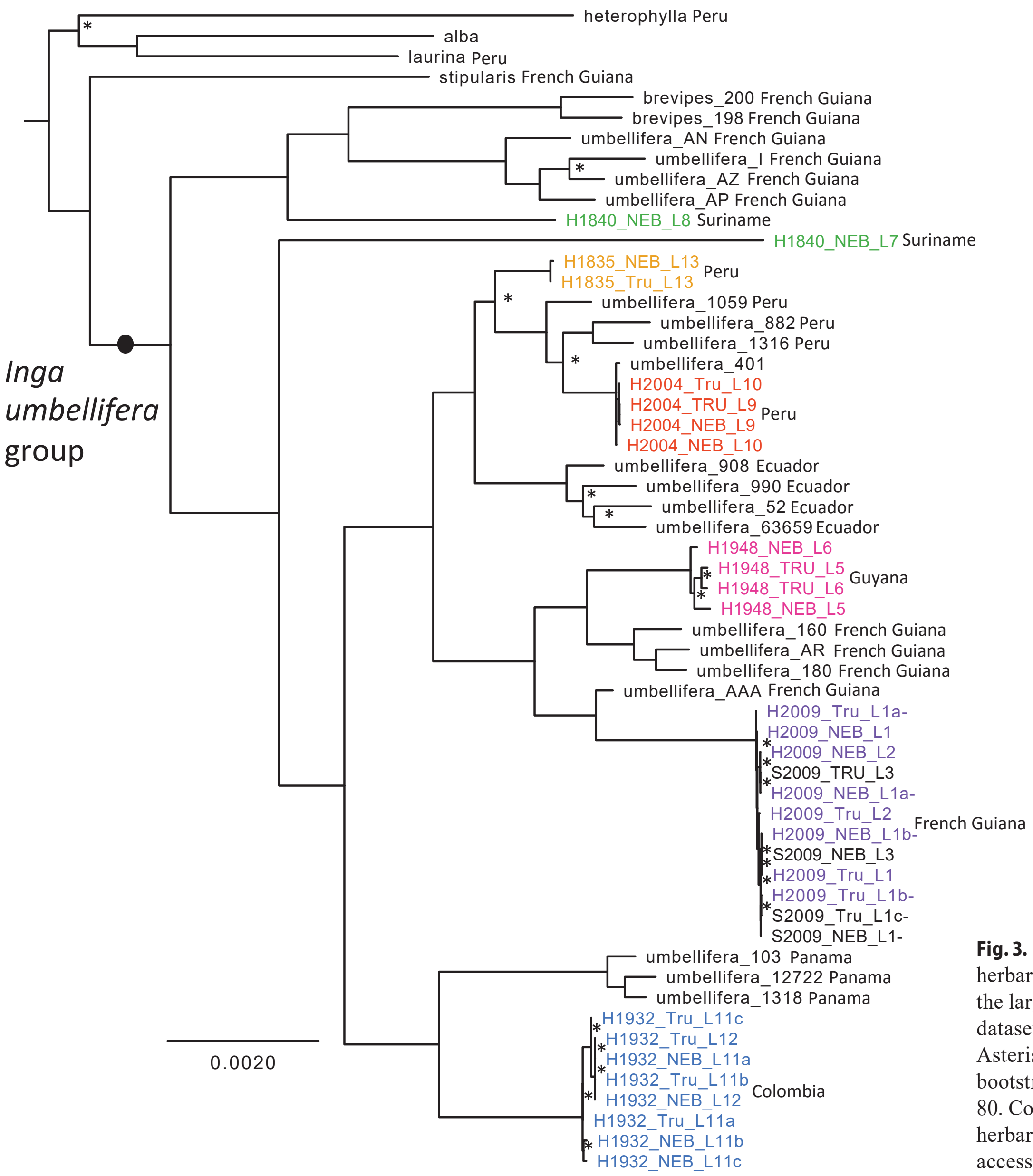

Fig. 3. Phylogeny placing herbarium samples within the larger *Inga umbellifera* dataset of Nicholls & al. (2015). Asterisks next to nodes indicate bootstrap support of less than 80. Coloured names indicate herbarium material (coloured by accession).

sister to three accessions from neighbouring Panama. This placement of herbarium samples in their correct geographic population within *I. umbellifera* provides further support of the robustness of data derived from the targeted enrichment process.

Plastid DNA. — Although we achieved high enrichment efficiencies, some reads were not of target loci. Of these, many were of plastid origin, and the proportion of these reads possibly reflects the plant sample rather than the method of DNA extraction or library preparation. Libraries from the two youngest samples (*Dexter 401* (E) 2004 and *16L 145* (E) 2009) have the most reads mapping to the plastid (8.2%–10.4%), regardless of the method of source tissue preservation or library preparation (Table 4). Fewer reads from the older herbarium specimens mapped to the plastid—*FDBG 5682* (K) 1948: 1.6%–3.9%; *Lawrance 260* (E) 1932: 2.3%–6.4%; *Hostmann 170* (K) 1841: 0.7%–0.8%; *Matthews 1593* (E) 1835: 4.3%–5.2%. This relationship may be due to more degraded DNA failing to map to the plastid (Electr. Suppl. Fig. S3). The wider range of values from the 1932 accession reflects the source DNA coming from two tissue types, inflorescence (2.3%–4.7%) and leaf (6.3%–6.4%), as the leaf would be expected to contain more plastids. The plastid reads could potentially be used to construct a plastid phylogeny; however in this group, plastid sequence does not include sufficient variation for resolution (Nicholls & al., 2015).

■ DISCUSSION

We have successfully demonstrated that herbarium specimens, some collected as long as 180 years ago, can be used to generate genomic-scale DNA sequence data. By using a targeted enrichment process, we can obtain high-quality, high-coverage sequence data from many hundreds of nuclear loci, and hence provide a route for utilisation of herbaria for a range of projects beyond their traditional role as repositories for specimens for morphological taxonomy. Regardless of the degree of degradation in source genomic DNA, the use or not of DNA repair, variation in size selection protocols and library preparation methods, 30 of the 32 libraries produced here gave high-quality sequence data, resulting in robust and reliable placement of the respective *Inga umbellifera* accessions within a phylogenetic and population genetic context.

The most important variable for successful sequencing of herbarium material appears to be a minimum threshold quantity of starting DNA. The reasons for low recovery of DNA from herbarium samples could be many, including how the collector treated the specimen in the field, how rapidly it was dried, and subsequent storage. The two replicate libraries generated from very low quantities of DNA (≤5 ng) from the same herbarium sheet produced poor-quality sequences and did not resolve as a monophyletic cluster, although the data were informative enough to place them in the correct species. However, libraries made from only three times this amount (16 ng) provided large volumes of high-quality sequence data. Technological advances may overcome the barrier even this small threshold imposes, for instance through the use of novel kits optimised for exceedingly low input DNA amounts (e.g., the NEBNext Ultra II DNA library prep kit, New England Biolabs, with sample inputs as low as 500 pg).

Our experience suggests several recommendations for future next-generation sequencing work, especially targeted enrichment, using herbarium material.

Firstly, of the two methods tested, although we cannot statistically test the difference, we suggest using standard DNeasy columns rather than QiaQuick columns for genomic DNA extractions, in order to recover even the smallest amounts of high molecular weight genomic DNA fragments.

Secondly, we suggest using more rather than less starting DNA when possible, to provide a greater chance that the genomic DNA sample contains sufficient copies of the targeted loci of interest. Our data suggest there is not a linear relationship between DNA input and library quality, but a threshold below which library quality rapidly falls off.

Thirdly, although size selection is recommended when working with sufficient quantities of DNA, for very degraded samples it is possible to generate good quality libraries without size selection.

Fourth, our testing only assessed the impact of repair in younger herbarium material and not very degraded older DNA samples, and does not show a statistical difference between repaired and unrepaired libraries. Our data show that the repair process does not introduce errors into DNA. However, the loss of starting DNA quantity during the repair process may impact on final data quality as seen in H1841. We have no evidence that DNA repair is necessary in targeted enrichment.

However, the most important result from this study is that, regardless of starting DNA quality (rather than quantity), the library making and enrichment processes are repeatable and robust, with minor modifications having little effect on subsequent phylogenetic analyses. Even DNA that appears to be highly degraded (Fig. 2B) can be successfully used with this methodology.

The approach to calling consensus sequence data from the capture reads which we use here is highly conservative, deriving a single sequence for each bait and limiting the sequence to the bait itself, losing information contained in flanking sequences. This method is ideal for recovery of sequence from degraded DNA as it gives sequence only where there is strong support from many high-quality reads for calling any individual base. However, it does reduce the amount of information which can be derived from each bait, making individual gene trees uninformative (Electr. Suppl.: Table S1; (see set of phylogenies for individual loci in the supplementary data). The phylogenetic approach we use here, simple concatenation, is used to show that robust sequence data can be derived from herbarium data, but an experimental use of such data would likely implement a more detailed, population genetics approach (Bi & al., 2013; Nicholls & al., 2015) for a full scale analysis.

Our study was facilitated by the availability of a targeted enrichment bait set, previously developed using transcriptomic data for *Inga* (Nicholls & al., 2015). Although such resources are not yet available for all plant groups, the number of publically accessible plant transcriptomes and genomes is increasing (e.g., through the 1KP project https://sites.google.com/a/ualberta.ca/onekp/), as well as being increasingly cheap to develop in house for specific taxa. Bioinformatic tools for locus selection and bait design are also becoming easier to use (e.g., Chamala & al., 2015; Schmickl & al., 2015). With increasing usage of targeted enrichment as a way of obtaining genome-scale data cheaply for multiple accessions, resources that can be applied to herbarium material will become increasingly common. For example, target bait sets designed from conserved regions (e.g., DEB-1240045 AToL: Assembling the Pleurocarp Tree of Life http://pleurocarps.uconn.edu/project-2/), that could be expected to work across large phylogenetic distances, are already available.

Robustly placing herbarium specimens on molecular phylogenies is an invaluable check of nomenclatural concepts (typically based on morphology), and for classification of species that are either now extinct, or grow in regions that have become difficult or dangerous to sample. For example, we demonstrate here that the two isotypes we sampled, the 1932 *Inga lawranceana* isotype *Lawrance 260* (E) and the 1841 *Inga sciadion* isotype *Hostmann 170* (K) were correctly synonymised in Pennington (1997), being nested within *I. umbellifera* in the phylogeny (Fig. 3). With this targeted enrichment method making herbarium specimens available for genomic DNA sequencing, the possibilities for using herbaria are vast—for instance, using this method to sample genomic data from extinct species for functional studies, for population genetic analyses over deep timescales (e.g., Bi & al., 2013), or to help pinpoint genetic changes that correlate with historical geographic or climatic variations.

■ ACKNOWLEDGMENTS

This work was funded by NBAF grant 883, the Scottish Government's Rural and Environment Science and Analytical Services Division, and NSF Standard and Dimensions of Biodiversity grant DEB-1135733. We thank Helen Gunter, Edinburgh Genomics, for help and advice, Graham Stone (University of Edinburgh) for laboratory space, David Harris (RBGE), Kyle Dexter (University of Edinburgh), Terry Pennington (RBG-Kew) and Toby Pennington (RBGE) for specimens, Juan Carlos Villarreal (Laval University) for assistance with the phylogenetic reconstruction, and Toby Pennington additionally for comments on an earlier draft of this manuscript. We are grateful for constructive comments from three anonymous reviewers, and from communicating editor Michael Pirie.

■ LITERATURE CITED

Bi, K., Linderoth, T., Vanderpool, D., Good, J.M., Nielsen, R. & Moritz, C. 2013. Unlocking the vault: Next-generation museum population genomics. *Molec. Ecol.* 22: 6018–6032. http://dx.doi.org/10.1111/mec.12516

Capella-Gutiérrez, S., Silla-Martínez, J.M. & Gabaldón, T. 2009. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25: 1972–1973. http://dx.doi.org/10.1093/bioinformatics/btp348

Chamala, S., García, N., Godden, G.T., Krishnakumar, V., Jordon-Thaden, I.E., De Smet, R., Barbazuk, W.B., Soltis, D.E. & Soltis, P.S. 2015. MarkerMiner 1.0: A new application for phylogenetic marker development using angiosperm transcriptomes. *Applic. Pl. Sci.* 3(4): 1400115. http://dx.doi.org/10.3732/apps.1400115

Cronn, R., Knaus, B., Liston, A., Maughan, P.J., Parks, M., Syring, J. & Udall, J. 2012. Targeted enrichment strategies for next generation plant biology. *Amer. J. Bot.* 99: 291–311. http://dx.doi.org/10.3732/ajb.1100356

Davey, J.W. & Blaxter, M.L. 2010. RADSeq: Next-generation population genetics. *Briefings Funct. Genomics* 9: 416–423. http://dx.doi.org/10.1093/bfgp/elq031

Drábková, L., Kirschnew, J. & Vlček, Č. 2002. Comparison of seven DNA extraction and amplification protocols in historical herbarium specimens of Juncaceae. *Pl. Molec. Biol. Reporter* 20: 161–175. http://dx.doi.org/10.1007/BF02799431

Enk, J.M., Devault, A.M., Kuch, M., Murgha, Y.E., Rouillard, J.M. & Poinar, H.N. 2014. Ancient whole genome enrichment using baits built from modern DNA. *Molec. Biol. Evol.* 31: 1292–1294. http://dx.doi.org/10.1093/molbev/msu074

Gassmann, M. & McHoull, B. 2015. *DNA Integrity Number (DIN) with the Agilent 2200 TapeStation System and the Agilent Genomic DNA ScreenTape Assay: Technical overview.* Agilent Technologies. http://www.agilent.com/cs/library/applications/5991-5258EN.pdf

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., Gabriel, S., Jaffe, D.B., Lander, E.S. & Nusbaum, C. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* 27: 182–189. http://dx.doi.org/10.1038/nbt.1523

Grover, C.E., Salmon, A. & Wendel, J.F. 2012. Targeted sequence capture as a powerful tool for evolutionary analysis. *Amer. J. Bot.* 99: 312–319. http://dx.doi.org/10.3732/ajb.1100323

Jones, M.R. & Good, J.M. 2015. Targeted capture in evolutionary and ecological genomics. *Molec. Ecol.* 25: 185–202. http://dx.doi.org/10.1111/mec.13304

Katoh, K., Asimenos, G. & Toh, H. 2009. Multiple alignment of DNA sequences with MAFFT. Pp. 39–64 in: Posada, D. (ed.), *Bioinformatics for DNA sequence analysis*. New York: Humana Press. http://dx.doi.org/10.1007/978-1-59745-251-9_3

Langmead, B. & Salzberg, S.L. 2012. Fast gapped-read alignment with Bowtie2. *Nature, Meth.* 9: 357–359. http://dx.doi.org/10.1038/nmeth.1923

Lemmon, A.R., Emme, S.A. & Lemmon, E.M. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Syst. Biol.* 61: 727–744. http://dx.doi.org/10.1093/sysbio/sys049

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & 1000 Genome Project Data Processing Subgroup 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25: 2078–2079. http://dx.doi.org/10.1093/bioinformatics/btp352

Linderholm, A. 2016. Ancient DNA: The next generation—chapter and verse. *Biol. J. Linn. Soc.* 117: 150–160. http://dx.doi.org/10.1111/bij.12616

Mamanova, L., Coffey, A.J., Scott, C.E., Kozarewa, I., Turner, E.H., Kumar, A., Howard, E., Shendure, J. & Turner, D.J. 2010. Target-enrichment strategies for next-generation sequencing. *Nature, Meth.* 7: 111–118. http://dx.doi.org/10.1038/nmeth.1419

McCormack, J.E., Hird, S.M., Zellmer, A.J., Carstens, B.C. & Brumfield, R.T. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. *Molec. Phylogen. Evol.* 66: 526–538. http://dx.doi.org/10.1016/j.ympev.2011.12.007

Mouttham, N., Klunk, J., Kuch, M., Fourney, R. & Poinar, H. 2015. Surveying the repair of ancient DNA from bones via high-throughput sequencing. *BioTechniques* 59: 19–25. http://dx.doi.org/10.2144/000114307

Nachman, M.W. 2013. Genomics and museum specimens. *Molec. Ecol.* 22: 5966–5968. http://dx.doi.org/10.1111/mec.12563

Nicholls, J.A., Pennington, R.T., Koenen, E.J.M., Hughes, C.E., Hearn, J., Bunnefeld, L., Dexter, K.G., Stone, G.N. & Kidner, C.A. 2015. Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the neotropical rain forest genus *Inga* (Leguminosae: Mimosoideae). *Frontiers Pl. Sci.* 6: 710. http://dx.doi.org/10.3389/fpls.2015.00710

Pennington, T.D. 1997. *The Genus* Inga: *Botany.* Richmond: The Royal Botanic Gardens, Kew.

Rowe, K.C., Singhal, S., Macmanes, M.D., Ayroles, J.F., Morelli, T.L., Rubidge, E.M., Bi, K. & Moritz, C.C. 2011. Museum genomics: Low-cost and high-accuracy genetic data from historical specimens. *Molec. Ecol. Resour*ces 11: 1082–1092. http://dx.doi.org/10.1111/j.1755–0998.2011.03052

Särkinen, T., Staats, M., Richardson, J.E., Cowan, R.S. & Bakker, F.T. 2012. How to open the treasure chest? Optimising DNA extraction from herbarium specimens. *PLoS ONE* 7: e43808. http://dx.doi.org/10.1371/journal.pone.0043808

Sass, C., Iles, W.J., Barrett, C.F., Smith, S.Y., Specht C.D. 2016. Revisiting the Zingiberales: Using multiplexed exon capture to resolve ancient and recent phylogenetic splits in a charismatic plant lineage. *PeerJ* 4: e1584. http://dx.doi.org/10.7717/peerj.1584

Schmickl, R., Liston, A., Zeisek, V., Oberlander, K., Weitemier, K., Straub, S.C.K., Cronn, R.C. & Suda, J. 2016. Phylogenetic marker development for target enrichment from transcriptome and genome skim data: The pipeline and its application in southern African *Oxalis* (Oxalidaceae). *Molec. Ecol. Resources* 16: 1124–1135. http://dx.doi.org/10.1111/1755-0998.12487

Staats, M., Erkens, R.H.J., Van de Vossenberg, B., Wieringa, J.J., Kraaijeveld, K., Stielow, B., Geml, J., Richardson, J.E. & Bakker, F.T. 2013. Genomic treasure troves: Complete genome sequencing of herbarium and insect museum specimens. *PLoS ONE* 8: e69189. http://dx.doi.org/10.1371/journal.pone.0069189

Stamatakis, A., Hoover, P. & Rougemont, J. 2008. A rapid bootstrap algorithm for the RAxML web-servers. *Syst. Biol.* 75: 758–771. http://dx.doi.org/10.1080/10635150802429642

Telle, S. & Thines, M. 2008. Amplification of *cox2* (~620 bp) from 2 mg of up to 129 years old herbarium specimens, comparing 19 extraction methods and 15 polymerases. *PLoS ONE* 3: e3584. http://dx.doi.org/10.1371/journal.pone.0003584

Uribe-Convers, S., Settles, M.L. & Tank, D.C. 2016. A phylogenomic approach based on PCR target enrichment and high throughput sequencing: Resolving the diversity within the South American species of *Bartsia* L. (Orobanchaceae). *PLoS ONE* 11(2): e0148203. http://dx.doi.org/10.1371/journal.pone.0148203

Wang, Z., Gerstein, M. & Snyder, M. 2009. RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10: 57–63. http://dx.doi.org/10.1038/nrg2484

Zhou, L. & Holliday, J.A. 2012. Targeted enrichment of the black cottonwood (*Populus trichocarpa*) gene space using sequence capture. *B. M. C. Genomics* 13: 703. http://dx.doi.org/10.1186/1471-2164-13-703

Vol. 65 (5) • October 2016

TAXON

International Journal of Taxonomy, Phylogeny and Evolution

Electronic Supplement to

Retrieval of hundreds of nuclear loci from herbarium specimens

Michelle L. Hart, Laura L. Forrest, James A. Nicholls & Catherine A. Kidner

Taxon 65

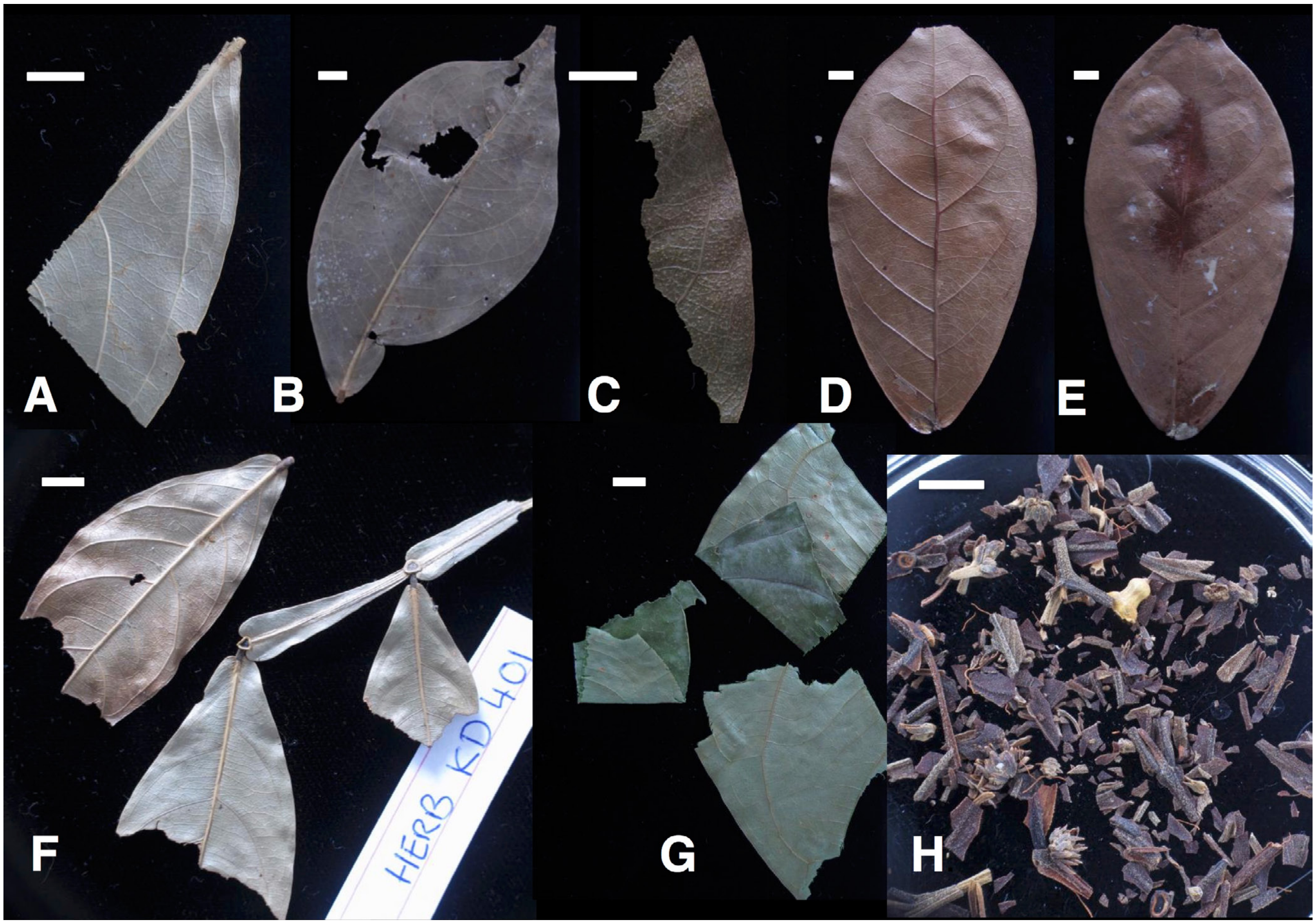

Fig. S1. *Inga umbellifera* herbarium tissue used for DNA extractions; white lines to left of plant material represent 5 mm scale bars. **A,** *Tello 2608* (E) (<5 ng of DNA was extracted from this material, so it was excluded from the rest of the study); **B,** *For. Dept. Brit. Guiana 5682* (K) 1948; **C,** *Matthews 1593* (E) 1835; **D & E,** *Hostmann 170* (K) 1841; **F,** *Dexter 401* (E) 2004; **G,** *Dexter 16L 145* (E) 2009; **H,** *Lawrance 260* (E) 1932.

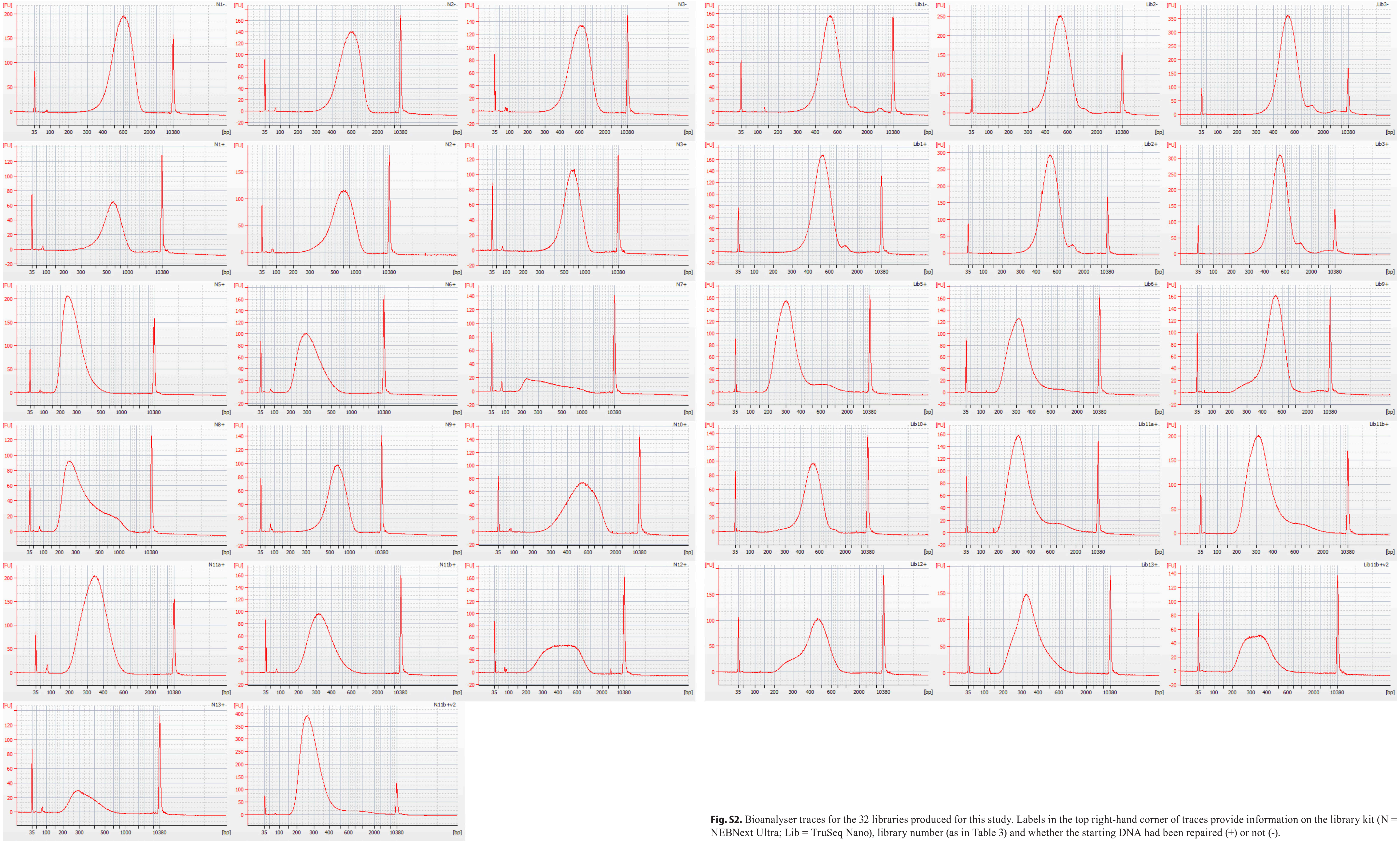

Fig. S2. Bioanalyser traces for the 32 libraries produced for this study. Labels in the top right-hand corner of traces provide information on the library kit (N = NEBNext Ultra; Lib = TruSeq Nano), library number (as in Table 3) and whether the starting DNA had been repaired (+) or not (-).

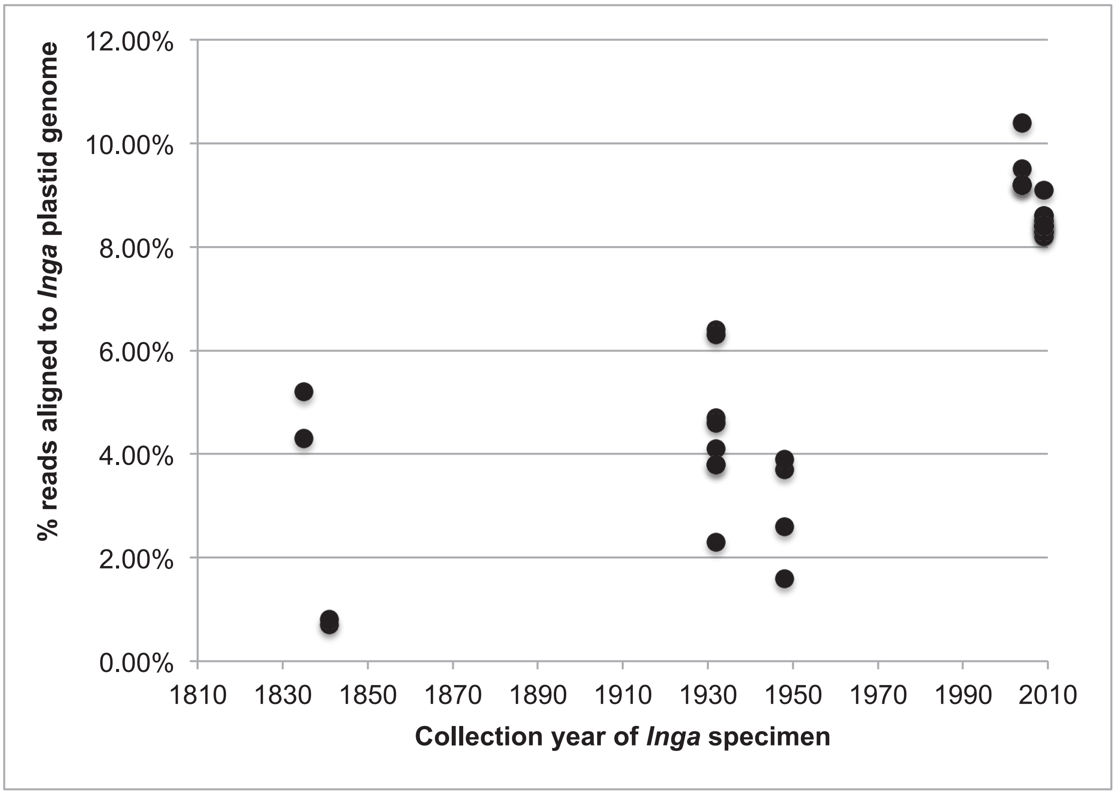

Fig. S3. Relationship between specimen collection date and the percentage of reads mapping to plastid after capture.

Table S1. Statistics for each individual bait alignment generated using AMAS.py (Borowiec, M.L. 2016. AMAS: A fast tool for alignment manipulation and computing of summary statistics. *PeerJ* 4:e1660).

Alignment_name	No. of taxa	Alignment length	Total matrix cells	Undetermined characters	Missing percent	No. variable sites	Proportion variable sites	Parsimony informative sites	Proportion parsimony informative	AT content	GC content	A	C	G	T	N	-
comp36444_c1_seq1	80	956	76480	8614	11,263	110	0,115	71	0,074	0,498	0,502	16269	19096	14959	17542	2809	5805
comp46465_c2_seq1	80	728	58240	2624	4,505	95	0,13	49	0,067	0,535	0,465	15040	12148	13715	14713	878	1746
comp546012_c0_seq1	80	686	54880	5800	10,569	95	0,138	46	0,067	0,467	0,533	11461	10878	15300	11441	4330	1470
comp46551_c0_seq1	80	1262	100960	18173	18	165	0,131	78	0,062	0,52	0,48	22723	18768	21004	20292	17129	1044
comp53516_c0_seq1	80	1267	101360	11971	11,81	172	0,136	75	0,059	0,624	0,376	30767	18543	15043	25036	9012	2959
comp44212_c0_seq1	80	687	54960	2016	3,668	67	0,098	40	0,058	0,571	0,429	12058	10356	12340	18190	1250	766
comp46465_c0_seq1	80	1024	81920	10236	12,495	117	0,114	59	0,058	0,512	0,488	18605	15496	19514	18069	3839	6397
comp338739_c0_seq1	80	1222	97760	22651	23,17	128	0,105	70	0,057	0,565	0,435	17759	13177	19497	24676	1427	21224
comp33962_c0_seq1	80	1423	113840	14304	12,565	137	0,096	79	0,056	0,571	0,429	28669	22715	19967	28185	9075	5229
comp45467_c0_seq1	80	1363	109040	5618	5,152	130	0,095	77	0,056	0,569	0,431	27481	20526	24065	31350	2661	2957
comp38281_c0_seq1	80	851	68080	1981	2,91	86	0,101	45	0,053	0,594	0,406	20375	10792	16053	18879	900	1081
comp50758_c0_seq4	80	243	19440	2032	10,453	27	0,111	13	0,053	0,645	0,355	3298	1675	4507	7928	1949	83
comp45038_c0_seq1	80	1252	100160	8321	8,308	157	0,125	65	0,052	0,553	0,447	27640	16926	24150	23123	6313	2008
comp35561_c0_seq1	80	468	37440	2613	6,979	55	0,118	24	0,051	0,582	0,418	9844	7202	7350	10431	2031	582
comp37377_c0_seq1	80	1739	139120	3309	2,379	158	0,091	88	0,051	0,539	0,461	34379	32336	30208	38888	1871	1438
comp41570_c0_seq1	80	1816	145280	12014	8,27	192	0,106	93	0,051	0,548	0,452	37368	28804	31413	35681	6894	5120
comp43290_c0_seq1	80	869	69520	2425	3,488	99	0,114	44	0,051	0,537	0,463	16905	15763	15332	19095	1265	1160
comp52736_c0_seq1	80	2628	210240	6880	3,272	267	0,102	130	0,049	0,543	0,457	56691	51619	41247	53803	3756	3124
comp45994_c1_seq1	80	537	42960	3352	7,803	64	0,119	25	0,047	0,473	0,527	8623	8782	12089	10114	1320	2032
comp53216_c0_seq2	80	1437	114960	38016	33,069	165	0,115	67	0,047	0,564	0,436	24991	14785	18776	18392	35010	3006
comp26820_c0_seq1	80	670	53600	14368	26,806	83	0,124	31	0,046	0,61	0,39	9505	6812	8471	14444	13465	903
comp42391_c0_seq1	80	1188	95040	4616	4,857	116	0,098	55	0,046	0,476	0,524	18752	23885	23454	24333	3434	1182
comp53952_c0_seq2	80	917	73360	1849	2,52	84	0,092	42	0,046	0,591	0,409	20681	11931	17315	21584	1441	408
comp55899_c0_seq1	80	1997	159760	6993	4,377	188	0,094	92	0,046	0,554	0,446	38121	34258	33900	46488	4038	2955
comp56254_c1_seq2	80	1425	114000	9681	8,492	109	0,076	66	0,046	0,59	0,41	29754	20511	22216	31838	5818	3863
comp46472_c0_seq1	80	1424	113920	7920	6,952	128	0,09	64	0,045	0,549	0,451	25949	19276	28527	32248	6469	1451
comp51314_c1_seq5	80	1110	88800	7918	8,917	126	0,114	50	0,045	0,586	0,414	20989	17742	15752	26399	7662	256
comp53881_c0_seq1	80	1408	112640	2874	2,551	135	0,096	63	0,045	0,503	0,497	26890	23679	30846	28351	2008	866
comp40581_c0_seq1	80	1246	99680	8704	8,732	102	0,082	55	0,044	0,551	0,449	23489	18453	22374	26660	8150	554

Table S1. Continued.

Alignment_name	No. of taxa	Alignment length	Total matrix cells	Undetermined characters	Missing percent	No. variable sites	Proportion variable sites	Parsimony informative sites	Proportion parsimony informative	AT content	GC content	A	C	G	T	N	-
comp40615_c0_seq1	80	833	66640	3412	5,12	75	0,09	37	0,044	0,568	0,432	19863	11386	15913	16066	2754	658
comp46489_c0_seq1	80	1563	125040	4721	3,776	154	0,099	67	0,043	0,537	0,463	30850	22265	33394	33810	3383	1338
comp37293_c0_seq1	80	1043	83440	5774	6,92	105	0,101	44	0,042	0,537	0,463	21485	19990	16006	20185	4757	1017
comp51262_c0_seq4	80	1208	96640	2309	2,389	115	0,095	51	0,042	0,575	0,425	27137	21754	18349	27091	652	1657
comp52829_c0_seq3	80	620	49600	11996	24,185	47	0,076	26	0,042	0,612	0,388	10796	6954	7640	12214	11687	309
comp43316_c0_seq1	80	657	52560	1486	2,827	58	0,088	27	0,041	0,622	0,378	15778	8937	10389	15970	701	785
comp43866_c0_seq1	80	1272	101760	5168	5,079	103	0,081	52	0,041	0,503	0,497	24928	20610	27423	23631	1142	4026
comp46025_c0_seq1	80	1207	96560	15413	15,962	97	0,08	49	0,041	0,476	0,524	18658	14257	28260	19972	13324	2089
comp46553_c1_seq1	80	1309	104720	15156	14,473	125	0,095	54	0,041	0,526	0,474	23283	19472	22967	23842	13843	1313
comp49929_c0_seq1	80	1591	127280	4610	3,622	140	0,088	65	0,041	0,492	0,508	31976	26523	35737	28434	2634	1976
comp51482_c0_seq2	80	2166	173280	2536	1,464	232	0,107	89	0,041	0,55	0,45	50955	37062	39849	42878	1615	921
comp37261_c0_seq1	80	1288	103040	98713	95,801	54	0,042	51	0,04	0,514	0,486	1195	605	1497	1030	38326	60387
comp39051_c0_seq2	80	227	18160	6255	34,444	13	0,057	9	0,04	0,5	0,5	2119	2724	3224	3838	5777	478
comp43423_c0_seq1	80	1402	112160	12375	11,033	118	0,084	56	0,04	0,504	0,496	25014	21077	28398	25296	10012	2363
comp49673_c0_seq1	80	1544	123520	17254	13,969	152	0,098	61	0,04	0,558	0,442	27277	23030	23952	32007	15672	1582
comp51236_c3_seq1	80	1958	156640	15176	9,688	194	0,099	79	0,04	0,57	0,43	38432	25751	35019	42262	12611	2565
comp41024_c0_seq1	80	746	59680	2087	3,497	62	0,083	29	0,039	0,561	0,439	17098	12391	12908	15196	1296	791
comp53167_c1_seq1	80	1744	139520	8598	6,163	134	0,077	68	0,039	0,541	0,459	35124	26770	33324	35704	4057	4541
comp55300_c0_seq3	80	1469	117520	20381	17,343	115	0,078	57	0,039	0,603	0,397	29046	17451	21148	29494	18480	1901
comp42737_c0_seq1	80	2152	172160	5280	3,067	211	0,098	82	0,038	0,548	0,452	40657	35629	39816	50778	4724	556
comp44609_c0_seq1	80	1216	97280	5906	6,071	96	0,079	46	0,038	0,457	0,543	20367	24319	25270	21418	3849	2057
comp48218_c0_seq1	80	1000	80000	13443	16,804	97	0,097	38	0,038	0,59	0,41	18376	12848	14416	20917	12388	1055
comp56258_c0_seq1	80	1578	126240	36409	28,841	121	0,077	60	0,038	0,518	0,482	21333	22525	20759	25214	34299	2110
comp41081_c0_seq1	80	1163	93040	25446	27,35	112	0,096	43	0,037	0,568	0,432	20500	15727	13491	17876	20060	5386
comp42274_c0_seq1	80	836	66880	12782	19,112	62	0,074	31	0,037	0,593	0,407	17549	8747	13284	14518	11790	992
comp45684_c0_seq1	80	2321	185680	12072	6,502	197	0,085	86	0,037	0,539	0,461	43947	36503	43564	49594	11051	1021
comp39487_c0_seq1	80	726	58080	20583	35,439	44	0,061	26	0,036	0,596	0,404	8553	7026	8125	13793	19958	625
comp53141_c1_seq1	80	2419	193520	5674	2,932	224	0,093	86	0,036	0,501	0,499	45560	42192	51492	48602	5061	613
comp37145_c0_seq1	80	1353	108240	6417	5,928	131	0,097	48	0,035	0,539	0,461	25080	20412	26504	29827	5705	712

Table S1. Continued.

Alignment_name	No. of taxa	Alignment length	Total matrix cells	Undetermined characters	Missing percent	No. variable sites	Proportion variable sites	Parsimony informative sites	Proportion parsimony informative	AT content	GC content	A	C	G	T	N	-
comp41473_c0_seq1	80	1063	85040	1099	1,292	74	0,07	37	0,035	0,533	0,467	17318	21894	17326	27403	938	161
comp46497_c0_seq2	80	1649	131920	9937	7,533	151	0,092	57	0,035	0,561	0,439	31463	25127	28364	37029	7901	2036
comp48873_c0_seq1	80	1570	125600	13349	10,628	132	0,084	55	0,035	0,568	0,432	27211	22546	25955	36539	12420	929
comp44856_c1_seq1	80	379	30320	491	1,619	29	0,077	13	0,034	0,585	0,415	8851	6953	5414	8611	271	220
comp45170_c0_seq2	80	933	74640	3782	5,067	84	0,09	32	0,034	0,566	0,434	18381	14389	16363	21725	3521	261
comp48120_c0_seq1	80	1275	102000	6410	6,284	110	0,086	43	0,034	0,429	0,571	18149	28672	25898	22871	3402	3008
comp50657_c0_seq1	80	1243	99440	40139	40,365	97	0,078	42	0,034	0,583	0,417	17007	13986	10741	17567	38356	1783
comp53340_c0_seq1	80	1349	107920	3764	3,488	105	0,078	46	0,034	0,49	0,51	24375	25514	27623	26644	2543	1221
comp55228_c0_seq8	80	812	64960	1568	2,414	71	0,087	28	0,034	0,622	0,378	16217	11556	12435	23184	347	1221
comp56609_c0_seq1	80	2677	214160	48972	22,867	209	0,078	90	0,034	0,636	0,364	53660	30731	29424	51373	39737	9235
comp27375_c0_seq1	80	242	19360	1249	6,451	16	0,066	8	0,033	0,569	0,431	4550	2781	5022	5758	926	323
comp44391_c1_seq1	80	552	44160	2394	5,421	45	0,082	18	0,033	0,586	0,414	11602	8470	8801	12893	2357	37
comp49109_c0_seq4	80	840	67200	8726	12,985	75	0,089	28	0,033	0,533	0,467	17971	14247	13044	13212	6529	2197
comp49588_c0_seq1	80	1329	106320	4152	3,905	106	0,08	44	0,033	0,57	0,43	26517	23656	20276	31719	1908	2244
comp53904_c0_seq1	80	1313	105040	3152	3,001	94	0,072	43	0,033	0,559	0,441	28218	20581	24358	28731	1801	1351
comp27897_c0_seq1	80	1251	100080	14924	14,912	91	0,073	40	0,032	0,514	0,486	20389	17563	23845	23359	14057	867
comp43766_c1_seq1	80	462	36960	1018	2,754	29	0,063	15	0,032	0,42	0,58	6883	8627	12212	8220	626	392
comp45125_c0_seq2	80	820	65600	4776	7,28	42	0,051	26	0,032	0,466	0,534	13733	13044	19417	14630	3487	1289
comp51566_c0_seq1	80	2346	187680	8656	4,612	192	0,082	75	0,032	0,525	0,475	37168	35405	49698	56753	8048	608
comp56887_c0_seq1	80	2561	204880	9698	4,734	167	0,065	81	0,032	0,563	0,437	54592	38615	46736	55239	9562	136
comp43819_c0_seq1	80	1583	126640	34801	27,48	145	0,092	49	0,031	0,524	0,476	23097	22364	21333	25045	26679	8122
comp45862_c0_seq1	80	491	39280	1094	2,785	44	0,09	15	0,031	0,636	0,364	11354	7430	6455	12947	817	277
comp51289_c0_seq1	80	1385	110800	20266	18,291	103	0,074	43	0,031	0,584	0,416	24169	15187	22476	28702	19697	569
comp53352_c0_seq1	80	1690	135200	4392	3,249	132	0,078	53	0,031	0,586	0,414	35935	28623	25499	40751	1545	2847
comp55182_c0_seq2	80	751	60080	7874	13,106	47	0,063	23	0,031	0,594	0,406	15519	8535	12686	15466	7154	720
comp46048_c0_seq1	80	797	63760	16875	26,466	44	0,055	24	0,03	0,586	0,414	12078	9019	10371	15417	16052	823
comp46275_c0_seq1	80	1350	108000	9103	8,429	90	0,067	40	0,03	0,541	0,459	23659	17985	27404	29849	8925	178
comp55873_c0_seq1	80	1456	116480	66384	56,992	84	0,058	43	0,03	0,554	0,446	12182	10830	11527	15557	64991	1393
comp42358_c0_seq1	80	1529	122320	20034	16,378	100	0,065	44	0,029	0,569	0,431	28058	21667	22411	30150	19856	178

Table S1. Continued.

Alignment_name	No. of taxa	Alignment length	Total matrix cells	Undetermined characters	Missing percent	No. variable sites	Proportion variable sites	Parsimony informative sites	Proportion parsimony informative	AT content	GC content	A	C	G	T	N	-
comp46036_c0_seq1	80	1101	88080	2209	2,508	77	0,07	32	0,029	0,589	0,411	23525	15381	19883	27082	1178	1031
comp48566_c2_seq1	80	1498	119840	1926	1,607	105	0,07	44	0,029	0,527	0,473	30017	24831	30981	32085	1564	362
comp49294_c0_seq2	80	1173	93840	2086	2,223	100	0,085	34	0,029	0,527	0,473	22759	19683	23718	25594	1521	565
comp52711_c0_seq1	80	1486	118880	11533	9,701	114	0,077	43	0,029	0,534	0,466	29717	22040	27944	27646	10705	828
comp54484_c0_seq1	80	1329	106320	25479	23,964	100	0,075	39	0,029	0,528	0,472	21519	17217	20929	21176	24392	1087
comp53194_c2_seq3	80	1881	150480	16912	11,239	104	0,055	53	0,028	0,608	0,392	41399	22289	30008	39872	15437	1475
comp826804_c0_seq1	80	1076	86080	1316	1,529	63	0,059	30	0,028	0,521	0,479	21285	17361	23240	22878	949	367
comp23076_c0_seq1	80	1273	101840	1784	1,752	73	0,057	35	0,027	0,597	0,403	27079	21095	19253	32629	1393	391
comp42706_c0_seq1	80	1212	96960	2733	2,819	105	0,087	33	0,027	0,506	0,494	26582	22150	24428	21067	2733	0
comp43262_c0_seq1	80	1204	96320	2159	2,241	81	0,067	33	0,027	0,497	0,503	21724	24709	22630	25098	952	1207
comp44802_c0_seq1	80	1385	110800	27728	25,025	99	0,071	38	0,027	0,624	0,376	23911	14300	16941	27920	26211	1517
comp49395_c0_seq1	80	1582	126560	21297	16,828	89	0,056	43	0,027	0,57	0,43	30320	20916	24370	29657	19958	1339
comp51015_c0_seq1	80	1729	138320	8255	5,968	126	0,073	46	0,027	0,529	0,471	33230	26948	34321	35566	7727	528
comp54662_c0_seq1	80	1577	126160	5436	4,309	111	0,07	43	0,027	0,555	0,445	33330	23271	30490	33633	5196	240
comp55034_c0_seq4	80	995	79600	32448	40,764	47	0,047	27	0,027	0,597	0,403	14287	7717	11283	13865	28062	4386
comp27108_c0_seq1	80	500	40000	8905	22,262	27	0,054	13	0,026	0,567	0,433	8970	6542	6916	8667	8089	816
comp36900_c2_seq1	80	385	30800	1292	4,195	20	0,052	10	0,026	0,494	0,506	7457	5451	9474	7126	336	956
comp41267_c0_seq2	80	704	56320	13109	23,276	48	0,068	18	0,026	0,501	0,499	9447	9996	11572	12196	12090	1019
comp43995_c0_seq2	80	1116	89280	9637	10,794	65	0,058	29	0,026	0,556	0,444	23851	15540	19831	20421	9213	424
comp44503_c0_seq2	80	935	74800	13229	17,686	60	0,064	24	0,026	0,542	0,458	14868	13623	14602	18478	11214	2015
comp50626_c0_seq4	80	1034	82720	1967	2,378	77	0,074	27	0,026	0,474	0,526	17626	20721	21718	20688	1044	923
comp52915_c0_seq2	80	1520	121600	3945	3,244	70	0,046	40	0,026	0,6	0,4	34585	20866	26141	36063	3495	450
comp56733_c0_seq5	80	1689	135120	37618	27,84	121	0,072	44	0,026	0,583	0,417	31317	21375	19277	25533	33744	3874
comp28617_c0_seq1	80	833	66640	1770	2,656	49	0,059	21	0,025	0,536	0,464	18152	12300	17824	16594	930	840
comp28839_c0_seq1	80	844	67520	4387	6,497	64	0,076	21	0,025	0,564	0,436	18432	12321	15180	17200	2633	1754
comp30607_c0_seq1	80	836	66880	26596	39,767	55	0,066	21	0,025	0,58	0,42	12359	7447	9474	11004	25754	842
comp36697_c0_seq1	80	1414	113120	5494	4,857	99	0,07	35	0,025	0,577	0,423	29095	25509	20000	33022	3658	1836
comp37644_c0_seq1	80	1058	84640	5904	6,975	82	0,078	26	0,025	0,527	0,473	18482	17769	19458	23027	614	5290
comp42557_c0_seq1	80	1614	129120	16368	12,677	113	0,07	41	0,025	0,564	0,436	30124	24096	25069	33463	15806	562

Table S1. Continued.

Alignment_name	No. of taxa	Alignment length	Total matrix cells	Undetermined characters	Missing percent	No. variable sites	Proportion variable sites	Parsimony informative sites	Proportion parsimony informative	AT content	GC content	A	C	G	T	N	-
comp44627_c0_seq1	80	1092	87360	7506	8,592	71	0,065	27	0,025	0,463	0,537	17889	21744	21175	19046	6255	1251
comp48019_c0_seq2	80	1467	117360	27881	23,757	107	0,073	37	0,025	0,514	0,486	20700	22206	21277	25296	23938	3943
comp53028_c0_seq2	80	2051	164080	29364	17,896	132	0,064	52	0,025	0,523	0,477	32993	32941	31329	37453	26639	2725
comp53595_c0_seq20	80	1344	107520	40896	38,036	94	0,07	34	0,025	0,581	0,419	19934	11930	15956	18804	39620	1276
comp39600_c0_seq2	80	1139	91120	4825	5,295	72	0,063	27	0,024	0,555	0,445	22708	21029	17351	25207	4070	755
comp42265_c0_seq1	80	1712	136960	65776	48,026	92	0,054	41	0,024	0,595	0,405	21060	14039	14796	21289	62220	3556
comp44244_c0_seq1	80	2126	170080	52828	31,061	122	0,057	52	0,024	0,596	0,404	32362	22708	24619	37563	52143	685
comp49523_c1_seq1	80	1120	89600	3197	3,568	76	0,068	27	0,024	0,544	0,456	24831	20611	18804	22157	1220	1977
comp53769_c0_seq1	80	1726	138080	41011	29,701	118	0,068	42	0,024	0,564	0,436	29757	20312	21963	25037	38572	2439
comp56192_c1_seq2	80	739	59120	2436	4,12	38	0,051	18	0,024	0,606	0,394	14690	12119	10201	19674	2436	0
comp39332_c0_seq1	80	683	54640	17027	31,162	32	0,047	16	0,023	0,593	0,407	9463	6803	8488	12859	16546	481
comp41658_c0_seq1	80	1416	113280	52569	46,406	79	0,056	32	0,023	0,56	0,44	18223	13284	13423	15781	51246	1323
comp46261_c0_seq1	80	1754	140320	29556	21,063	127	0,072	40	0,023	0,512	0,488	30618	22968	31084	26094	28682	874
comp47751_c0_seq1	80	1696	135680	2943	2,169	94	0,055	39	0,023	0,574	0,426	36740	26340	30153	39504	1972	971
comp48510_c0_seq6	80	511	40880	2768	6,771	39	0,076	12	0,023	0,557	0,443	9321	7769	9129	11893	2010	758
comp49386_c0_seq1	80	1668	133440	28839	21,612	123	0,074	38	0,023	0,526	0,474	29193	25127	24439	25842	28378	461
comp50170_c0_seq12	80	902	72160	17585	24,369	58	0,064	21	0,023	0,555	0,445	14330	9153	15146	15946	16933	652
comp51335_c0_seq1	80	1591	127280	60673	47,669	79	0,05	37	0,023	0,619	0,381	19828	11846	13519	21414	57464	3209
comp52686_c0_seq2	80	2020	161600	82184	50,856	94	0,047	46	0,023	0,612	0,388	20315	12015	18800	28286	79927	2257
comp53258_c0_seq3	80	1630	130400	41421	31,765	90	0,055	37	0,023	0,538	0,462	22132	21788	19353	25706	38341	3080
comp54299_c0_seq1	80	1490	119200	54814	45,985	72	0,048	34	0,023	0,606	0,394	17616	11138	14206	21426	52560	2254
comp37141_c0_seq1	80	312	24960	528	2,115	17	0,054	7	0,022	0,574	0,426	8008	4961	5458	6005	255	273
comp46121_c0_seq1	80	894	71520	20332	28,428	67	0,075	20	0,022	0,607	0,393	14334	9638	10460	16756	19969	363
comp46351_c1_seq1	80	1092	87360	2044	2,34	78	0,071	24	0,022	0,437	0,563	19706	18201	29873	17536	1764	280
comp49369_c0_seq1	80	2158	172640	3879	2,247	154	0,071	48	0,022	0,56	0,44	41906	30832	43420	52603	3070	809
comp52112_c0_seq3	80	1674	133920	2072	1,547	87	0,052	36	0,022	0,572	0,428	36761	30337	26149	38601	947	1125
comp54031_c0_seq1	80	1748	139840	7231	5,171	90	0,051	39	0,022	0,533	0,467	33329	32607	29340	37333	5877	1354
comp55479_c0_seq1	80	1626	130080	15858	12,191	104	0,064	36	0,022	0,578	0,422	34562	23873	24317	31470	15690	168
comp710440_c0_seq1	80	543	43440	3235	7,447	32	0,059	12	0,022	0,52	0,48	10081	10255	9060	10809	1433	1802

Table S1. Continued.

Alignment_name	No. of taxa	Alignment length	Total matrix cells	Undetermined characters	Missing percent	No. variable sites	Proportion variable sites	Parsimony informative sites	Proportion parsimony informative	AT content	GC content	A	C	G	T	N	-
comp41859_c0_seq1	80	1308	104640	40256	38,471	55	0,042	28	0,021	0,523	0,477	19234	12323	18412	14415	39987	269
comp42092_c0_seq1	80	1191	95280	21526	22,592	79	0,066	25	0,021	0,552	0,448	20546	17244	15790	20174	20334	1192
comp42703_c0_seq1	80	1478	118240	1398	1,182	94	0,064	31	0,021	0,604	0,396	34210	21157	25147	36328	1398	0
comp52981_c1_seq1	80	2801	224080	29171	13,018	154	0,055	60	0,021	0,586	0,414	51630	40494	40166	62619	29156	15
comp53604_c0_seq2	80	1594	127520	6190	4,854	74	0,046	34	0,021	0,485	0,515	29191	28823	33706	29610	4596	1594
comp53857_c0_seq1	80	2461	196880	40262	20,45	129	0,052	52	0,021	0,55	0,45	41507	34415	36089	44607	36515	3747
comp54142_c0_seq1	80	3393	271440	20212	7,446	212	0,062	71	0,021	0,584	0,416	75540	52424	51980	71284	19730	482
comp44153_c0_seq1	80	935	74800	36444	48,722	57	0,061	19	0,02	0,573	0,427	10524	7312	9070	11450	35269	1175
comp45867_c0_seq1	80	1354	108320	8348	7,707	82	0,061	27	0,02	0,574	0,426	28015	18619	23946	29392	6941	1407
comp53451_c0_seq1	80	1737	138960	65374	47,045	67	0,039	34	0,02	0,598	0,402	20921	14303	15260	23102	64329	1045
comp55651_c0_seq1	80	1787	142960	6114	4,277	141	0,079	35	0,02	0,583	0,417	37749	26197	30801	42099	5120	994
comp43405_c0_seq1	80	1492	119360	4530	3,795	93	0,062	29	0,019	0,576	0,424	29732	24575	24129	36394	4007	523
comp49023_c0_seq1	80	1605	128400	28240	21,994	116	0,072	31	0,019	0,59	0,41	25441	15876	25222	33621	28236	4
comp56022_c2_seq1	80	1809	144720	89562	61,886	79	0,044	34	0,019	0,605	0,395	19050	9171	12614	14323	83751	5811
comp56474_c0_seq2	80	2547	203760	138357	67,902	91	0,036	48	0,019	0,6	0,4	17343	11426	14749	21885	134876	3481
comp56022_c2_seq1	80	1809	144720	89562	61,886	79	0,044	34	0,019	0,605	0,395	19050	9171	12614	14323	83751	5811
comp41758_c0_seq1	80	1451	116080	28784	24,797	67	0,046	26	0,018	0,594	0,406	28554	17296	18143	23303	27039	1745
comp44887_c0_seq1	80	1568	125440	22313	17,788	91	0,058	29	0,018	0,521	0,479	23351	19944	29416	30416	21968	345
comp53688_c0_seq1	80	2605	208400	8701	4,175	144	0,055	47	0,018	0,593	0,407	58604	32864	48450	59781	6757	1944
comp31780_c0_seq1	80	1861	148880	75956	51,018	75	0,04	32	0,017	0,496	0,504	17751	16344	20378	18451	74746	1210
comp43079_c0_seq1	80	1599	127920	58898	46,043	73	0,046	27	0,017	0,559	0,441	20995	13265	17180	17582	55104	3794
comp43779_c0_seq1	80	1428	114240	1691	1,48	67	0,047	24	0,017	0,584	0,416	29770	20885	25928	35966	1039	652
comp46303_c0_seq1	80	1615	129200	61125	47,31	89	0,055	27	0,017	0,544	0,456	18460	12541	18468	18606	59351	1774
comp46343_c0_seq1	80	233	18640	818	4,388	16	0,069	4	0,017	0,568	0,432	5009	2865	4827	5121	716	102
comp49083_c0_seq1	80	1757	140560	46195	32,865	86	0,049	30	0,017	0,563	0,437	26436	21860	19333	26736	44547	1648
comp50204_c0_seq1	80	2761	220880	38527	17,443	143	0,052	48	0,017	0,557	0,443	44838	32030	48727	56758	37610	917
comp53540_c0_seq3	80	1725	138000	2687	1,947	93	0,054	29	0,017	0,597	0,403	36388	30053	24536	44336	2305	382
comp53279_c0_seq1	80	897	71760	46635	64,987	30	0,033	14	0,016	0,66	0,34	7486	3184	5370	9085	45998	637
comp53978_c0_seq1	80	1381	110480	64921	58,763	44	0,032	22	0,016	0,618	0,382	12738	9658	7759	15404	63600	1321

Table S1. Continued.

Alignment_name	No. of taxa	Alignment length	Total matrix cells	Undetermined characters	Missing percent	No. variable sites	Proportion variable sites	Parsimony informative sites	Proportion parsimony informative	AT content	GC content	A	C	G	T	N	-
comp55820_c0_seq1	80	1610	128800	53738	41,722	67	0,042	26	0,016	0,596	0,404	22940	14437	15887	21798	52966	772
comp55885_c0_seq14	80	737	58960	33746	57,235	35	0,047	12	0,016	0,649	0,351	7337	3664	5184	9029	32618	1128
comp36008_c0_seq1	80	584	46720	11452	24,512	41	0,07	9	0,015	0,588	0,412	8971	6915	7611	11771	10640	812
comp53459_c0_seq1	80	1579	126320	84353	66,777	54	0,034	24	0,015	0,567	0,433	10303	8054	10136	13474	81470	2883
comp56903_c0_seq3	80	1944	155520	57549	37,004	70	0,036	29	0,015	0,556	0,444	31283	22126	21350	23212	54997	2552
comp1160393_c0_seq1	80	434	34720	5112	14,724	21	0,048	6	0,014	0,591	0,409	7052	5722	6401	10433	4728	384
comp40678_c0_seq1	80	662	52960	14581	27,532	45	0,068	9	0,014	0,611	0,389	8792	5891	9027	14669	13428	1153
comp46880_c0_seq1	80	1526	122080	42606	34,9	77	0,05	22	0,014	0,528	0,472	22871	16222	21296	19085	41948	658
comp52492_c1_seq2	80	1251	100080	50661	50,621	59	0,047	18	0,014	0,629	0,371	12883	7784	10554	18198	49706	955
comp54453_c0_seq1	80	2682	214560	105247	49,052	112	0,042	38	0,014	0,563	0,437	30853	24970	22837	30653	102028	3219
comp28983_c0_seq1	80	2294	183520	92222	50,252	57	0,025	29	0,013	0,638	0,362	26392	19630	13436	31840	88213	4009
comp30427_c0_seq1	80	1209	96720	59071	61,074	33	0,027	16	0,013	0,607	0,393	11877	6047	8744	10981	58948	123
comp51118_c0_seq1	80	3390	271200	156930	57,865	125	0,037	45	0,013	0,542	0,458	29235	24269	28077	32689	155739	1191
comp53279_c1_seq1	80	753	60240	37848	62,829	22	0,029	10	0,013	0,569	0,431	5794	4398	5245	6955	36298	1550
comp40592_c0_seq5	80	519	41520	4791	11,539	26	0,05	6	0,012	0,478	0,522	9796	7920	11267	7746	4068	723
comp48138_c0_seq1	80	1330	106400	66100	62,124	44	0,033	16	0,012	0,572	0,428	11267	8213	9023	11797	64770	1330
comp53868_c0_seq1	80	673	53840	746	1,386	29	0,043	8	0,012	0,506	0,494	10944	14380	11852	15918	310	436
comp54072_c0_seq2	80	2158	172640	77969	45,163	84	0,039	25	0,012	0,584	0,416	26299	16030	23372	28970	75808	2161
comp1072929_c0_seq1	80	271	21680	938	4,327	6	0,022	3	0,011	0,504	0,496	4797	3279	7006	5660	759	179
comp41904_c0_seq1	80	1387	110960	34777	31,342	67	0,048	15	0,011	0,595	0,405	23707	14382	16478	21616	34669	108
comp46777_c0_seq1	80	1004	80320	50409	62,76	27	0,027	11	0,011	0,597	0,403	8746	5848	6203	9114	49618	791
comp50371_c0_seq3	80	1372	109760	3005	2,738	57	0,042	15	0,011	0,58	0,42	30563	18450	26418	31324	2927	78
comp51757_c0_seq1	80	1679	134320	97880	72,871	52	0,031	19	0,011	0,587	0,413	10145	7885	7149	11261	94759	3121
comp43766_c2_seq1	80	585	46800	22185	47,404	25	0,043	6	0,01	0,482	0,518	6791	5766	6993	5065	19627	2558
comp50161_c0_seq1	80	629	50320	15301	30,407	8	0,013	6	0,01	0,614	0,386	10716	8084	5422	10797	14636	665
comp55210_c1_seq1	80	1156	92480	5801	6,273	25	0,022	12	0,01	0,524	0,476	26187	19357	21875	19260	4651	1150
comp40970_c0_seq1	80	583	46640	28982	62,14	15	0,026	5	0,009	0,672	0,328	5475	1875	3916	6392	28287	695
comp41589_c0_seq1	80	926	74080	28451	38,406	35	0,038	8	0,009	0,573	0,427	12874	8472	11033	13250	27325	1126
comp51608_c0_seq1	80	2767	221360	143688	64,911	81	0,029	25	0,009	0,612	0,388	21903	14534	15620	25615	142493	1195

Table S1. Continued.

Alignment_name	No. of taxa	Alignment length	Total matrix cells	Undetermined characters	Missing percent	No. variable sites	Proportion variable sites	Parsimony informative sites	Proportion parsimony informative	AT content	GC content	A	C	G	T	N	-
comp54430_c2_seq4	80	1521	121680	80609	66,247	41	0,027	12	0,008	0,555	0,445	10384	8399	9868	12420	58995	21614
comp39192_c0_seq1	80	615	49200	25482	51,793	23	0,037	4	0,007	0,433	0,567	5408	4068	9381	4861	25101	381
comp46270_c0_seq1	80	1808	144640	39841	27,545	49	0,027	11	0,006	0,537	0,463	31143	24569	23985	25102	37963	1878
comp52180_c0_seq2	80	1210	96800	71688	74,058	25	0,021	7	0,006	0,572	0,428	6516	5177	5570	7849	18245	53443
comp47631_c0_seq1	80	1646	131680	84199	63,942	27	0,016	9	0,005	0,484	0,516	9273	11403	13093	13712	81544	2655
comp56397_c0_seq2	80	2768	221440	186837	84,374	39	0,014	12	0,004	0,645	0,355	12064	5891	6406	10242	183602	3235
comp56747_c1_seq2	80	535	42800	29941	69,956	9	0,017	2	0,004	0,508	0,492	2169	3394	2936	4360	29381	560
comp39985_c0_seq4	80	718	57440	51391	89,469	3	0,004	1	0,001	0,559	0,441	1569	1395	1270	1815	50003	1388
comp1585458_c0_seq1	80	562	44960	44390	98,732	0	0	0	0	0,584	0,416	171	147	90	162	21136	23254
comp36654_c0_seq1	80	161	12880	12276	95,311	0	0	0	0	0,583	0,417	136	60	192	216	12276	0
comp49874_c0_seq1	80	221	17680	17550	99,265	0	0	0	0	0,638	0,362	25	17	30	58	17550	0