MOLECULAR ECOLOGY RESOURCES

Molecular Ecology Resources (2014) doi: 10.1111/1755-0998.12258

Cost-effective enrichment hybridization capture of chloroplast genomes at deep multiplexing levels for population genetics and phylogeography studies

CÉDRIC MARIAC,* NORA SCARCELLI,* JULIETTE POUZADOU,* ADELINE BARNAUD,*†‡ CLAIRE BILLOT,§ ADAMA FAYE,*¶ AYITE KOUGBEADJO,* VINCENT MAILLOL,** GUILLAUME MARTIN,§ FRANÇOIS SABOT,* SYLVAIN SANTONI,** YVES VIGOUROUX* and THOMAS L. P. COUVREUR*¶

*Institut de Recherche pour le Développement, UMR DIADE, BP 64501, 34394 Montpellier, France, †Laboratoire National de Recherche sur les Productions Végétales, Institut Sénégalais de Recherches Agricoles, Centre de Recherche de Bel Air, Dakar, Senegal, ‡Laboratoire mixte international Adaptation des Plantes et microorganismes associés aux Stress Environnementaux, Institut de Recherche pour le Développement/Institut Sénégalais de Recherches Agricoles/Université Cheikh Anta Diop, Centre de Recherche de Bel Air, Dakar, Senegal, §UMR AGAP, CIRAD, F-34398 Montpellier, France, ¶Université de Yaoundé I, Ecole Normale Supérieure, Département des Sciences Biologiques, Laboratoire de Botanique, Systématique et d'Ecologie, B.P.047, Yaoundé, Cameroon, **UMR AGAP, Equipe Diversité et Adaptation de la Vigne et des Espèces Méditerranéennes, INRA, 2 Place Viala, 34060 Montpellier, France

Abstract

Biodiversity, phylogeography and population genetic studies will be revolutionized by access to large data sets thanks to next-generation sequencing methods. In this study, we develop an easy and cost-effective protocol for in-solution enrichment hybridization capture of complete chloroplast genomes applicable at deep-multiplexed levels. The protocol uses cheap in-house species-specific probes developed via long-range PCR of the entire chloroplast. Barcoded libraries are constructed, and in-solution enrichment of the chloroplasts is carried out using the probes. This protocol was tested and validated on six economically important West African crop species, namely African rice, pearl millet, three African yam species and fonio. For pearl millet, we also demonstrate the effectiveness of this protocol to retrieve 95% of the sequence of the whole chloroplast on 95 multiplexed individuals in a single MiSeq run at a success rate of 95%. This new protocol allows whole chloroplast genomes to be retrieved at a modest cost and will allow unprecedented resolution for closely related species in phylogeography studies using plastomes.

Keywords: DNA probes, long-range PCR, MiSeq, next-generation sequencing, plastomes, whole chloroplast sequencing

Received 20 December 2013; revision received 7 March 2014; accepted 17 March 2014

Introduction

Understanding the evolutionary dynamics of natural species provides important insights into speciation, extinction and colonization of populations through time leading to a better knowledge of species reactions to past and future climatic changes (Avise 2000). For cultivated plants and their wild relatives, such knowledge helps to understand domestication processes as well as the origin and spread of crops across the world. The detection and analyses of genetic variation play a central role in such studies. Next-generation sequencing (NGS) is revolutionizing research in plant evolution and genetic diversity allowing for the generation of large quantities of sequence data in cost-effective ways (Varshney *et al.* 2009; McCormack *et al.* 2013).

Complete chloroplast genomes, or plastomes, have been used to infer plant relationships at interfamilial (phylogenetics, Barrett *et al.* 2013; Moore *et al.* 2007; Parks *et al.* 2009; Straub *et al.* 2012; Stull *et al.* 2013) and interspecific levels (Parks *et al.* 2009; Yang *et al.* 2013; Bock *et al.* 2014) providing good levels of variability. Chloroplast data can provide important insights into population genetic or phylogeography studies. Indeed, the plastome is straight forward to use being nonrecombinant, generating few problems of paralogy, uniparentally inherited allowing to trace in most cases seed-mediated lineages and generally presents a short

Correspondence: Cedric Mariac, Yves Vigouroux and Thomas Couvreur Fax: 33-0-467416222; E-mails: cedric.mariac@ird.fr, yves.vigouroux@ird.fr and thomas.couvreur@ird.fr

coalescent time (Petit & Vendramin 2007). However, because of the perception of low variability for this organelle, its use at infraspecies levels has lagged behind.

Data obtained from the comparative analyses of full chloroplast sequences within species are starting to overturn the concept of low infraspecific chloroplast variability at the genomic scale (Whittall *et al.* 2010; Besnard *et al.* 2011; Kane *et al.* 2012). The full chloroplast has also recently been suggested as a ultra-barcode for plants compartments (identification of agronomical varieties), underlining its variability at the genomic level (Kane *et al.* 2012). The general conclusion of these studies is that variability of chloroplast DNA (cpDNA) cannot be concluded from a few regions (e.g. *rbcL*) and that sequencing the whole genome will undoubtedly be the best way to exploit this resource to completion.

Although sequencing the whole chloroplast genome is feasible, its application to large numbers of individuals in a single sequencing pool of an NGS technology (e.g. MiSeq) remains technically challenging. Indeed, the ability to multiplex individuals into a single run is of central importance for population genetics and phylogeography where sample numbers are often in the hundreds. Genome-skimming methods (where total genomic DNA is directly sequenced generally leading to full chloroplast coverage) do allow for limited multiplexing of the chloroplast genomes (Zhang *et al.* 2011; Kane *et al.* 2012; Straub *et al.* 2012; Bock *et al.* 2014). In a recent study, Bock *et al.* (2014) sequenced 34 plastomes of the Jerusalem Artichoke tuber crop species in a single HiSeq 2000 lane without prior enrichment. This resulted in overall good coverage depth for the plastome for all individuals, in addition to other nuclear, mitochondrial and ribosomal data. However, genome-skimming approaches to plastome sequencing are conditioned by the ratio between chloroplast and nuclear DNA which can vary manifold between species (e.g. Rauwolf *et al.* 2010). Thus, genome-skimming methods to retrieve plastome information provide a limited option for population genetics or phylogeography studies that require deep multiplexing levels.

Another increasingly popular method in plants is targeted enrichment (Mamanova *et al.* 2010). In this approach, one increases the ratio of cpDNA vs. nuclear DNA (nrDNA) resulting in higher specificity and better coverage, which in turn allows for deep multiplexing of samples. Several enrichment methods have been explored either via direct multiple traditional PCR amplification of the plastid genome followed by NGS (Cronn *et al.* 2008; Whittall *et al.* 2010) or via chloroplast isolation protocols followed by whole genome amplification (e.g. Atherton *et al.* 2010; Shi *et al.* 2012). In the former, one is quickly confronted with thousands of PCRs (Straub *et al.* 2012; even when using long-range PCRs, Uribe-Convers *et al.* 2014), while in the latter, chloroplast isolation protocols are long and tedious to undertake (Mariac *et al.* 2000), although new methods are being published (Shi *et al.* 2012). In addition, chloroplast extraction approaches are not always available, for example if the DNA has already been extracted, which is frequently the case for crops species (DNA banks) or in the case of herbarium specimens (low DNA yields and isolation of chloroplasts impossible). In either case, both these approaches quickly become difficult for large-scale population genetics or phylogeographic studies.

A more promising approach is targeted enrichment hybridization capture (Cronn *et al.* 2012). This method has been extensively used in human genomics research (Ng *et al.* 2009) with very good results. In plant evolutionary biology, however, this approach has been rarely applied (Cronn *et al.* 2012; Stull *et al.* 2013) but has proven very promising at low multiplexing levels (Parks *et al.* 2012).

Here, we detail a protocol for in-solution enrichment hybridization capture of chloroplast genomes applicable to deep multiplexing levels. This protocol consists of generating in-house cheap probes via long-range PCR (LRPCR) which are then used to enrich libraries for numerous individuals prior to sequencing. First, we undertook the development and validation of the protocol at low multiplexing levels using six economically important crop species in Africa: *Oryza glaberrima* Steud. (African rice), *Pennisetum glaucum* R.Br. (Pearl millet), *Dioscorea rotundata* Poir., *Dioscorea praehensilis* Benth., *Dioscorea abyssinica* Hochst. ex Kunth (African yams) and *Digitaria exilis* Stapf. (Fonio). We then show for the first time that complete chloroplast genomes for 95 individuals can be sequenced in a cost-effective way in a single MiSeq run. Finally, as a proof of concept, we analyse infra- and interspecies diversity of chloroplast genomes for four selected species.

Materials and methods

Samples

For this study, we concentrated on economically important plant species in West Africa. First, to develop and validate the protocol, individuals from three different monocot genera and six different species were used (Table S1, Supporting information): *Oryza glaberrima*, *Pennisetum glaucum*, *Dioscorea rotundata*, *Dioscorea praehensilis*, *Dioscorea abyssinica* and *Digitaria exilis*. All the species are diploid with a genome size ranging from 2 Gb for *Pennisetum glaucum* to 400 Mb for *Oryza glaberrima*. Second, we selected 95 individuals of *Pennisetum* *glaucum* to test deep multiplexing levels of full chloroplast genome sequencing in a single MiSeq run (see below). Finally, we undertook diversity analyses on eight individuals of *Pennisetum glaucum* (taken from the above run) and eight individuals of the three *Dioscorea* species (newly sequenced). DNA extractions from green leaf for all individuals followed Mariac *et al.* (2006) and Scarcelli *et al.* (2006).

General probe design

To capture chloroplast sequences directly from the genomic DNA libraries, biotinylated probes were produced (Fig. 1) using a protocol adapted from Cronn *et al.* (2012). Probe production is undertaken once for each genus following the same protocol. First, and for one individual per genus, an initial full length chloroplast was amplified by LRPCR using 11 primer pairs (Table S2, Supporting information). These primer pairs were taken from Scarcelli *et al.* (2011). However, for difficult fragments, new primer sequences were designed (Table S2, Supporting information). LRPCR was carried out using the LongAmp Taq PCR kit (E5200S New England Biolabs) following the manufacturer's instruction in a final volume of 50 μL and using 100 ng of DNA. We checked whether resulting amplified sizes were those expected from a previously sequenced closely related species (*Dioscorea elephantipes* GenBank number: NC_009601 for the *Dioscorea* species and *O. sativa* GenBank number: NC_001320 for the three other species). Probe production is detailed in Appendix S1 (Supporting information, protocol 1). Briefly, LRPCR fragments were equimolary bulked and sheared to mean target size of 400 bp. Fragments were then end-repaired, and a ligation-mediated PCR was performed with a 5′ TEG-biotinylated primer to produce the probe (Appendix S1, Supporting information).

Library preparations, in-solution hybridization for chloroplast enrichment, multiplexing and sequencing

Libraries were constructed following the Rohland & Reich (2012) protocol using 6-bp barcodes to allow for multiplexing at different degrees (Fig. 1). No Illumina indexes were used. Extra steps were added to the Rohland & Reich (2012) protocol to allow for in-solution hybridization as follows (Fig. 1; Appendix S1, Supporting information): total DNA for each individual was sheared using a Covaris E220 sonicator (Covaris, Woburn, Massachusetts, USA) to a mean target size of 400 bp. DNA was then repaired, ligated and nicked fill-in before a six cycles prehybridation PCR was performed (Appendix S1, Supporting information; protocol 2). After clean-up and quantification, library preparation was mixed with biotin-labelled probes and hybridized to the targeted regions. The hybridized biotin-labelled probes were then immobilized using streptavidin-coated magnetic beads. A magnetic field was applied, and supernatant containing unbounded DNA was discarded. Enriched plastid DNA fragments were then eluted from the beads and amplified in real-time PCR to complete adapters and generate libraries ready for cluster generation and sequencing.

As a control, libraries were also produced using only the probes and genomic DNA (i.e. without the hybridization step). Thus, for each genus, three different libraries were sequenced: enriched chloroplast DNA, probes and nonenriched genomic DNA (Table S1, supporting information). This setup allowed us to test whether the probes effectively covered the whole plastid genome, to control for the homogeneity of its coverage compared with the reference, and estimate the degree of enrichment compared to a genomic (nonenriched) library.

To limit costs and workload, we undertook bulking of eight barcoded individuals in equimolar concentrations just before the enrichment step. Finally, we used this protocol to undertake multiplexing of 95 individuals on a single MiSeq run. All sequencing results presented here were performed on a Illumina MiSeq v3 platform at CIRAD facilities (Montpellier, France) with about 12 pmol of the capture-amplified DNA libraries deposited on the flowcell.

Bioinformatic analyses

Demultiplexing and data cleaning. Demultiplexing was undertaken using a script that sorts pair-end reads in function of a given list of barcodes. The script is freely available at https://github.com/Maillol/demultadapt. Here, we used a 0-mistmatch threshold for demultiplexing. For each library, adapters were removed using the cutadapt 1.2.1 software with the following parameters: cut-off = 20, length minimum = 7, length overlap = 35. Reads were filtered based on their length and their quality mean values ($Q > 30$).

De novo assembly of chloroplast sequences. For the three nonmodel species (*Digitaria, Dioscorea rotundata* and *Pennisetum*), *de novo* chloroplast sequences were assembled because no reference full chloroplasts were available. We used either total genomic DNA libraries (*Digitaria exilis* and *Dioscorea rotundata*) or a LRPCR library (*Pennisetum glaucum*) for this step. The software MITOBIM (v. 1.5, Hahn *et al.* 2013) was used to reconstruct *de novo* plastid genomes after the algorithm was adapted to deal with chloroplast genomes. In the first step, MITOBIM maps reads using MIRA V3.4.1.1 (Chevreux *et al.* 1999) to highly conserved regions of a closely related reference genome

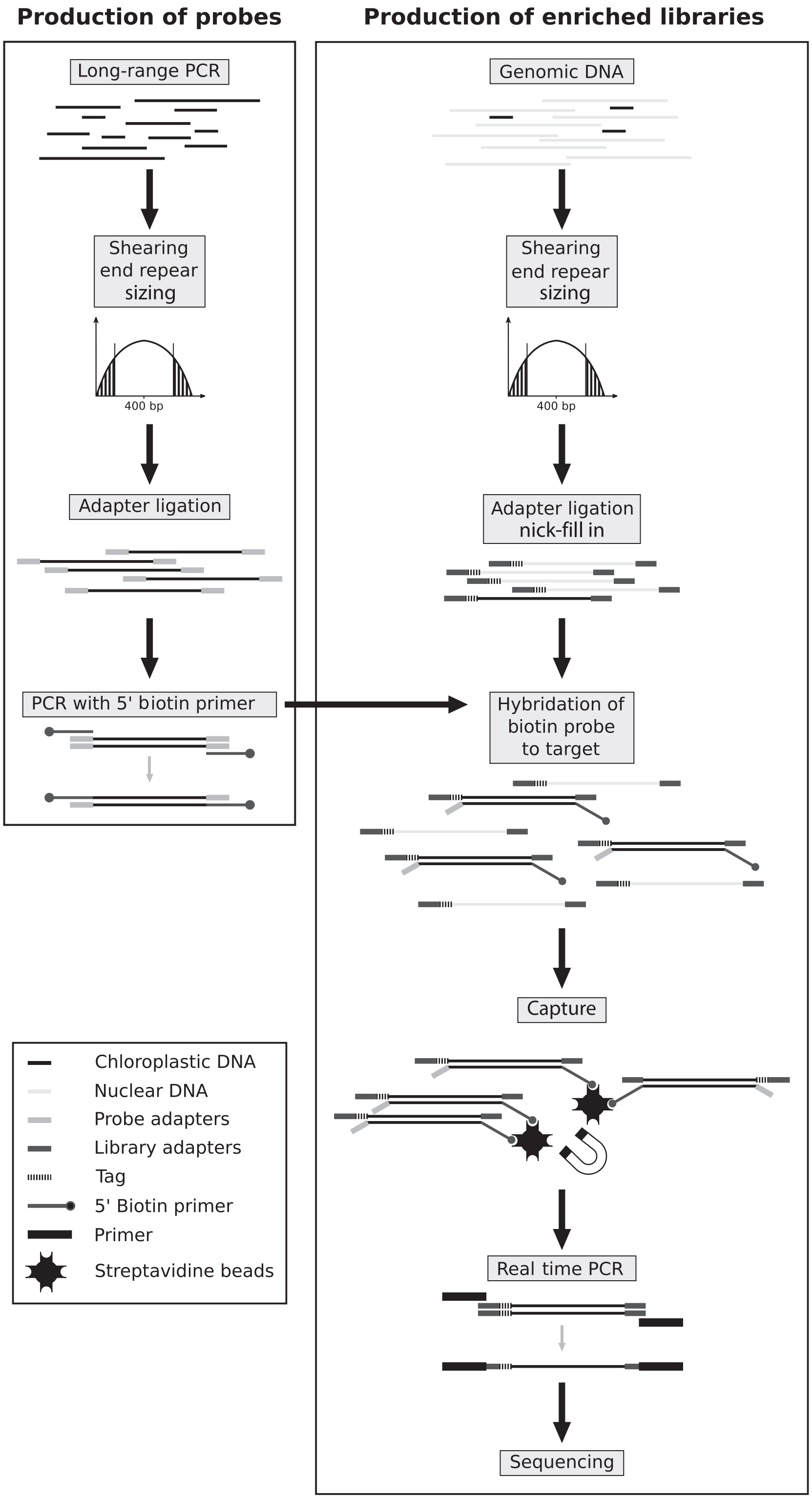

Fig. 1 Protocol workflow for probe construction and for library preparation and hybridization.

(hence, no reference genome of the focal species is required). For this step, we used as initial references: *Zea mays* (NC_001666) for *Pennisetum glaucum*; *Dioscorea elephantipes* (NC_009601) for *Dioscorea rotundata* and the *Pennisetum glaucum* chloroplastic genome reconstructed in this study for *Digitaria exilis*. This step produces a gapped reference sequence containing several contigs. In the second step, in silico baiting is performed with a 31 bp overlap with the reference contigs using all reads. In the third step, the reads identified in step two are mapped to the gapped reference resulting in an extension of the contigs and a reduction in the gaps between contigs. This process was itinerated as many times as needed until a complete *de novo* genome was achieved for all species.

To ensure the quality and validity of the assembly method, we performed a preliminary MITObim run using 500 000 pair-end reads from RAM63 of 75 bases size (Illumina), an *O. sativa* ssp *indica* and the *Brachypodium dystachion* chloroplast (NC_011032) as bait. The standard conditions of MITObim were used for running job. After we removed stretches of N's larger than 10, we obtained a final reconstructed reference 95.5% identical to *O. sativa* ssp. *indica* control (NC_008155) and with a final 4.5% of gaps (Figs S2A et S2B, Supporting information).

Annotation of de novo assembly chloroplast genomes. A first automated annotation was conducted by aligning the chloroplast gene sequences of the species, listed in Table S3 (Supporting information), against the chloroplast of *Digitaria exilis*, *Dioscorea rotundata*, *Oryza glaberrima* and *Pennisetum glaucum*. The alignment was performed using blast (v2.2.22). Identification of protein-encoding sequences was performed using blastp (e-value threshold: 10^{-10}) against chloroplastic open-reading frames (ORFs) extracted from plastome sequences using the perl script get_orf.pl designed by Paul Stothard (University of Alberta). Identification of ribosomal RNA (rRNA) and transfer RNA (tRNA) sequences was performed using blastn with e-value threshold 10^{-10} and 10^{-5}, respectively. The automated annotation was verified and corrected by performing manual alignment using Geneious Pro (version 4.8.5; created by Biomatters, available from http://www.geneious.com/). The resulting annotated sequences have been deposited in GenBank under accession nos KJ490011, KJ490012, KJ513090, KJ513091.

Mapping. De novo genomes for each species generated with MITOBIM were used as references for subsequent mapping. Reads were mapped to their reference using BWA 0.6.2 (Li & Durbin 2009). We then used the TABLET software v. 1.13 (Milne *et al.* 2013) to retrieve coverage information. For each species, we calculated for enriched libraries, probe libraries and nonenriched genomic libraries, the percentage of reads mapped to the reference, the percentage of the reference sequence covered and average and maximum depths per base. Enrichment was calculated as the ratio between reads mapped to the chloroplast of the enriched library vs. the reads mapped to the chloroplast of nonenriched library.

SNP detection. Single-nucleotide polymorphisms (SNP) as well as microsatellite variations were called on 9 individuals in *Dioscorea* and *Pennisetum* each based on their full chloroplast sequences. The mpileup command from the SAMTOOLS program 0.1.18 without indel calling was initially used (Li *et al.* 2009). A VCF file was then produces with bcftools 0.1.17 and filtered with the GenomeAnalysisTK variant filtration program 2.4-7 (McKenna *et al.* 2010). Heterozygote genotypes were assigned as missing data. Sequence data diversity for *Dioscorea* and *Pennisetum* were calculated using DNAsp v.5 (Librado & Rozas 2009). We calculated the number of variants (single-nucleotide polymorphism or microsatellite variants), the number of haplotypes, the estimation of θ the product of two times the effective size and the mutation rate ($4N\mu$) based on the number of segregating sites (θ_W) and, finally, the average nucleotide diversity (π).

Results

Probe validation

LRPCRs for probe construction ranged from 7 to 22 kb (Table S2, Supporting Information). Most of the primers used were designed in a previous study (Scarcelli *et al.* 2011). However, new primers for some regions were redesigned for successful amplification (Table S2, Supporting Information).

The first probe was constructed using a single individual of African rice because it is a well-studied cereal with an already available independently sequenced reference plastome for *Oryza sativa*. Probes developed for African rice (Fig. 2A) were sequenced to confirm their identity, resulting in a total of 95.9% of the reads mapping to *Oryza sativa* (Table S2, Supporting Information). The sequences covered 99.7% of the nucleotides of the chloroplast genome (Fig. 2A, Table S2, Supporting information). This confirmed that the probes were mainly composed of African rice chloroplast DNA. Maximum difference in coverage observed along the chloroplast is 2.1-fold for the LRPCR fragments. Coverage within a LRPCR fragment was relatively homogeneous. Overlapping regions between LRPCR fragments lead to local coverage increase, which was expected (Fig. 2). For the inverted region (IR) a and IRb, a local decrease in coverage was observed in a specific 5000-bp sequence identical between the two IRs which contain relatively high levels GC content (55%). For reference, the mean GC content in the *Oryza* chloroplast DNA is 39%.

Enrichments protocol validation

Nonenriched sequencing of the African rice library resulted in 8.6% of reads mapping to the chloroplast with an average coverage depth of 110 (Fig. S2, Table S1, Supporting Information). Applying our enrichment protocol during the construction of the library led to 96% of reads

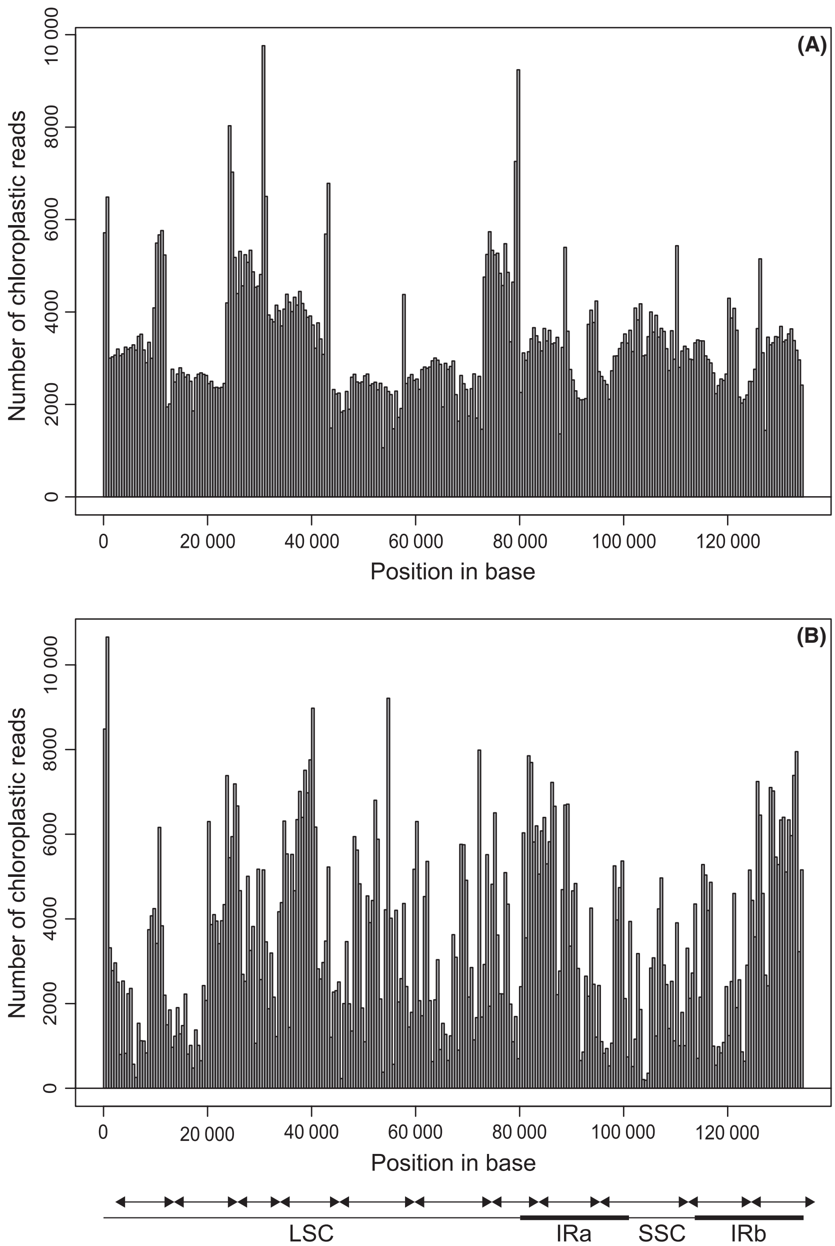

Fig. 2 Sequence coverage of the whole African rice chloroplast. The graphic represents the number of reads mapped to the chloroplast reference genome (*Oryza sativa* NC_001320) for the long-range PCR (LRPCR) probe library (A) and for an individual after enrichment (B). The LRPCR fragment sizes are illustrated with horizontal lines with arrows. The position of the large single copy (LSC), inverted repeat (thick lines, IRa and IRb) and small single copy regions (SSC) are also reported.

mapping to the rice chloroplast. The percentage of bases covered was 99.8% with an average coverage depth of 146 (Fig. 1B, Table S2, Supporting Information).

Application to other species

We then applied the same protocol to the other three species used in our study. Because these species do not have a very close relative with a complete chloroplast to serve as a reference, we first reassembled the chloroplast genomes *de novo* using MITOBim (Hahn *et al.* 2013). The assembled chloroplast of *P. glaucum*, *Dioscorea rotundata* and *Digitaria exilis* was 140 718, 155 406 and 140 908 bp long, respectively. Very few ambiguities were observed in the assembly with only 195 N bases for *Pennisetum* (among them, 191 were manually checked and corrected), 24 N bases for *Digitaria* and none for *Dioscorea*.

For the nonenriched libraries, reads mapped to the *de novo* reference ranged from 6.3% in *Digitaria* and 6.2% in *Dioscorea* to just 0.53% in *Pennisetum* (Table S2, Fig. S1, Supporting information). Our enrichment protocol led to a large increase in mapped chloroplast DNA for the three species (Fig. S1, Table S2, Supporting Information) with 95.8%, 95.9% and 63.0% for each species, respectively. The enrichment fold is the highest for *Pennisetum*, with a

116-fold increase and lower for *Oryza* (11), *Dioscorea* (16) and *Digitaria* (15). The fold increase is related to the initial frequency of chloroplasts sequenced as well as the fact that for the three last species, the enrichment is almost complete (95%).

Bulking of individuals

To reduce cost and workload, we simplified our protocol by bulking eight barcoded individuals together prior to enrichment step (Fig. 1). For *Dioscorea*, the enrichment was similar to the single individual (no bulking) protocol with an average percentage of reads mapped of 95.5% and chloroplast covered averaged 97.7%. For *Pennisetum*, the bulk enrichment leads to an average of 45.3% of reads mapping the chloroplast which was lower than the value found for a single individual of 60%. However, the percentage of the chloroplast covered is identical for the eight bulked samples than for the single individual sample, corresponding to 99.5%. These results indicate that bulking for eight individuals can be performed easily with no loss in results, decreasing the complexity of the protocol step-up.

Deep multiplexing of individuals

To validate our protocol for deep multiplexing, we sequenced 95 enriched *Pennisetum* samples in a single MiSeq run. One individual (PE03815, TAG7) did not work and was excluded from subsequent analyses (Fig. 3D, Table S1, Supporting Information). All the other samples had a number of reads varying from 10925 to 566545, with an average value of 250000 (Fig. 3D, Table S1, Supporting Information). The number of reads mapping to their reference varied from 1930 to 379 422 with an average value of 148 000. The percentage of reads mapping to the reference varied from 16% to 83%, with a mean value of 56% (Fig. 3A, Table S1, Supporting Information). The average coverage was 99% across the 94 individuals (Fig. 3B). So if we considered only the individuals with at least 95% coverage of the chloroplast, 90 of the 95 samples had sufficient data. The success rate for retrieving a whole chloroplast genome covering at least 95% is 94.7%. Finally, our results show that with under 20 000 reads, we can reconstruct 99% of the genome (Fig. 3C).

Diversity analyses

We did a preliminary diversity study to assess whole chloroplast variability for population genetics (within species) and phylogeographies studies (between closely related species or species complexes). We detected 70 SNPs and 23 variable microsatellites in *Dioscorea*. Considering only SNPs, nucleotide diversity was 1.96×10^{-4} and theta of Watterson was estimated at 1.88×10^{-4} for *Dioscorea*. For *Pennisetum*, 22 SNP sites are detected as well as 11 polymorphic microsatellite loci. The nucleotide diversity is estimated at 6.14×10^{-5}, and theta of Watterson was estimated at 5.8×10^{-5}.

Discussion

We present a protocol for complete chloroplast genome sequencing using an in-solution hybridization-based enrichment approach applicable to deep multiplexing levels, with a main application to population genetics and phylogeography of plants. We show that this protocol can be used to capture whole chloroplast genomes across several monocot species in a cost- and time-effective way. Previously published protocols for plastome enrichment via hybridization depended on short probes (18–120 bp, oligoprobes) spanning either a fraction (numerous primer pairs; Cronn *et al.* 2012) or the entire genome (120-bp long overlapping RNA probes; Stull *et al.* 2013). In addition, probes have been constructed by PCR amplifying the whole plastome and concatenating them into a single probe several kbp long (Cronn *et al.* 2012). In all cases, enrichment provided good results in terms of chloroplast coverage and depth at shallow multiplexing levels. In the present protocol, probes are generated from 11 long-range PCR fragments implying fewer total reactions, and thus, less complex setups and less time spent fine tuning compared with previous protocols. It is also cheaper than synthesizing numerous short RNA probes (Stull *et al.* 2013). Additionally, our homemade probe construction can be PCR amplified numerous times eliminating the need to be resynthesized or reordered and reducing costs over time compared with probes that need to be commercially synthesized (e.g. short DNA or RNA probes).

Although the above-mentioned protocols indicate that deep multiplexing is feasible, multiplexing has never exceeded 24 samples using hybridization protocols (Stull *et al.* 2013). Bock *et al.* (2014) sequenced 34 plastomes without prior enrichment with good coverage (average 95x, min: 30x; max 355x) suggesting that genome skimming could be a useful and easy way to generate large numbers of plastomes. However, these results will largely depend on the ratio between cpDNA vs. nrDNA per cell. This ratio is known to vary greatly across angiosperm species as well as between and within individuals (e.g. Rauwolf *et al.* 2010). For example, in monocotyledons, the size of genomes varies 1000-fold (see http://data.kew.org/cvalues). Our results on pearl millet are a good case in point of the limits of genome skimming as genomic nonenriched DNA generated only a fraction of reads mapping to the reference chloroplast genome

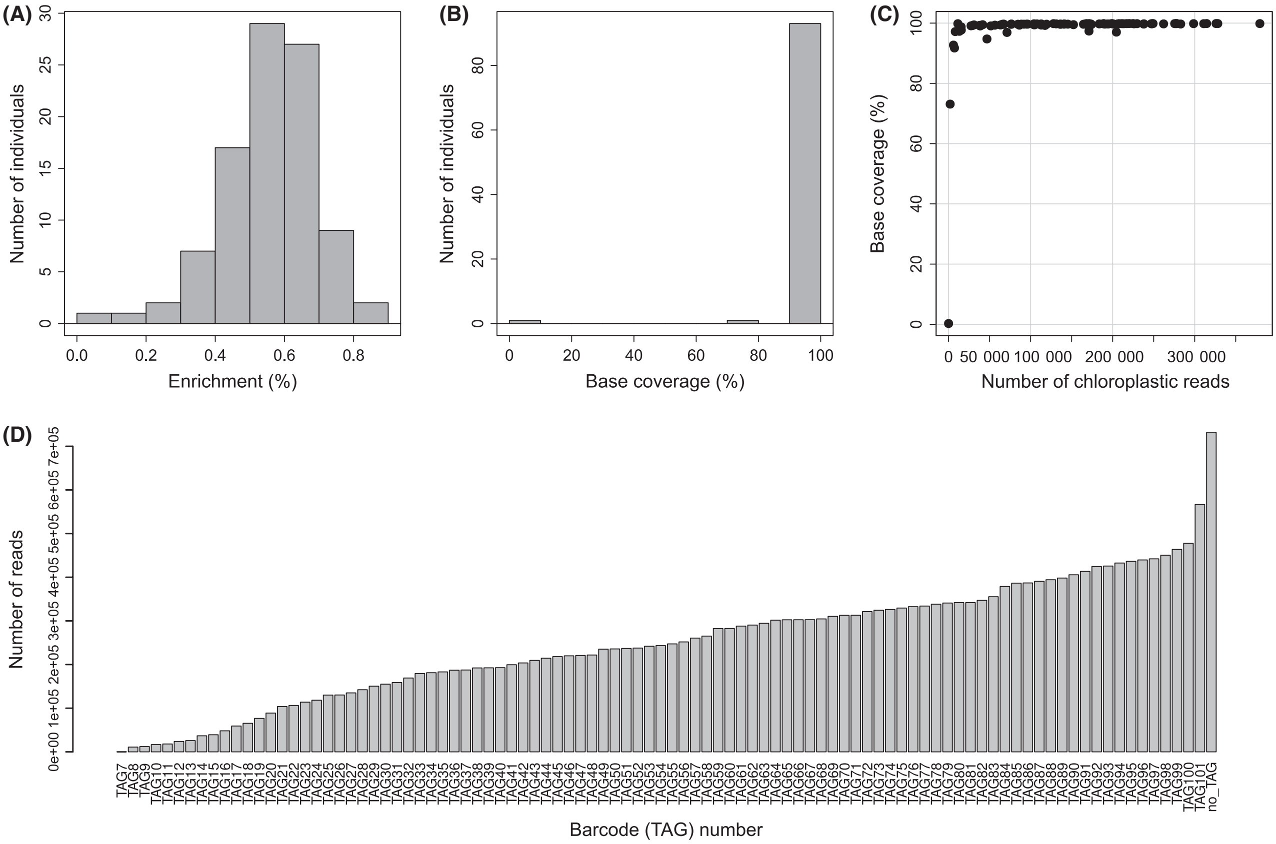

Fig. 3 Statistics of the deep-multiplexed MiSeq run for 95 *Pennisetum glaucum* individuals. (A): Histogram representing the percentage of enrichment for the 95 individuals of pearl millet. (B): Histogram representing the percentage of coverage of the whole chloroplast for the 95 individuals of pearl millet. (C): The number of chloroplast reads in function of the coverage of the reference chloroplast genome (base coverage%). (D): Total number of reads per TAG. The last bar (no_TAG) indicates the number of reads without a TAG.

Table 1 Plastome diversity in *Pennisetum* and *Dioscorea* individuals. Diversity was calculated using nine individuals for each species. Polymorphism values were calculated using the whole data set [including short sequence repeats (SSRs) and indels] or only using SNPs (indicated with *)

Species	Number of sequences used	Number of polymorphism sites	Number of haplotypes	Haplotype diversity	Nucleotide diversity	Theta-W (per site)
Pennisetum	9	33	6	0.889	8.89E-05	8.67E-05
		22*	6*	0.889*	6.14E-05*	5.84E-05*
Dioscorea	9	93	3	0.556	2.55E-04	2.44E-04
		70*	3*	0.556*	1.96E-04*	1.88E-04*

(*Pennisteum glaucum*: 0.5%, Table S1, Fig. S1, Supporting Information). We predict than in the case of pearl millet, deep multiplexing of individuals using the genome-skimming approach would be unsuccessful. We do note, however, that genome skimming does provide a good approach when one is also interested in nuclear sequences and not just plastome data (Bock *et al.* 2014).

Deep multiplexing has already been achieved for mitochondrial genomes (Hancock-Hanser *et al.* 2013) but has never been undertaken for chloroplasts. Here, we show for the first time that sequencing 95 plastomes in a single MiSeq run is achievable using our protocol. Our results confirm that hybridization via probes is a cost-effective way to sequence entire organelle genomes at deep multiplexing levels. Not considering the development of the probe, in this study, we produced a complete plastome for a total cost of just 22$ *per individual* (*including library preparation at* 10$ per individual and one

MiSeq run at 1150$). Higher multiplexing levels are certainly achievable. Our results indicate that with 10 000 reads mapped to the chloroplast, a complete plastid genome is recovered at a depth of 100x (Fig. 3C, Table S1, Supporting information). Thus, in the present run, given 26 million useful reads (Table S1, Supporting information), we could have multiplexed up to 1500 individuals for complete plastome sequencing with a depth of 100x. This would be a conservative estimation, as they are based on an average enrichment of 56% for *Pennisetum* (Table S1, Supporting information). However, for the other species enrichment exceeded 93% (Table S1, Supporting information), thus significantly more individuals could be multiplexed. The main problem now is the number of available barcodes. A total of 148 hexamere barcodes are currently available (Craig *et al.* 2008; Rohland & Reich 2012), but multiplexing can be drastically increased by combing the barcodes with single or dual Illumina indexes (24 available to date) or by extending the barcodes to 8 bp.

One limitation of our protocol is that primer pairs used to generate the probes via LRPCR have to cover the whole genome of the focal taxon, implying knowledge of the chloroplast genome beforehand. This is a problem whenever one has to construct probes in general and can be problematic for nonmodel species. However, the primer list published by Scarcelli *et al.* (2011) provided a good source for LRPCR primer selection for four monocotyledon genera with <20% of these primers had to be redesigned based on closely related reference genomes (Table S2, Supporting information). Recently, Uribe-Convers *et al.* (2014) published a list of 16 LRPCR primer pairs potentially usable across angiosperms. In other groups, published plastid genomes of even distally related species could provide sufficient data to design appropriate primers, although more tests would have to be undertaken. In any case, sequencing a chloroplast genome beforehand from a single individual is relatively easy and would require little investment, for example using genome-skimming approaches (Cronn *et al.* 2012; Straub *et al.* 2012).

In terms of portability of the probes to other species, our results indicate that there is no difference in enrichment within the *Dioscorea* species complex always being higher than 93% (Table S1, Supporting information). This indicates that our approach is useful to enrich chloroplasts in phylogeography studies of closely related species (Terauchi *et al.* 1992; Scarcelli *et al.* 2011). However, they provide little information on the portability towards more distantly related species. Other studies have shown that probes designed in one species could be useful to enrich organelles of even distantly related species in both animals (mitochondria; Hancock-Hanser *et al.* 2013) and plants (chloroplasts; Cronn *et al.* 2012; Stull *et al.* 2013), although diminishing in efficiency with genetic distance. Cronn *et al.* (2012) showed that probes designed in the species *Pinus thunbergii* (gymnosperm) could efficiently enrich chloroplasts in cotton (*Gossypium raimondii*, Malvaceae). These two groups diverged around ca. 330 million years ago (Magallón & Sanderson 2005) indicating that species-specific probes could be used on broader phylogenetics scales.

The protocol developed here allows the analysis of plastid diversity within and between closely related species. For *Pennisetum* and based on nine individuals, we recovered 33 and 22 SNPs including or excluding microsatellite regions, respectively, across the entire plastome (Table 1). For nine individuals of the *Dioscorea* species complex, we retrieved 93 and 70 SNPs (Table 1). These levels are in the range (*Dioscorea*) or lower than (*Pennisetum*) the one found for other species such as Cacao (78 SNPs, nine individuals; Kane *et al.* 2012). However, this level of diversity is similar or higher to the D-loop mitochondrial, one of the most used and useful markers in animal phylogeography studies (e.g. cattle: Beja-Pereira *et al.* 2006; horses: Cieslak *et al.* 2010). These preliminary data strongly support the idea that plastomes have sufficient variability at the genomic level to be useful in population genetic and phylogeographic studies (Cronn *et al.* 2012; Kane *et al.* 2012). Our method could be largely applied for phylogeography and population genetic studies within species or between closely related species.

Acknowledgements

This project is supported by Agropolis Foundation through the « Investissements d'avenir » programme (ANR-10-LABX-0001-01) under the reference ID 1202-040 to the senior author. The work of Vincent Maillol at INRA, UMR AGAP was supported by the project GRAPERESEQ (European Plant KBBE 2008). We wish to thank Valerie Laucou and Catherine Breton for their help with the bioinformatics, Morgane Ardisson and Marie Couderc for QPCR setup, Hélène Vignes for her assistance with the Miseq runs, Eric Desmarais for discussions during the project. We also thank Christoph Hahn for adapting the MITOBim program to deal with chloroplast genomes.

References

Atherton R, McComish B, Shepherd L *et al.* (2010) Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform. *Plant Methods*, **6**, 22.

Avise JC (2000) *Phylogeography: The History and Formation Of Species* Harvard University Press, Cambridge.

Barrett CF, Davis JI, Leebens-Mack J, Conran JG, Stevenson DW (2013) Plastid genomes and deep relationships among the commelinid monocot angiosperms. *Cladistics*, **29**, 65–87.

Beja-Pereira A, Caramelli D, Lalueza-Fox C *et al.* (2006) The origin of European cattle: evidence from modern and ancient DNA. *Proceedings of the National Academy of Sciences*, **103**, 8113–8118.

Besnard G, Hernandez P, Khadari B, Dorado G, Savolainen V (2011) Genomic profiling of plastid DNA variation in the Mediterranean olive tree. *BMC Plant Biology*, **11**, 80.

Bock DG, Kane NC, Ebert DP, Rieseberg LH (2014) Genome skimming reveals the origin of the Jerusalem Artichoke tuber crop species: neither from Jerusalem nor an artichoke. *New Phytologist*, **201**, 1021–1030.

Chevreux B, Wetter T, Suhai S(1999) Genome Sequence Assembly Using Trace Signals and Additional Sequence Information Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB) vol: 99, 45-56.

Cieslak M, Pruvost M, Benecke N *et al.* (2010) Origin and history of mitochondrial DNA lineages in domestic horses. *PLoS ONE*, **5**, e15311.

Craig DW, Pearson JV, Szelinger S *et al.* (2008) Identification of genetic variants using bar-coded multiplexed sequencing. *Nature methods*, **5**, 887–893.

Cronn R, Liston A, Parks M *et al.* (2008) Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Research*, **36**, e122.

Cronn R, Knaus BJ, Liston A *et al.* (2012) Targeted enrichment strategies for next-generation plant biology. *American Journal of Botany*, **99**, 291–311.

Hahn C, Bachmann L, Chevreux B (2013) Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Research*, **41**, e129.

Hancock-Hanser BL, Frey A, Leslie MS *et al.* (2013) Targeted multiplex next-generation sequencing: advances in techniques of mitochondrial and nuclear DNA sequencing for population genomics. *Molecular Ecology Resources*, **13**, 254–268.

Kane N, Sveinsson S, Dempewolf H *et al.* (2012) Ultra-barcoding in cacao (*Theobroma* spp.; *Malvaceae*) using whole chloroplast genomes and nuclear ribosomal DNA. *American Journal of Botany*, **99**, 320–329.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

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

Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, **25**, 1451–1452.

Magallón SA, Sanderson MJ (2005) Angiosperm divergence times: the effect of genes, codon positions, and time constraints. *Evolution*, **59**, 1653–1670.

Mamanova L, Coffey AJ, Scott CE *et al.* (2010) Target-enrichment strategies for next-generation sequencing. *Nature methods*, **7**, 111–118.

Mariac C, Trouslot P, Poteaux C, Bezancon G, renno J-F, (2000) Chloroplast DNA extraction from herbaceous and woody plants for direct restriction fragment length polymorphism analysis. *BioTechniques*, **28**, 110–113.

Mariac C, Luong V, Kapran I *et al.* (2006) Diversity of wild and cultivated pearl millet accessions (*Pennisetum glaucum* [L.] R. Br.) in Niger assessed by microsatellite markers. *Theoretical and Applied Genetics*, **114**, 49–58.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2013) Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, **66**, 526–538.

McKenna A, Hanna M, Banks E *et al.* (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.

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

Moore MJ, Bell CD, Soltis PS, Soltis DE (2007) Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 19363–19368.

Ng SB, Turner EH, Robertson PD *et al.* (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, **461**, 272–276.

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

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

Petit RJ, Vendramin GG (2007) Plant phylogeography based on organelle genes: an introduction. In: *Phylogeography of Southern European Refugia* (eds Weiss S & Ferrand N), pp. 23–97. Springer, Dordrecht, the Netherlands.

Rauwolf U, Golczyk H, Greiner S, Herrmann R (2010) Variable amounts of DNA related to the size of chloroplasts III. Biochemical determinations of DNA amounts per organelle. *Molecular Genetics and Genomics*, **283**, 35–47.

Rohland N, Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research*, **22**, 939–946.

Scarcelli N, Tostain S, Mariac C *et al.* (2006) Genetic nature of yams (*Dioscorea* sp.) domesticated by farmers in Benin (West Africa). *Genetic Resources and Crop Evolution*, **53**, 121–130.

Scarcelli N, Barnaud A, Eiserhardt W *et al.* (2011) A set of 100 chloroplast DNA primer pairs to study population genetics and phylogeny in monocotyledons. *PLoS ONE*, **6**, e19954.

Shi C, Hu N, Huang H *et al.* (2012) An improved chloroplast DNA extraction procedure for whole plastid genome sequencing. *PLoS ONE*, **7**, e31468.

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

Stull GW, Moore MJ, Mandala VS *et al.* (2013) A targeted enrichment strategy for massively parallel sequencing of angiosperm plastid genomes. *Applications in Plant Sciences*, **1**, 1200497.

Terauchi R, Chikaleke V, Thottappilly G, Hahn S (1992) Origin and phylogeny of Guinea yams as revealed by RFLP analysis of chloroplast DNA and nuclear ribosomal DNA. *Theoretical and Applied Genetics*, **83**, 743–751.

Uribe-Convers S, Duke JR, Moore MJ, Tank DC (2014) A long PCR–based approach for DNA enrichment prior to next-generation sequencing for systematic studies. *Applications in Plant Sciences*, **2**, 1300063.

Varshney RK, Nayak SN, May GD, Jackson SA (2009) Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends in Biotechnology*, **27**, 522–530.

Whittall JB, Syring J, Parks M *et al.* (2010) Finding a (pine) needle in a haystack: chloroplast genome sequence divergence in rare and widespread pines. *Molecular Ecology*, **19**, 100–114.

Yang J-B, Yang S-X, Li H-T, Yang J, Li D-Z (2013) Comparative chloroplast genomes of *Camellia* Species. *PLoS ONE*, **8**, e73053.

Zhang Y-J, Ma P-F, Li D-Z (2011) High-throughput sequencing of six bamboo chloroplast genomes: phylogenetic implications for temperate woody bamboos (*Poaceae*: *Bambusoideae*). *PLoS ONE*, **6**, e20596.

C.M., Y.V., S.S. and T.L.P.C. designed the study; C.M. and J.P. produced the data; C.M., Y.V., F.S., A.K. and G.M. analysed the data; N.S., A.B., C.B., A.F. contributed materials; V.M. wrote the demultiplexing script; C.M., Y.V. and T.L.P.C. wrote the study.

Data Accessibility

The R script is available at https://github.com/Maillol/demultadapt

Sequences data are available at:

DNA sequence: NCBI SRA accessions listed in Table S1 (project SRP033144). GenBank accessions: KJ490011, KJ490012, KJ513090, KJ513091 (annotated chloroplasts).

DRYAD entry doi:10.5061/dryad.t6b05. Sequence alignment map in BAM file format and references listed in Table S1. RAM63, 500 K forward and reverse reads used for assembly validation.

Supporting Information

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

Table S1 Description of samples.

Table S2 Description of long-range PCR conditions.

Table S3 List of species used for the database gene construction in order to perform the automatic annotations.

Figure S1 Percentage of reads mapping to the chloroplast reference for the genomic nonenriched library (grey) and the genomic enriched library (dark) for *Oryza*, *Pennisetum*, *Dioscorea* and *Digitaria*.

Figure S2 Validation of the assembly method.

Appendix S1 Protocol 1: Construction of biotinylated probes. Protocol 2: Construction of enriched library based on Rohland & Reich (2012).