BETHUNE et al. Long fragment capture and sequencing

Long-fragment targeted capture for long read sequencing of plastomes

Bethune Kevin[1*]; Mariac Cedric[1*]; Couderc Marie[1]; Scarcelli Nora[1]; Santoni Sylvain[2]; Ardisson Morgane[2]; Martin Jean-Francois[3]; Montufar Rommel[4]; Valentin Klein[1]; Francois Sabot[1]; Vigouroux Yves[1]; Couvreur L.P. Thomas[1]

[1]IRD, DIADE, Univ Montpellier, Montpellier, France

[2]UMR AGAP, Equipe Diversitée et Adaptation de la Vigne et des Espèces Méditerranéennes, INRA, 2 Place Viala, 34060 Montpellier, France

[3]CBGP, Montpellier SupAgro, INRA, CIRAD, IRD, Univ Montpellier, Montpellier, France

[4] Facultad de Ciencias Exactas y Naturales, Pontificia Universidad Católica del Ecuador, Quito, Ecuador

* both authors contributed equally to the work

Manuscript received ___; revision accepted ___.

Number of words: 3678

ABSTRACT

Third generation sequencing methods generate significantly longer reads than those produced using alternative sequencing methods. This provides increased possibilities to better study biodiversity, phylogeography and population genetics. We developed a protocol for in-solution enrichment hybridization capture of long DNA fragments applicable to complete chloroplast genomes. The protocol uses cost effective in-house probes developed via long-range PCR and was used in six non-model monocot species (Poaceae: African rice, pearl millet, fonio; and three palm species). DNA was extracted from fresh and silicagel dried leaves. Our protocol successfully captured long read chloroplast fragments (up to 4 264 bp median) with an enrichment rate ranging from 15% to 98%. DNA extracted from silicagel dried leaves led to low quality plastome assemblies when compared to freshly extracted DNA. Our protocol could also be generalized to capture long sequences from specific nuclear fragments.

Keywords: MinION, DNA probes, long-range PCR, whole chloroplast sequencing, De Novo assembly

INTRODUCTION

High throughput sequencing is revolutionizing research in plant evolutionary biology. The development of second generation sequencing (SGS) led to a massive amount of sequence data to be generated in a cost effective way (Straub et al. 2012). Besides the many advantages of SGS one shortcoming is that they generate short reads (between 100-400 base pairs (bp)). This is problematic for *de novo* assemblies of plant genomes that prove difficult in resolving repetitive sequences due to transposable elements, polyploidy and large genome sizes.

In contrast to SGS, third generation sequencing (TGS) directly targets single DNA molecules without prior PCR, enabling “real time sequencing” (Bleidorn 2016). The main improvement of TGS is the significant increase in read length from tens to tens of thousands of bases per single read (termed ‘long reads’). This provides important advantages to improve *de novo* assemblies (Jiao and Schneeberger 2017), gap filling (Eckert et al. 2016) or phasing (Laver et al. 2016). Technologies such as Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) are able to generate mean read lengths ranging from 5kbp to 200kbp in standard analyses (and peak up to 2 mbp) depending on the quality of the DNA (Lee et al. 2016). One drawback is that most TGS technologies have high error rates when compared to SGS (~ 10% for ONT MinION versus 0.1% for Illumina, Goodwin et al. 2016). However, new base calling algorithms, associated with *a posteriori* corrections, allow for a significant decrease sequence errors. With sufficient coverage and proper algorithms, TGS can lead to assemblies with consensus nucleotide accuracy of 99.90% (Lee et al. 2016).

The application, however, of TGS using MinION to complex genomes such as plants is problematic mainly because of the generally low output of data currently available (10-20 Gb versus 1,500 Gb for a HiSeq4000, Illumina). Thus, efficiently sequencing specific regions will depend on genome reduction approaches, such as targeted sequencing (Cronn et al. 2012; Jones and Good 2016). Genome reduction *via* sequence capture refers to DNA fragments (nuclear, ribosomal or plastid) that are directly captured from a total genomic library using probes binding to the complementary DNA sequences. This approach has the advantage of being cost effective, optimizes read depth on the targeted region and allows to analyze more samples per run. However, sequence capture is only routinely undertaken on short DNA fragments (Mamanova et al. 2010; Cronn et al. 2012), limiting its usefulness for long read based TGS.

Sequencing of complete chloroplasts or plastomes have been shown to be a marker of choice for the study plant evolution (Mariac et al. 2014; Twyford & Ness 2017). *De novo* assembly of plastomes based on short reads can be problematic (Mariac et al. 2014) leading to low quality reference plastomes. This is especially true for non-model taxa where no high quality reference genomes are available. Given the low output of data from MinION, this technology cannot be easily used to sequence plastomes directly from genomic DNA (e.g. genome skimming). The main challenge in order to efficiently apply TGS to the study of plant evolution will be based on our ability to capture long DNA fragments. To date long read targeted capture has mainly been undertaken on simple organism such as bacteria or virus (e.g.

Eckert et al. 2016) and rarely in complex organisms such as plants. Protocols for DNA enrichment for segments in excess of 20kbp in length have also been developed (Dapprich et al. 2016). In plants, few studies have undertaken long read targeted capture (Giolai et al. 2016, 2017). These protocols prove that capturing long fragments is possible but has yet to be routinely developed for plants.

Here, we present a protocol to capture long reads for plastome sequencing and reassembling using ONT MinION technology. We first developed our protocol for the model plant species *Oryza sativa* (Asian rice). We then applied the protocol to sequence plastomes in several wild species and non-model but economically important crops. Finally, we tested the ability to capture and assemble plastomes from DNA extracted silicagel dried leaves.

MATERIAL AND METHODS

Sampling strategy and DNA extraction

For this study, we focused on seven economically important plant species from Asia, Africa and South America. First, we developed and validated our long read capture protocol using the model plant species *Oryza sativa* (Asian rice). We then applied our protocol to several other plant species from the same genus (*Oryza*), family (*Poaceae*) and finally super-order (*Lilianae* or Monocotyledons): African rice (*Oryza glaberrima* Steud.), a close relative to Asian rice, Pearl Millet (*Cenchrus americanus* (L.) Morrone (*Pennisetum glaucum*)), Fonio (*Digitaria exilis* Stapf.), and three species of palms: *Podococcus acaulis* Hua, *Raphia textilis* Welw. and *Phytelephas aequatoralis* Spruce (Table 1, Table S1). Export of *Podococcus acaulis* and *Phytelephas aequatoralis* silicagel dried leaves were authorized by the Centre national de la recherche scientifique (CENAREST, Gabon) and the Ministerio del Ambiente (Ecuador), respectively.

DNA was extracted from fresh leaves for *O. sativa*, *O. glaberrima*, *C. americanus* and *D. exilis*; while silicagel dried leaves were used for DNA extraction for *Podococcus aucaulis*, *Raphia textilis* and *Phytelephas aequatoralis*. In both cases DNA extraction was performed using a MATAB lysis buffer and chloroform isoamyl alcohol (24:1) purification method following Mariac et al. (2006).

General probe design

Long fragment chloroplast sequences were captured from the total genomic DNA extracts using two different sets of biotinylated probes: one based on *O. sativa* and used on related Poaceae species (*O. glaberrima, C. americanus*, *D. exilis*) and one based on *P. barteri* Mann & H.Wendl. and used for *P. aucaulis*, *R. textilis* and *P. aequatoralis*. Probe production (Figure 1, Supplementary file for detailed steps) was undertaken following the protocol described elsewhere (Cronn et al. 2012; Mariac et al. 2014) and lead to an average probe size of 300 bp: first, an initial full length chloroplast was amplified by long range PCR (LRPCR), using 11 primer pairs taken from Scarcelli et al. (2011) for *O. sativa* (Table S2), and another set of 11 primer pairs taken from Faye et al (2016) for *P. barteri* (Table S2). LR-PCR were carried out using the LongAmp Taq PCR kit (New England BioLabs® Inc., #E5200S) following the manufacturer's instruction in a final volume of 50 µL and using 300 ng of DNA. For each probe set LR-PCR amplicons were equimolary pooled and sheared to reach a mean size fragment of 300 bp, then ligated to adapters so that they can be PCR amplified with biotinyladed primers.

Library preparation, in-solution hybridization, multiplexing and sequencing

Illumina type libraries were constructed following the Rohland & Reich (2012) protocol using 6-bp barcodes and Illumina indexes with some extra steps added to allow for amplification and in-solution hybridization (Figure 1, Table S3). Briefly, each high molecular weight DNAs were sheared using G-tubes (Covaris®) to a mean target size of 10 Kb. DNA fragments below 2,000 bp were removed by a sizing step performed with 0.4X ampure beads. DNA was then end-repaired, ligated with adapters (allowing PCR amplifications) and then nick filled-in before performing a pre-hybridation PCR. Optimal cycle number (ranging from 5 to 12) was defined by real-time amplification (KAPA Biosystems, KK2700). After clean-up and quantification using NanoQuant and QIAxcel, library preparations were mixed with biotin-labelled probes for hybridization of the targeted regions. DNA-probe hybridization complexes were then immobilized with 100 µg of streptavidine coated magnetic beads. This step was performed using the Dynabeads^{TM} M-280 kilobaseBINDER^{TM} Kit (Invitrogen, ThermoFisher Scientific, #60101), which is designed for immobilizing double stranded DNA molecules longer than 2 kbp.

A magnetic field was applied to the resulting solution and the supernatant containing unbounded DNA was discarded. Enriched DNA fragments were then dehybridized from the beads and amplified in a 12 to 15 cycles real-time PCR in order to obtain requested quantity for the Nanopore library preparation. The final libraries were then constructed following the Nanopore library preparation detailed in the 1D Amplicon by ligation (SQK-LSK108) protocol for single samples and also in the 1D Native barcoding genomic DNA (with EXP-NBD103 and SQK-LSK108) protocol. Briefly 1µg of enriched DNA was end-repaired, extended with a dA-tailing, ligated with Nanopore barcodes and then with Nanopore tether-adapter required previous to loading and sequencing on the MinION flowcell. To benefit from multiplexing and limit costs and workload, up to four individuals were equimolarily pooled using Oxford Nanopore barcodes. Prior to each run, flowcells (FLO-MIN106 R9.4 Version) were quality-tested using the MinKNOW software version-1.2.8 to ensure the presence of at least 50% (256) of active channels. Flow cells were loaded with around 275±100 fmol of capture-amplified DNA libraries.

Non-enriched MiSeq data

To estimate enrichment rate, we used single sample non-enriched library datasets originating from various Illumina MiSeq sequencing runs for *O. sativa*, *O. glaberrima*, *C. americanus* and *P. aequatoralis*. For *D. exilis*, *P. aucaulis*, and the *R. textilis*, we merged 10, 2 and 16 samples, respectively, of non-enriched libraries to provide adequate read counts. Forward sequencing read outputs from each MiSeq runs, namely R1 files, were first demultiplexed using demultadapt script (https://github.com/Maillol/demultadapt) to sort reads according to a given barcodes list. Adapters at the beginning of each read from the R2 and demultiplexed R1 files were removed using cutadapt-1.2.1 software (Martin 2011) with the default parameters. Reads were then filtered on their length (size > 35bp) and mean quality values (Q > 30) before being paired using compare_fastq_paired_v5.pl, (https://github.com/SouthGreenPlatform/arcad-hts/blob/master/scripts/arcad_hts_3_synchronized_paired_fastq.pl and /arcad_hts_2_Filter_Fastq_On_Mean_Quality.pl). A last trimming step using the fastx-trimmer command from the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) was undertaken onto the R2 paired files to remove the last six bases of each read to ensure removal of any possible barcode present on short reads.

Bioinformatics

All command lines are available in the Supplementary file.

Using the MinION Fast5 output format, base-calling and demultiplexing were undertaken using the Albacore program v2.5.11 (https://github.com/Albacore/albacore). This generated a fastq file from which reads were filtered out. The average quality score was lower than 7. For each barcode a quality control using the MinionQC R script (https://github.com/roblanf/minion_qc) was done to check for read mean length and quality scores. Reads were then trimmed using Porechop (https://github.com/rrwick/Porechop) in order to remove the sequencing adapters and barcodes. The only non-default setting is that splitting reads containing middle adapters was disabled, in order to avoid issues during the polishing step using Nanopolish (see below).

For each library the percentage of chloroplastic reads was estimated by mapping reads to a reference chloroplast genome using the Burrows-Wheeler alignment tool (bwa mem, https://github.com/lh3/bwa) with "-B 1" option for non-enriched short reads data and "-x ont2d" option for long reads data (Li and Durbin 2009). We then calculated the X-fold enrichment to evaluate capture efficiency (the ratio of chloroplastic reads obtained with capture relative to chloroplastic reads obtained without capture). Coverage and depth values were calculated using Bedtools (Quinlan and Hall 2010) genomecov (https://github.com/arq5x/bedtools2). Mismatch percentage values between mapped reads and references were recovered using Tablet v. 1.17.08.17 (Milne et al. 2010).

De novo assembly of chloroplast genomes

De novo assembly of plastomes based on the long MinION reads, we used the Flye assembler version 2.3 (Kolmogorov et al. 2018). For *O. sativa*, all available reads (17,129) were assembled. For the other species, the number of reads was too high, in excess of 3000x of the reference coverage for some datasets, which caused memory usage issues. To alleviate this, the reads were randomly split into sets of approximately equal size. Each set was then assembled individually using the raw nanopore reads mode. The "min_overlap" parameter (i.e. the minimum overlap between reads), in Flye was adjusted on a species by species basis ranging from 3 000 pb (the default value for our genome size) to 1 000 bp, depending on the medium read length for each species. This was done in order to ensure that a sufficient amount of overlaps were detected for the assembly. The draft assemblies were then polished using Nanopolish version 0.9.1 (https://github.com/jts/nanopolish), using minimap2 on the "map-ont" preset for the overlapping step. Finally, the assemblies were mapped on the reference sequence of each species using the dnadiff tool of MUMmer version 4.0beta2 (Kurtz et al. 2004), which directly provides alignment coordinates and global statistics such as the mean identity percentage of alignments.

Besides read length, the uniformity of coverage of the reference by the reads could also have an impact on the correct assembly of plastomes. This is especially problematic for low molecular weight DNA extractions (in our case from silicagel dried leaves) which resulted in shorter read lengths on average when compared to low molecular weight DNA extractions (in our case from fresh tissue). To test for the impact of the uniformity of reference coverage on the assembly, simulated reads for *P. aequatoralis* (DNA extracted from silicagel dried leaves) were generated using NanoSim v2.1.0 (Yang et al. 2017). A model was first trained on the raw real reads and then 40’000 simulated reads were generated, ensuring they have approximately the same length distribution and error model as the real reads (see results). The simulated reads were then assembled using the same workflow as above.

RESULTS

Plastome enrichment protocol validation on Oryza sativa

After read filtering (Q>7) the median length of the 12 227 mapped plastome reads was of 4 264 bp (Table 1, Figure 2A). We recovered the whole plastome with an average coverage depth of 364X for the enriched MinION library, with a standard deviation increasing from 0.25 to 0.37 between enriched and non-enriched libraries (Figure 2A, B, Table S1). The average mismatch was 11.80% (Table S1). Finally, 70.8% of the reads mapped to the reference chloroplast (Figure 2D, Table S1) representing a ~5-fold increase in chloroplast reads when compared to the non-enriched MiSeq sequenced library (13.32% mapped, Table S1). The longest plastome read recovered was 25 828 bp long (Table 1).

Plastome enrichment protocol applied to non-model species

DNA extraction qualities were variable deepening on the source of the leaf material used. Freshly extracted DNA always produced single bands (not degraded) with fragments higher than 20 kb (Table

S1). Silicagel extracted DNA in contrast was of lower quality generally degraded (smear present) with fragments under 20 kb long (Table S1). For the six non-model species, sequencing of the non-enriched libraries resulted in 0.63% to 7.94% of chloroplast reads (Table S1). In contrast, enriched libraries resulted in 15.7% to 98.2% chloroplast reads, corresponding to a 12 to 161-fold increase in chloroplast DNA sequences (Figure 3, Table 1, Table S1). The mean average of fragments sequenced from freshly extracted DNA was 4 279 bp versus 2 525 bp for DNA extracted from silicagel dry leaves (Figure 4A). Sequences mapped to the reference plastomes ranged mainly from 2 kb and 8 kb, depending on the species (Figure 4A). Average coverage depth was 1,988X for enriched libraries (Figure 4B, Table S1). The longest read mapped to the plastome ranged from 10 405 bp to 25 167 bp for *R. textilis* and *C. americanus*, respectively (Table 1).

De novo assembly of chloroplast genome

When DNA was extracted from fresh leaves, the chloroplast was assembled in two contigs covering most of the reference (Table 1, Figure S1 for a visual exemple in *C. americanus*). Assembled contig lengths varied from 81 053 to 12 5727 bp long. However, the assembler never managed to achieve the full assembly and circularization of a chloroplast into a single contig. For DNA extracted from silicagel dried leaves, where reads were shorter and the coverage more heterogeneous, assembly was suboptimal (Table 1, Figure S2 for a visual exemple in *Phytelephas aequatorialis*) with more final contigs (10-17), uncovered regions and sometimes misassemblies. The longest assembled contigs were also much short than for fresh DNA (Table 2). In addition, the Inverted Repeats (IRs) were also often not differentiated.

Using a simulated dataset of reads uniformly distributed across the chloroplast (Figure S3) and based on the same quality as *P. aequatorialis* significantly improved assembly (Table 2). The assembler resulted in four contigs (versus 13) covering almost 99.72% (versus 87.60%) of the reference and the longest of

107 633 bp (versus 21 797 bp). However, the existence of two distinct repeated regions was still not resolved.

CONCLUSIONS

Here, we show that targeted capture hybridization of long plastome DNA fragments with good coverage (362 x to 3318 x) is possible in plants (Table 1, Table S1). In addition, we show a significant enrichment of our target region (the plastome) when compared to non-enriched data (Figure 2D, 3, Table S1). The different steps of our protocol (Figure 1, Supplementary file) are not fundamentally different from previous chloroplast short read capture protocols (e.g. Mariac et al. 2014) based on in-house probe preparation, shearing, adapter ligation, hybridization and finally capture (Figure 1, Supplementary file). Thus, our approach requires minimal adaptation from previous cost and time effective protocols and should therefore be of broad interest. The main technical change focused on the beads used to capture long DNA fragments. For that we used the kilobaseBINDER^{TM} Kit of Invitrogen which is said to capture DNA fragments longer than 2 kbp. The sizing step we performed at 0.4X using ampure removes fragments smaller than 2 kb and corresponds to the maximum allowed size with the ampure beads. However, other approaches are possible to achieve sizing with higher molecular weight and could be tested (e.g. gel extraction, Automated Size Selection System).

When capturing plastomes across a range of difference species, we find a difference in enrichment percentage ranging from 15.7% to 98.2% of useful reads (Table 1, Figure 4). Differences in genome versus plastome ratios between species can explain the variation of on-target mapped reads percentage compared to non-enriched libraries. In general, species with smaller genomes show higher mapped read percentages. Alternatively, the material used for DNA extractions, the cellular type and the degradation state, can also explain such variations. The low enrichment observed for *P. acaulis* (Table 1, Figure 3) could potentially be linked to a large genome size, although we do not have an estimate of its genome.

A common coverage gap is observed among the chloroplasts of the three palm species due to a region that wasn't covered by the probes (see Faye et al. 2016). Coverage depressiveness of other regions can be explained with biases that occur during DNA shearing, PCR amplification and hybridization capture, considering a CG content effect. Probe bulk normalization from long-range PCR also has to be taken into account. However, global decrease of standard deviation of enriched libraries proved a slight increase of the whole target coverage homogeneity compared with non-enriched libraries. This means that the hybridization capture performed in our protocol didn't introduce more on-target coverage heterogeneity. Nevertheless, applying an alternative capture method such as region-specific extraction (Dapprich et al. 2016) could help maintaining an overall good coverage by accessing high complex, variable, repeat-masked or unknown regions that forbid adequate probe binding.

Probes were designed to hybridize across the whole targeted region (Figure 1), as is generally done using short read approaches (Stull et al. 2013; Mariac et al. 2014). However, a recent study showed that probes targeting small regions are also effective to capture long reads surrounding the targeted region. Indeed, Gasc & Peyret (2017) were able to reconstruct a 21.6 kbp fragment using probes designed for a small 471 bp microbial gene target. This shows that long read capture will also be very useful for targeted sequence capture of nuclear regions.

We demonstrated the capacity of heterologous plastome probes to capture target DNA in other species or genera in Arecacece and Poaceae. For example, probes designed on *P. barteri* hybridized well to other palm genera in different sub families. This underlines the good portability of probes for capturing plastomes across a broad evolutionary spectrum (Stull et al. 2013), even for long fragment capture.

Limits and challenges

Although we were able to successfully capture long plastid fragments using our enrichment protocol, assembling plastomes from this data remains challenging. Indeed, the best assembly resulted in 2 mapped contigs, and the worst one in 17 (Table 2). Assembly of plastomes is well known to be problematic (Twyford & Ness 2017) mainly because of the presence of near identical inverted repeats (IR). Indeed, the similarity of the two IRs is too high for assemblers to decipher between IRs when resolving the assembly graph for the entry and exit point of those sequences. Thus, when the size of the sequenced reads are shorter than the IRs themselves it becomes hard to the correctly assemble of the plastome into a single contig. This is visible for exemple in *C. americanus* (Figure S1) where the resulting two contigs do not across one of the IR regions leading to a failure in reaching a single contig. Of course, this problem is enhanced when dealing with overall shorter reads sequenced from low molecular weight DNA (see Figure S2 for an exemple). In our case, fragment length recovered of DNA extracted from silicagel dried leaves was shorter than those extracted from fresh leaves (Table 1). Moreover, we observed a decrease of the average library fragment size during preparation steps and mainly after PCR because of preferential amplification of shorter fragments, as observed by Giolai et al. (2016) and Eckert et al. (2016).

Optimizing read length in such a way that single reads are longer than the entire IR region should significantly help in the assembly process. In this sense, DNA shearing could be removed in order to increase the average size of the reads. Technical limitations would however be 1) the ability of streptavidin beads to immobilize fragments of tens of thousands of base pairs and 2) the long range PCR amplification step of the enriched fragments which is necessary to produce an input of several hundred ng for the construction of nanopore libraries. The latter is probably the most limiting because it is difficult to produce amplicons of several tens of Kb, and even if we achieve this, representation bias are to be considered. Finally, we show, via simulations, that the uniformity of read coverage across the reference are important for assembly (Table 2). Indeed, uniformly distributed reads, even of lower quality, lead to better assemblages than poor coverage of the reference (Table 2). Therefore, uniform coverage of the reference by the captured reads plays a big role in the correct and improved assembly even for suboptimal DNA extractions.

ACKNOWLEDGEMENTS

This work was publicly funded through ANR (the French National Research agency) under the "Investissement d'Avenir" programme with the reference ANR-10-LABX-001-01 Labex Agro and coordinated by Agropolis Fondation. We thank Oliver Lucas from Oxford Nanopore Technologies and Christian-Julian Villabona-Arenas for help and advice. The exportation and used of the *Phytelephas aequatorialis* sample was authorized by Ecuador under permit MAE-DNB-CM_2018-0082; *Podococcus barteri* under permits AR0020/16; AR0036/15 (CENAREST) authorized by Gabon.

AUTHOR CONTRIBUTIONS

CM, YV, TLPC conceived the idea; RM, TLPC, CM, YV provided material; CM, YV, JFM, SS, AM designed the protocol; KB, CM, MC undertook the experiments; CM, KB, FS, VK analyzed the data. KB, TLPC led the writing; all authors read and commented on the final version.

LITERATURE CITED

Bleidorn C. 2016. Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. Syst. Biodivers. 14:1–8.

Cronn R., Knaus B.J., Liston A., Maughan P.J., Parks M., Syring J.V., Udall J. 2012. Targeted enrichment strategies for next-generation plant biology. Am. J. Bot. 99:291–311.

Dapprich J., Ferriola D., Mackiewicz K., Clark P.M., Rappaport E., D'Arcy M., Sasson A., Gai X., Schug J., Kaestner K.H., others. 2016. The next generation of target capture technologies-large DNA fragment enrichment and sequencing determines regional genomic variation of high complexity. BMC Genomics. 17:486.

Eckert S.E., Chan J.Z.-M., Houniet D., the PATHSEEK consortium, Breuer J., Speight G. 2016. Enrichment by hybridisation of long DNA fragments for Nanopore sequencing. Microb. Genomics. 2.

Faye A., Deblauwe V., Mariac C., Richard D., Sonké B., Vigouroux Y., Couvreur T.L.P. 2016. Phylogeography of the genus Podococcus (Palmae/Arecaceae) in Central African rain forests: Climate stability predicts unique genetic diversity. Mol. Phylogenet. Evol. 105:126–138.

Gasc C., Peyret P. 2017. Revealing large metagenomic regions through long DNA fragment hybridization capture. Microbiome. 5:33.

Giolai M., Paajanen P., Verweij W., Percival-Alwyn L., Baker D., Witek K., Jupe F., Bryan G., Hein I., Jones J.D.G., Clark M.D. 2016. Targeted capture and sequencing of gene-sized DNA molecules. BioTechniques. 61:315–322.

Giolai M., Paajanen P., Verweij W., Witek K., Jones J.D.G., Clark M.D. 2017. Comparative analysis of targeted long read sequencing approaches for characterization of a plant's immune receptor repertoire. BMC Genomics. 18:564.

GOODWIN, S., MCPHERSON, J.D., & MCCOMBIE, W.R. 2016. Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics, 17, 333–351.

JIAO W.-B., SCHNEEBERGER K. 2017. The impact of third generation genomic technologies on plant genome assembly. Curr. Opin. Plant Biol. 36:64–70.

KARAMITROS T., MAGIORKINIS G. 2015. A novel method for the multiplexed target enrichment of MinION next generation sequencing libraries using PCR-generated baits. Nucleic Acids Res. 43:e152–e152.

KOLMOGOROV M., YUAN J., LIN Y., PEVZNER P. 2018. Assembly of Long Error-Prone Reads Using Repeat Graphs. bioRxiv.:247148.

KOREN S., WALENZ B.P., BERLIN K., MILLER J.R., BERGMAN N.H., PHILLIPPY A.M. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27:722–736.

KURTZ S., PHILLIPPY A., DELCHER A.L., SMOOT M., SHUMWAY M., ANTONESCU C., SALZBERG S.L. 2004. Versatile and open software for comparing large genomes. Genome Biol. 5:R12.

LEE H., GURTOWSKI J., YOO S., NATTESTAD M., MARCUS S., GOODWIN S., MCCOMBIE W.R., SCHATZ M. 2016. Third-generation sequencing and the future of genomics. bioRxiv.:048603.

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., FENNELL T., RUAN J., HOMER N., MARTH G., ABECASIS G., DURBIN R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25:2078–2079.

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. Nat Meth. 7:111–118.

Mariac C., Scarcelli N., Pouzadou J., Barnaud A., Billot C., Faye A., Kougbeadjo A., Maillol V., Martin G., Sabot F., Santoni S., Vigouroux Y., Couvreur T.L.P. 2014. Cost effective enrichment hybridization capture of chloroplast genomes at deep multiplexing levels for population genetics and phylogeography studies. Mol. Ecol. Resour. 14:1103–1113.

Milne I., Bayer M., Cardle L., Shaw P., Stephen G., Wright F., Marshall D. 2010. Tablet—next generation sequence assembly visualization. Bioinformatics. 26:401–402.

Quinlan A.R., Hall I.M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26:841–842.

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

Scarcelli N., Barnaud A., Eiserhardt W., Treier U.A., Seveno M., d’Anfray A., Vigouroux Y., Pintaud J.-C. 2011. A Set of 100 Chloroplast DNA Primer Pairs to Study Population Genetics and Phylogeny in Monocotyledons. PLoS ONE. 6:e19954.

Straub S.C.K., Parks M., Weitemier K., Fishbein M., Cronn R.C., Liston A. 2012. Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. Am. J. Bot. 99:349–364.

Stull G.W., Moore M.J., Mandala V.S., Douglas N.A., Kates H.-R., Qi X., Brockington S.F., Soltis P.S., Soltis D.E., Gitzendanner M.A. 2013. A Targeted Enrichment Strategy for Massively Parallel Sequencing of Angiosperm Plastid Genomes. Appl. Plant Sci. 1:1200497.

YANG, C., CHU, J., WARREN, R. L., & BIROL, I. 2017. NanoSim: Nanopore sequence read simulator based on statistical characterization. GigaScience.

Tables

Table 1: MinION plastome enriched library output data. The percentage of “plastome mapped reads” was calculated using BWA to indicated reference plastomes.

Species	DNA	probe origin	total number of reads	Median read length	longest read	% of plastid reads	longest plastome read	Median plastome read length
Oryza sativa	fresh	*O. sativa*	17129	4627	26128	70,8	25828	4264
Oryza glaberrima	fresh	*O. sativa*	81361	3695	24804	98,2	24504	3398
Cenchrus americanus	fresh	*C. americanus*	105760	4914	25468	97,0	25167	4623
Digitaria exilis	fresh	*C. americanus*	141250	3783	19378	94,4	19078	3489
Podococcus acaulis	silicagel	*P. barteri*	202924	2486	13103	15,7	12805	2129
Raphia textilis	silicagel	*P. barteri*	83833	2322	10705	87,5	10405	1997
Phytelephas aequatorialis	silicagel	*P. barteri*	202925	2437	15132	79,0	14832	2158

Table 2: De novo assembly results from real (6) and simulated (1) data in number of contigs, and coverage and identity percentages to the respective reference plastome genomes (see Table 1). Min overlap is the minimum overlap between reads as defined in Flye. The simulated data was based on the out pu results of *P. aequatorialis*.

Species	Min overlap	Plastid contigs	Coverage %	Identity %	Longest contig
Oryza glaberrima	3000	2	92.32	99.14	109087
Cenchrus americanus	3000	2	99.91	98.52	81053
Digitaria exilis	3000	2	99.97	99.18	125727
Podococcus acaulis	1000	17	81.24	98.86	22803
Raphia textilis	1000	10	83.87	98.84	21797

Phytelephas aequatorialis	1000	13	87.60	98.31	20700
Simulated assembly	1000	4	99.72	99.05	107633

Figures (next page)

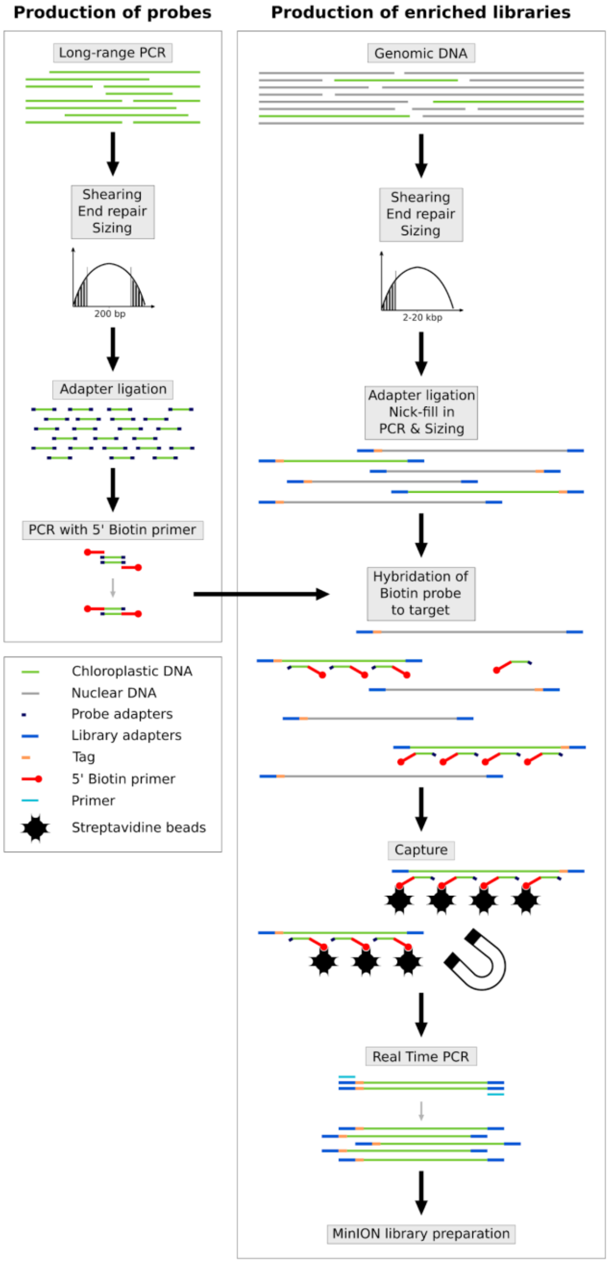

Figure 1: Schematic representation of the protocol used for long sequence capture of plastomes (modified from Mariac et al. 2014).

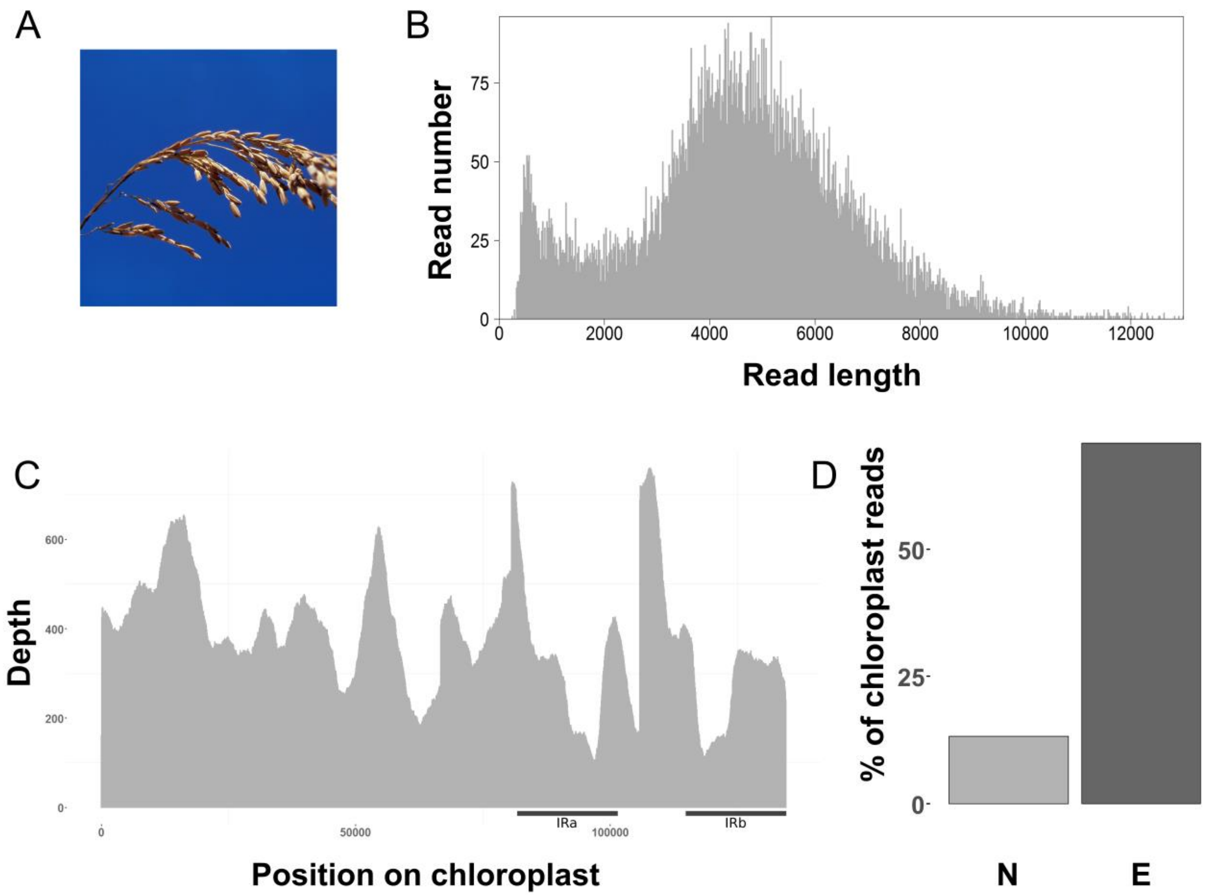

Figure 2: Long fragment capture results for *Oryza sativa*. A: Panicule of *Oryza sativa* (Jean-Pierre Montoroi IRD ©). B: number of reads per read length before mapping. C: Plastome coverage after mapping. Black bars indicate approximate position of both inverted repeats (IR). D: Percentage of useful reads mapped to *Oryza sativa* reference plastome (KT289404.1) between non enriched (N, light grey) and enriched (E, dark grey) libraries.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27411v1 | CC BY 4.0 Open Access | rec: 5 Dec 2018, publ: 5 Dec 2018

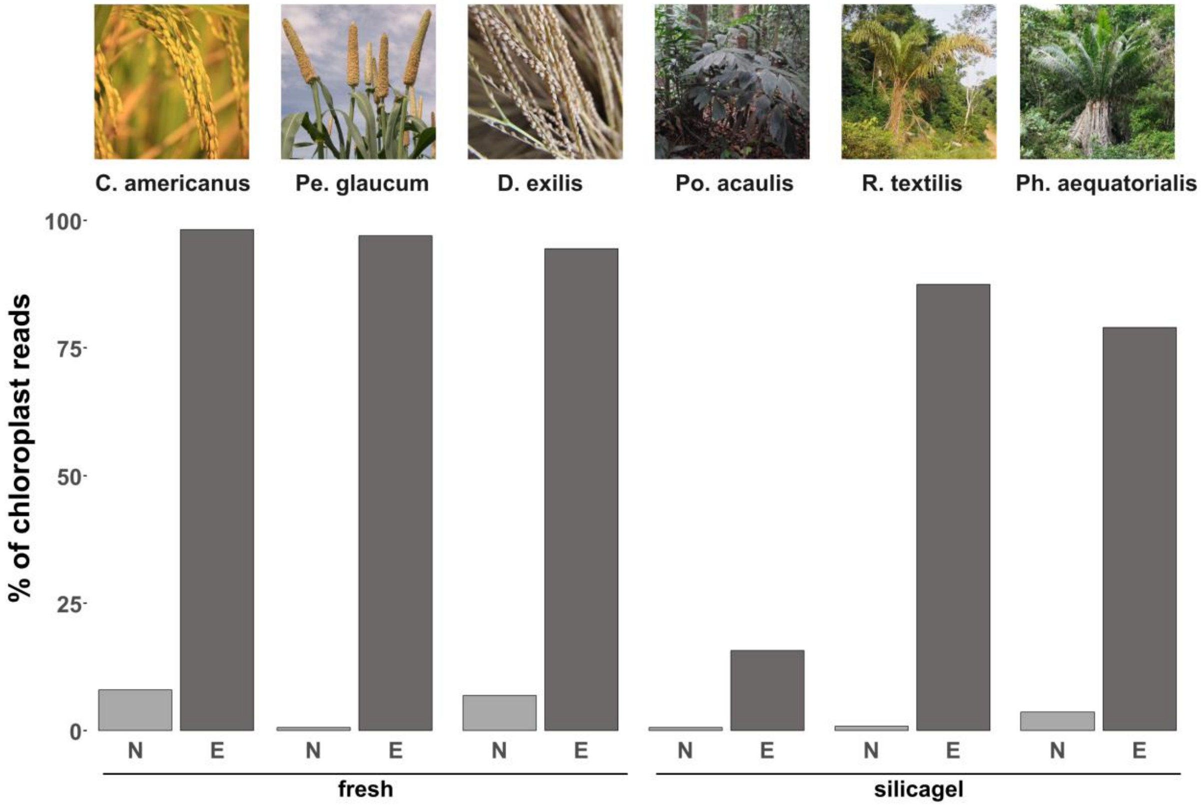

Figure 3: Percentage of useful reads mapped to their respective reference plastome (see Table 1) between Illumina non enriched (N, light grey) and MinION enriched (E, dark grey) protocols for the 6 non model species in our study. Photos: *O. glaberrima*: https://pxhere.com/fr/photo/706162; CC0 public domain; *Cenchrus americanus*: C. Mariac IRD ©; *D. exilis*: A. Barnaud IRD © ; *Po. acaulis, R. textilis, Ph. aequatorials*: TLP Couvreur IRD ©.

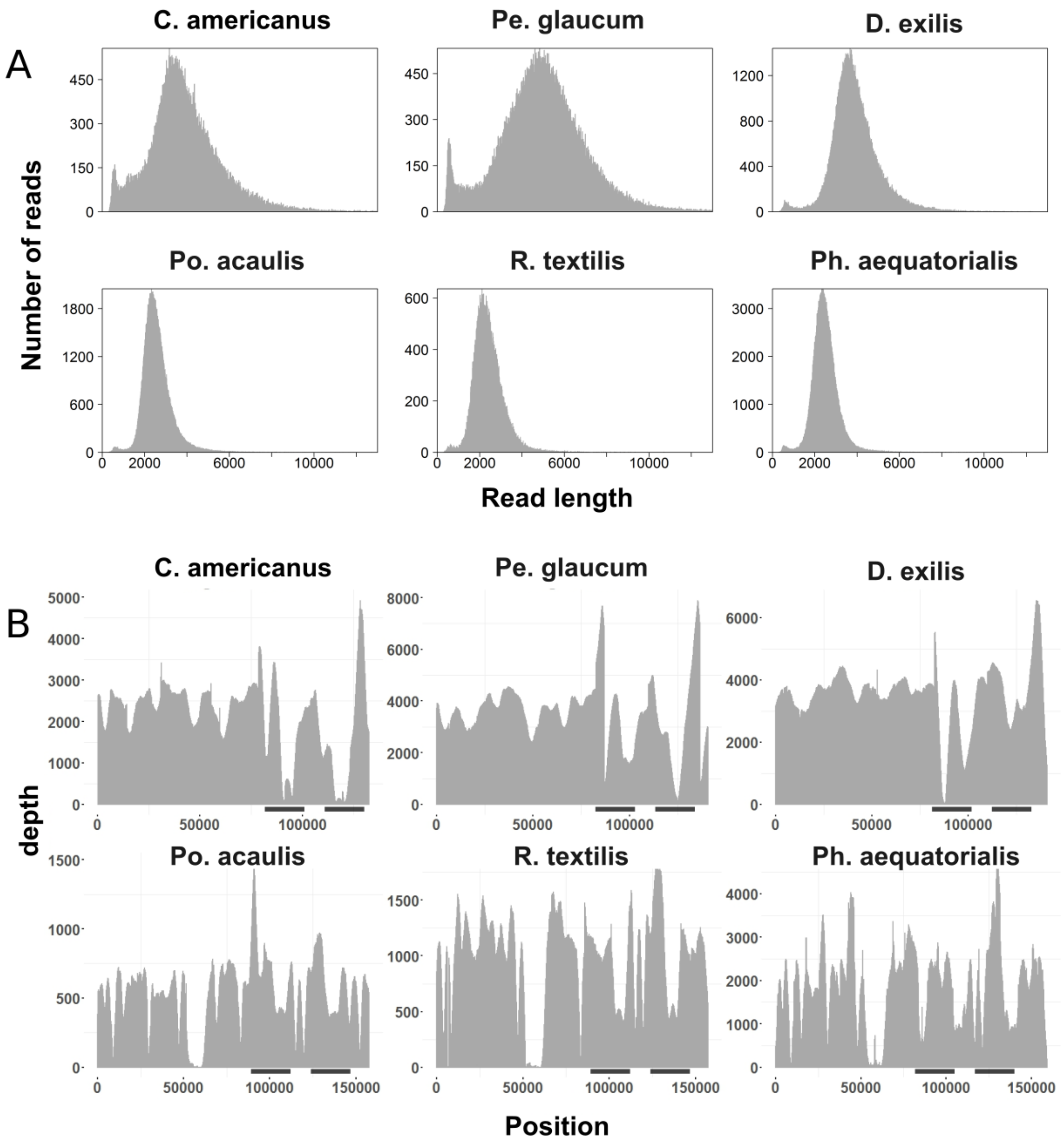

Figure 4: Long fragment capture results for six non model plant species. A: Number of reads per read length before mapping. B: Plastome coverage results from the enriched long read capture protocol. Black bars indicate approximate position of both inverted repeats (IR).