FrangiPANe, a tool for creating a panreference using left behind reads

Tranchant-Dubreuil Christine[1*†], Chenal Clothilde[1,2,3†], Blaison Mathieu[1], Albar Laurence[4], Klein Valentin[1], Mariac Cédric[1], Wing Rod A. [5], Vigouroux Yves [1*], Sabot Francois[1*]

[1] DIADE, Univ Montpellier, CIRAD, IRD, 911 Avenue Agropolis, 34934/34830 Montpellier Cedex 5, France.
[2] MIVEGEC, Univ Montpellier, CNRS, IRD, 911 Avenue Agropolis, 34934/34830 Montpellier Cedex 5, France
[3] ISEM, Univ Montpellier, CNRS, IRD, EPHE, CIRAD, INRAP, 1093-1317 Route de Mende, 34090 Montpellier, France
[4] PHIM Plant Health Institute, Univ Montpellier, CIRAD, INRAE, Institut Agro, IRD, Montpellier, France
[5] Center for Desert Agriculture, Biological and Environmental Sciences & Engineering Division (BESE), King Abdullah University of Science and Technology (KAUST), Thuwal, 23955-6900, Saudi Arabia.
[†]: Co-first author. The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

* To whom correspondence should be addressed. Email: christine.tranchant@ird.fr

Correspondence may also be addressed to francois.sabot@ird.fr, yves.vigouroux@ird.fr

ABSTRACT

We present here FrangiPANe, a pipeline developed to build panreference using short reads through a map-then-assemble strategy. Applying it to 248 African rice genomes using an improved CG14 reference genome, we identified an average of 8 Mb of new sequences and 5,290 new contigs per individual. In total, 1.4 G of new sequences, consisting of 1,306,676 contigs, were assembled. We validated 97.7% of the contigs of the TOG5681 cultivar individual assembly from short reads on a newly long reads genome assembly of the same TOG5681 cultivar.

FrangiPANe also allowed the anchoring of 31.5% of the new contigs within the CG14 reference genome, with a 92.5% accuracy at 2kb span. We annotated in addition 3,252 new genes absent from the reference.

FrangiPANe was developed as a modular and interactive application to simplify the construction of a panreference using the map-then-assemble approach. It is available as a Docker image containing (i) a Jupyter notebook centralizing codes, documentation and interactive visualization of results, (ii) python scripts and (iii) all the software and libraries requested for each step of the analysis.

INTRODUCTION

Nowadays, an increasing number of studies highlights the limit of using a single individual genome to assess genomic diversity within a species (1, 2, 3, 4). For instance, in plants, between 8% and 27% of genes varied in presence/absence across individuals from the same species (5, 6, 7). Pangenomics offers an alternative way to study gene content variations and more broadly the whole genomic

variations within a population. Initially introduced in bacteria by Tettelin et al. (8), the pangenome concept refers to the replete genomic content of a species, consisting in (i) the core genome, shared among all individuals, and (ii) the dispensable genome, shared only in a subset of individuals. With the decrease in sequencing costs, pangenomics analyses are more and more frequent in plants (9, 10, 11, 12) and animals (13, 14, 15, 16).

The pivotal step in any pangenomic analysis is the construction of a panreference that captures the (almost) full diversity of a large set of genomes. However, the pangenome construction remains a cumbersome and challenging process, especially for Eukaryotes due to their large genome size and complexity (e.g. repeat content or polyploidy). Although long reads sequencing technologies are increasingly used to directly detect large structural variations, generally through reassembly of genomes (12, 17, 18, 19), short reads ones currently remain less expensive and is still widely used. In addition, the numerous short reads datasets already available on many organisms provide an important source of data to perform large-scale pangenomic analyses. Two approaches were mainly used for the pangenome construction from individuals sequencing with such short reads : (i) *de novo* genome assembly followed by genome comparison (here referred as the "assemble-then-map" approach; 10, 20), and (ii) the "map-then-assemble" approach, based on the mapping of resequencing short reads followed by the *de novo* assembly of the unmapped reads (6, 13, 21).

Very few tools are publicly available to perform all the steps at once, being either developed for bacteria (22, 23, 24), or based on the *de novo* "assemble-then-map" approach (25).

We present here a method based on the "map-then-assemble" approach to (i) identify large fragments absent from a reference genome using short reads data, (ii) locate these variations on the reference, and (iii) build a panreference. For that purpose, we developed frangiPANe, a pipeline tool to easily apply this approach and to create an accurate panreference for any organism using short reads data.

To validate our method and frangiPANe, we used the resequencing data from 248 genomes (Cubry et al, 2018; Monat et al, 2016) of the cultivated African rice (*Oryza glaberrima*) and of its closest wild relative (*O. barthii*) as a proof-of-concept to build the first panreference for African rice.

MATERIAL AND METHODS

Sample sequencing

Short-Read sequencing data. We used whole genome sequencing data (Table 1) from 248 African rice accessions previously described in Cubry et al. (2018) and Monat et al. (2016) (Illumina technology TrueSeqv3, 100-150 bp paired-end reads), including 164 domesticated and 84 wild relative individuals. These samples covered the full range of genetic diversity in the two African rice species *O. glaberrima* and *O. barthii* (28).

TOG5681 long reads sequencing and assembly. DNA was extracted following the protocol from Serret et al. (29) using an adapted CTAB-lysis approach to ensure high molecular weight DNA. DNA quality and concentration were controlled using PFGE and Qubit, and subject to a LSK-109 library as

recommended by suppliers (Oxford Nanopore Technology, Inc, Oxford, United Kingdom). The library was loaded on two 9.4.1 flowcells, raw FAST5 base-called using Guppy 4.0.5 (hac model) with a cut-off at PhredQ 7, and FASTQ data were controlled using NanoPlot 1.38.1 (30). FASTQ were then assembled using Flye 2.8 (31) with the –nano-raw mode and standard options. Initial polishing was ensured by 3 turns of Racon 1.3 (32) under standard conditions using the initial set of nanopore reads and mapping performed by Minimap2 v2.10 (33) in -x map-ont mode. Final polishing was performed using Medaka 1.2 (https://Github.com/Nanoporetech/Medaka) with the standard model. Contamination was checked using Blobtools 1.1 (34) and remapping of short reads, as recommended, in the same way as described below. Final chromosome-scale scaffolding was done using RagTag 2.1 (35) using the CG14 OMAPv2 as reference sequence. BUSCO v5.0 (36) with *Viridiplantae* database v10 was used for computing the gene space completion, and all basic statistics on contigs and scaffolds were obtained using QUAST 5.0 (37).

"Map-then-assemble" approach

This approach starts with mapping of resequencing pair-ended short reads on a reference genome, followed by *de novo* assembly of unmapped reads for each sample. Next, all contigs are pooled after filtering. The last two steps consist in reducing the sequence redundancy at intra- and inter-population level and into placing the non-redundant sequences on the reference genome. Figure 1 provides an overview of the "map-then-assemble" approach.

Alignment to the reference genome. The CG14 reference genome was downloaded from the European Nucleotide Archive (ENA) (Accession GCA_000147395, https://www.ebi.ac.uk/ena/browser/view/GCA_000147395). For each accession, pair-ended short reads were aligned to the CG14 reference with bwa aln (option -n=5) and bwa sampe (v. 0.7.15) (38). Mapping results were sorted with Picardtools sortSam (version 2.6.0). Samtools view (version 1.3.1) (39) was used to extract unaligned reads (option -F 2).

Unmapped reads assembly and filtering. For each sample, unmapped pair-ended short reads were assembled with Abyss-pe (v. 2.0.2) (40). We first optimized k-mer size for assembly in the AA accession (Supplementary notes and Supplementary table 1), and choosed a k-mer size (option -k 64) maximizing N50 and minimizing both contigs number and L50. Contigs shorter than 300 bp were excluded.

We screened contigs for vector sequences using Vecscreen (https://www.ncbi.nlm.nih.gov/Tools/Vecscreen/about/) and the NCBI UniVec_core database (V.build 10.0, http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html). Contigs were aligned with blastn against the NCBI NT nucleotide database (Oct 27, 2019) and the rice organites genomes (mitochondrial and chloroplast). Contigs with best hits from outside the green plants taxon (Viridiplantae) or on rice organites genomes were removed.

Reducing redundancy. Contigs from all individuals were clustered using CD-HIT (version 4.6, options -c 0.80 -s 0.95) (41). Only the longest sequence for each cluster was conserved (Supplementary Notes and Supplementary Table 2).

Contigs position on the genome reference. Pair-ended shorts reads from each individual were remapped to a new cumulative reference formed of the CG14 assembly and of the non-redundant contigs. Mapping to this panreference was performed using bwa aln and sampe with the same parameters as described before. Pair-ended short reads mapping on both a contig and a chromosome were used to anchor contigs (9; 13). Briefly, all reads aligned within the first or last 300 bases of a contig and for which mates mapped on a CG14 chromosome were pre-selected. Contigs position within chromosomes was considered as valid if : (i) at least 10 reads with MAPQ 10 are aligned on the same contig and their mates on the same chromosome, and (ii) the positions of the 10 mate reads on the chromosome are all located in a span shorter than 2kb.

Assembly validation and position validation on chromosomes. We used the TOG5681 genome assembly based on long reads sequencing to validate (i) our contigs assembly from TOG5681 pair-ended short reads, and (ii) their position on chromosomes. We used the nucmer tool (MUMmer v4.0beta3) (42) and kept only alignment showing 90% identity and 80% coverage of the contig, with a minimum aligned sequence length of 300 bp.

Panreference annotation

Transposable Element Identification. Transposable elements were detected using RepeatMasker (v.4.0.7) (43) with the RiTE-db (v1.1) (44) and the RepBase (v23.11, *Oryza* section) (45) databases.

Genes mapping. Annotation of the panreference was performed using Liftoff (v1.6.1) (46) with annotated genes from the Nipponbare reference (*Oryza sativa* ssp. *japonica* cv. Nipponbare, IRGSP-1.0-1-2021-11-11 release). Genes were considered as successfully mapped if a minimum of 50% of the Nipponbare gene was aligned to the panreference with a sequence identity higher than 50% (options -s 0.5 -a 0.5). Gene copies were annotated using a minimum of sequence identity threshold of 95% (options -copies -sc 0.95).

Gene Ontology annotation. Genes sequences were aligned to the NCBI NR protein database (Sept 09, 2021 - *Viridiplantae* section) using blastx (options -e-value 1e-6). Genes with protein domain signatures were recovered using InterProScan(v5.53.87; options –goterms –iprlookup –pathway) (47). GO annotation and enrichment analysis were carried out through the Blast2GO (v.0.3, with default options, Fisher' exact test with a cutoff of *p*-value 0.05;) (48).

A tool to build panreference from scratch

FrangiPANe was developed as a modular and interactive application to simplify the construction of a panreference using the map-then-assemble approach (Figure 1). It is available as a Docker image

containing (i) a Jupyter notebook centralizing codes, documentation and interactive visualization of results, (ii) python scripts and (iii) all the softwares and libraries requested for each step of the analysis. Supplementary Table 3 presents the main list of tools required by FrangiPANe.

The code, documentation, installation manual and test data are available under the GPLv3 and CC4.0 BY-NC license at https://github.com/tranchant/frangiPANe. A dedicated virtual machine is also available on the BioSphere Cloud of the French Institute of Bioinformatics https://biosphere.france-bioinformatique.fr/catalogue/).

RESULTS

The CG14 and TOG5681 genomes

We relied on an improved reference genome of the cultivar CG14 from *Oryza glaberrima* from the OMAP consortium (Accession GCA_000147395, https://www.ebi.ac.uk/ena/browser/view/GCA_000147395) and on a new whole genome assembly of the cultivar TOG5681 (see below for details), both accessions being themselves part of the 248 accessions sequenced using short reads (27).

TOG5681 control genome. We obtained 509,485 ONT long reads of minimal PhredQ 7, for 6.612 Gb of data (~18x) with a N50 of 23.8 kb. After assembly and polishing, the final dataset represents 148 contigs, for a total assembly size of 348,131,590 bases, a N50 of 15,386,152 bases (L50 of 9), and 99.5% of the assembly being comprised in contigs larger than 50kb. The BUSCO score for this assembly is 95%, including 2.1% of duplicated target genes. Blobtools indicated only 3 contamination contigs, representing less than 0.01% of the total size. After removal of these contaminated contigs, RagTag was used on the remaining 145 contigs to scaffold the TOG5681 genome using the CG14 one as reference, with 99.4% of the bases placed, leading to a final chromosome scale assembly of 59 contigs (12 chromosomes + 47 unplaced contigs) representing 348,140,190 bases.

Building african rice panreference

To identify sequences absent from the CG14 genome, we used short reads sequencing data of 164 domesticated and 84 wild relatives, all of which exceeding a sequencing depth of 20X (Table 1). The mean mapping rate of these 248 genomes was high, with 96% and 97.8% for *O. barthii* and *O. glaberrima*, respectively. The mapping rate decreased respectively to 93.7% and 96.2% considering only reads correctly mapped in pairs (Supplementary Figure 1).

Unmapped reads assembly produced a total of 2.9 Gb of sequences and 9,887,127 contigs. After filtering for adapter (1% of sequences), alien sequences (0.01%) and minimal size (3.3 %), we ended up with 1.65 Gb and 1,306,676 contigs. On average, 8 Mb of sequences and 5,290 contigs were obtained per individual (from 1.4 to 25.2 Mb assembled per individual and a contigs number ranging from 1,008 and 49,949 per individual, Table 2). The exception was CG14, for which we assembled a few 633 contigs, each with a very small size (303 bp on average).

After reducing redundancy, we identified 513.5 Mb (484,394 contigs) with an average sequence size of 1,060 bp (ranging from 301 bp up to 83,704 bp, Supplementary Figure 1). 56.4% of these non-redundant sequences were identified as singleton (Supplementary Table 4, Supplementary Figure 2).

Contigs anchoring on the reference genome. We remapped all pair-ended short reads on the panreference consisting of the cumulation of the CG14 genome and of the newly deduplicated assembled sequences (484,394 contigs). We increased the mapping rate by 0.9 and 2.3% for the domesticated and the wild relative accessions, respectively (98.7% and 98.3%).

Using the pair-end mapping information, we accurately placed 31.5% of the non-redundant contigs (152,411 contigs) at a unique position on the reference genome (145 Mb; Fig. 2). A total of 39,630 contigs (8.2 %) were placed at multiple positions, on the same chromosome or not (31 Mb). Finally, 292,353 contigs (60.3%) remained unplaced (representing a total of 337 Mb).

The assembled contigs from TOG5681 short reads data (7.9Mb, 5,318 contigs) were recovered at 97.7% on the corresponding long reads assembled genome. A total of 1,696 contigs (31.9%) from this accession were placed with a high confidence at a unique position on the CG14 genome. 95.1% of these 1,696 contigs also mapped against the TOG5681 genome with a coverage of 100%. We realigned on the CG14 genome the TOG5681 1kbp-flanking sequences surrounding the aligned contigs and 92.5% of them were found at the same position on the CG14 genome, thus validating the anchoring approach.

Panreference Annotation

In total, 52.1 % of the panreference was annotated as repetitive elements, including retrotransposons (25.3%), DNA transposons (16.3%) and unclassified elements (10.5%) (Supplementary Table 5). The transposable elements (TEs) content ratio was twice higher in contigs (67.6%) than in the genome reference (29.2%). We also observed a higher percentage of DNA transposons within the contigs (34%) than in the reference genome (26.5%), especially regarding the ones being anchored on the genome (42.5%) (Supplementary Table 5 and Supplementary Figure 3).

Out of 37,864 genes annotated from the Nipponbare genome, 95.5% of genes (36,159 genes) were successfully mapped on the panreference, including 35,252 genes on chromosomes and 907 on all non redundant contigs. The average sequence identity in exons of mapped genes was 96% and the average alignment coverage was 98% (Supplementary Table 6 and Supplementary Figure 4).

98.7% of these genes are placed on the same chromosome on the Nipponbare and the CG14 genomes respecting the co-linearity of gene order between genomes (Supplementary Figure 5).

In addition to the successfully mapped genes, we found 2,631 additional gene copies, 281 and 2,345 on the CG14 genome and on contigs sequences, respectively.

Genes present in the contigs were enriched in GO related to detoxification, response to chemical and response to toxic substance (Supplementary Table 7).

DISCUSSION

Understanding plant genomic diversity requires reliable tools to rapidly build up pangenome sequences. We present here a framework to develop such an approach and apply it on 248 African rice genomes.

Overall, the results of FrangiPANe on African rice are in agreement with the ones from its cousin species the Asian rice (2, 12). In total, we identified 513 Mb of new sequences, in addition to the 344 Mb reference genome. The new sequenced part is in accordance with the Asian rice one, which ranges from 268 Mb (2) to 1.3 Gbp (12). The lowest value (268 Mb) is based on short reads from 3,010 Asian rice genomes (2), an approach similar to ours. We found twice as many new sequences for a smaller set of individuals, but we also included wild rice species. Generally, newly assembled dispensable genomic sequences are generally enriched in TEs. For instance in Asian rice, 52.7% of the newly assembled sequence were TEs (12), compared to an expected number of 35% in the reference genome (49). Our re-assembled sequence using short reads data was composed of 53% of TEs in African rice, almost identical to the one estimated with a long reads approach on Asian rice (52.7%; (12)). In terms of gene number, using 66 Asian rice, Zhao et al. (2018) found 10,872 new genes (50), so, roughly in average, 165 genes per individual. Using the 3,010 Asian rice genomes (2), a total of 12,465 novel full-length genes were detected, representing an average of 4 per genome. Here, we found 13 genes per genome (3252 genes in total), three times more than the 3,010 genome study and 10 times less than in the 66 ones. The large disparity between these estimations might lay in the stringency of gene calling and in the procedure of annotation. In our case, we certainly underestimate the number of genes, as we only used a transfer of annotation. *De novo* annotation should thus allow identification of additional new genes specific to the African rice.

Our tool presents several improvements compared to other available tools. These were developed primarily for bacterial species (24, 51, 52) using short reads sequencing data (53,54) such as PanSeq (22), PGAP (55), roary (23) or PanX (24). They are mostly gene-oriented tools, however, and were specifically designed and tested on the small and simple bacterial genomes. In Eukaryotes, HUPAN (25), a command line tool, has been developed and applied to rice and human (25, 20). This tool starts with *de novo* genome assembly of each individual, followed by mapping of contigs upon the reference genome, and finally clustering of all unaligned contigs ("assemble-then-map" approach). However, such assembly based on short reads lead to missing regions and repeat compression.

We proposed here FrangiPANe as a new solution relying on a massively parallelizable approach, based on the "map-then-assemble" pathway. Our tool proved to be particularly accurate with 97.7% of assembled contigs from the TOG5681 accession also present in a new long reads genome assembly of the same TOG5681 individual.

FrangiPANe also provides a complete environment for panreference creation through an unique and interactive interface, without requiring huge programming skills or the installation of numerous bioinformatic softwares. Based on Docker (https://docs.docker.com/get-docker) and Jupyter

(https://jupyter.org/), it streamlines the whole process involving multiple analysis steps and the data visualization in different way (e.g. tables or plot) within a single well-documented notebook.

While long reads *de novo* genome assembly offers new opportunities to perform pangenome analysis (11,18,19), the vast majority of currently available datasets are from short reads Illumina technology, and are generally very large in term of number of individuals. FrangiPANe offers opportunities to take advantage of these datasets to gain a better understanding of plant and animal genomic diversity.

AVAILABILITY

frangiPANe is freely available in the GitHub repository https://github.com/tranchant/frangiPANe, under the double licence CeCiLL-C (http://www.cecill.info/licences/Licence_CeCILL-C_V1-en.html) and GNU GPLv3.

A virtual machine is also available at the BioSphere service of the French Institute of Bioinformatics (https://biosphere.france-bioinformatique.fr/catalogue).

The sequences have been deposited in the IRD dataverse: Tranchant, Christine; Chenal, Clothilde; Blaison, Mathieu; Albar, Laurence; Klein, Valentin; Mariac, Cédric; Wing, Rod; Vigouroux, Yves; Sabot, Francois, 2022, "Supporting data for the African Rice Panreference produced by the frangiPANe software", https://doi.org/10.23708/93OQMD, DataSuds, V1.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENT

The authors acknowledge Ndomassi Tando and the ISO 9001 certified IRD itrop HPC (member of the South Green Platform) at IRD montpellier as well as the TGCC platform for providing HPC resources that have contributed to the research results reported within this paper. (URL: https://bioinfo.ird.fr/-http://www.southgreen.fr). They also thank Christophe Blanchet and the French Institute of bioinformatics (IFB) to provide access to the aplliance frangiPANe through the cloud Biosphere.

FUNDING

This work was supported by a grant from the France Genomique French National infrastructure and funded as part of ‘‘Investissement d’avenir’’ (ANR-10-INBS-09) and the IRIGIN project (http://irigin.org).

CONFLICT OF INTEREST. None declared.

REFERENCES

1. Springer,N.M., Ying,K., Fu,Y., Ji,T., Yeh,C.-T., Jia,Y., Wu,W., Richmond,T., Kitzman,J., Rosenbaum,H., *et al.* (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet.*, **5**, e1000734.

2. Wang,W., Mauleon,R., Hu,Z., Chebotarov,D., Tai,S., Wu,Z., Li,M., Zheng,T., Fuentes,R.R., Zhang,F., *et al.* (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*, **557**, 43–49.

3. Tranchant□Dubreuil,C., Rouard,M. and Sabot,F. (2019) Plant Pangenome: Impacts on Phenotypes and Evolution. In *Annual Plant Reviews online*. Wiley, pp. 453–478.

4. Bayer,P.E., Golicz,A.A., Scheben,A., Batley,J. and Edwards,D. (2020) Plant pan-genomes are the new reference. *Nat. Plants 2020 68*, **6**, 914–920.

5. Schatz,M.C., Maron,L.G., Stein,J.C., Wences,A., Gurtowski,J., Biggers,E., Lee,H., Kramer,M., Antoniou,E., Ghiban,E., *et al.* (2014) Whole genome de novo assemblies of three divergent strains of rice, Oryza sativa , document novel gene space of aus and indica. *Genome Biol.*, **15**, 506.

6. Gao,L., Gonda,I., Sun,H., Ma,Q., Bao,K., Tieman,D.M., Burzynski-Chang,E.A., Fish,T.L., Stromberg,K.A., Sacks,G.L., *et al.* (2019) The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet.*, **51**, 1044–1051.

7. Hübner,S., Bercovich,N., Todesco,M., Mandel,J.R., Odenheimer,J., Ziegler,E., Lee,J.S., Baute,G.J., Owens,G.L., Grassa,C.J., *et al.* (2019) Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. *Nat. Plants*, **5**, 54–62.

8. Tettelin,H., Masignani,V., Cieslewicz,M.J., Donati,C., Medini,D., Ward,N.L., Angiuoli,S. V., Crabtree,J., Jones,A.L., Durkin,A.S., *et al.* (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial “pan-genome.” *Proc. Natl. Acad. Sci.*, **102**, 13950–13955.

9. Golicz,A.A., Bayer,P.E., Barker,G.C., Edger,P.P., Kim,H., Martinez,P.A., Chan,C.K.K., Severn-Ellis,A., McCombie,W.R., Parkin,I.A.P., *et al.* (2016) The pangenome of an agronomically important crop plant Brassica oleracea. *Nat. Commun.*, **7**, 13390.

10. Gordon,S.P., Contreras-Moreira,B., Woods,D.P., Des Marais,D.L., Burgess,D., Shu,S., Stritt,C., Roulin,A.C., Schackwitz,W., Tyler,L., *et al.* (2017) Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure. *Nat. Commun.*, **8**, 2184.

11. Liu,Y., Du,H., Li,P., Shen,Y., Peng,H., Liu,S., Zhou,G.-A., Zhang,H., Liu,Z., Shi,M., *et al.* (2020) Pan-Genome of Wild and Cultivated Soybeans. *Cell*, **182**, 162-176.e13.

12. Qin,P., Lu,H., Du,H., Wang,H., Chen,W., Chen,Z., He,Q., Ou,S., Zhang,H., Li,X., *et al.* (2021) Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell*, **184**, 3542-3558.e16.

13. Sherman,R.M., Forman,J., Antonescu,V., Puiu,D., Daya,M., Rafaels,N., Boorgula,M.P., Chavan,S., Vergara,C., Ortega,V.E., *et al.* (2019) Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat. Genet.*, **51**, 30–35.

14. Gerdol,M., Moreira,R., Cruz,F., Gómez-Garrido,J., Vlasova,A., Rosani,U., Venier,P., Naranjo-Ortiz,M.A., Murgarella,M., Greco,S., *et al.* (2020) Massive gene presence-absence variation shapes an open pan-genome in the Mediterranean mussel. *Genome Biol. 2020 211*, **21**, 1–21.

15. Li,R., Fu,W., Su,R., Tian,X., Du,D., Zhao,Y., Zheng,Z., Chen,Q., Gao,S., Cai,Y., *et al.* (2019) Towards the Complete Goat Pan-Genome by Recovering Missing Genomic Segments From the Reference Genome. *Front. Genet.*, **10**, 1169.

16. Tian,X., Li,R., Fu,W., Li,Y., Wang,X., Li,M., Du,D., Tang,Q., Cai,Y., Long,Y., *et al.* (2020) Building a sequence map of the pig pan-genome from multiple de novo assemblies and Hi-C data. *Sci. China. Life Sci.*, **63**, 750–763.

17. Song,J.M., Guan,Z., Hu,J., Guo,C., Yang,Z., Wang,S., Liu,D., Wang,B., Lu,S., Zhou,R., *et al.* (2020) Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. *Nat. Plants 2020 61*, **6**, 34–45.

18. Jayakodi,M., Padmarasu,S., Haberer,G., Bonthala,V.S., Gundlach,H., Monat,C., Lux,T., Kamal,N., Lang,D., Himmelbach,A., *et al.* (2020) The barley pan-genome reveals the hidden legacy of mutation breeding. *Nat. 2020 5887837*, **588**, 284–289.

19. Walkowiak,S., Gao,L., Monat,C., Haberer,G., Kassa,M.T., Brinton,J., Ramirez-Gonzalez,R.H., Kolodziej,M.C., Delorean,E., Thambugala,D., *et al.* (2020) Multiple wheat genomes reveal global variation in modern breeding. *Nat. 2020 5887837*, **588**, 277–283.

20. Hu,Z., Sun,C., Lu,K.C., Chu,X., Zhao,Y., Lu,J., Shi,J. and Wei,C. (2017) EUPAN enables pan-genome studies of a large number of eukaryotic genomes. *Bioinformatics*, **33**, 2408–2409.

21. Hufnagel,B., Soriano,A., Taylor,J., Divol,F., Kroc,M., Sanders,H., Yeheyis,L., Nelson,M. and Péret,B. (2021) Pangenome of white lupin provides insights into the diversity of the species. *Plant Biotechnol. J.*, **19**, 2532–2543.

22. Laing,C., Buchanan,C., Taboada,E.N., Zhang,Y., Kropinski,A., Villegas,A., Thomas,J.E. and Gannon,V.P.J. (2010) Pan-genome sequence analysis using Panseq: An online tool for the rapid analysis of core and accessory genomic regions. *BMC Bioinformatics*, **11**, 1–14.

23. Page,A.J., Cummins,C.A., Hunt,M., Wong,V.K., Reuter,S., Holden,M.T.G., Fookes,M., Falush,D., Keane,J.A. and Parkhill,J. (2015) Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, **31**, 3691.

24. Ding,W., Baumdicker,F. and Neher,R.A. (2018) panX: pan-genome analysis and exploration. *Nucleic Acids Res.*, **46**, e5.

25. Duan,Z., Qiao,Y., Lu,J., Lu,H., Zhang,W., Yan,F., Sun,C., Hu,Z., Zhang,Z., Li,G., *et al.* (2019) HUPAN: a pan-genome analysis pipeline for human genomes. *Genome Biol. 2019 201*, **20**, 1–11.

26. Cubry,P., Tranchant-Dubreuil,C., Thuillet,A.-C., Monat,C., Ndjiondjop,M.-N., Labadie,K., Cruaud,C., Engelen,S., Scarcelli,N., Rhoné,B., *et al.* (2018) The Rise and Fall of African Rice Cultivation Revealed by Analysis of 246 New Genomes. *Curr. Biol.*, 10.1016/j.cub.2018.05.066.

27. Monat,C., Pera,B., Ndjiondjop,M.-N., Sow,M., Tranchant-Dubreuil,C., Bastianelli,L., Ghesquière,A. and Sabot,F. (2016) de novo assemblies of three Oryza glaberrima accessions provide first insights about pan-genome of African rices. *Genome Biol. Evol.*, 10.1093/gbe/evw253.

28. Orjuela,J., Sabot,F., Chéron,S., Vigouroux,Y., Adam,H., Chrestin,H., Sanni,K., Lorieux,M. and Ghesquière,A. (2014) An extensive analysis of the African rice genetic diversity through a global genotyping. *Theor. Appl. Genet.*, **127**, 2211–23.

29. Serret, Julien; Couderc, Marie; Mariac, Cedric; Albar,L.; and Sabot,F. From low cost plant HMW DNA extraction to MinION sequencing.

30. De Coster,W., D'Hert,S., Schultz,D.T., Cruts,M. and Van Broeckhoven,C. (2018) NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*, **34**, 2666–2669.

31. Kolmogorov,M., Yuan,J., Lin,Y. and Pevzner,P.A. (2019) Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol. 2019 375*, **37**, 540–546.

32. Vaser,R., Sović,I., Nagarajan,N. and Šikić,M. (2017) Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.*, **27**, 737–746.

33. Li,H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.

34. Laetsch,D.R. and Blaxter,M.L. (2017) BlobTools: Interrogation of genome assemblies. *F1000Research 2017 61287*, **6**, 1287.

35. Alonge,M., Soyk,S., Ramakrishnan,S., Wang,X., Goodwin,S., Sedlazeck,F.J., Lippman,Z.B. and Schatz,M.C. (2019) RaGOO: Fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol.*, **20**, 1–17.

36. Manni,M., Berkeley,M.R., Seppey,M., Simão,F.A. and Zdobnov,E.M. (2021) BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol. Biol. Evol.*, **38**, 4647–4654.

37. Mikheenko,A., Prjibelski,A., Saveliev,V., Antipov,D. and Gurevich,A. (2018) Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics*, **34**, i142–i150.

38. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–60.

39. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and 1000 Genome Project Data Processing Subgroup,1000 Genome Project Data Processing (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–9.

40. Simpson,J.T., Wong,K., Jackman,S.D., Schein,J.E., Jones,S.J.M. and Birol,I. (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res.*, **19**, 1117–23.

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

42. Marçais,G., Delcher,A.L., Phillippy,A.M., Coston,R., Salzberg,S.L. and Zimin,A. (2018) MUMmer4: A fast and versatile genome alignment system. *PLOS Comput. Biol.*, **14**, 1–14.

43. Smit, A. F. A., R. Hubley, and P. Green. 1999. RepeatMasker. 1999. http://repeatmasker.org.

44. Copetti,D., Zhang,J., El Baidouri,M., Gao,D., Wang,J., Barghini,E., Cossu,R.M., Angelova,A., Maldonado L.,C.E., Roffler,S., *et al.* (2015) RiTE database: A resource database for genus-wide rice genomics and evolutionary biology. *BMC Genomics*, **16**, 1–10.

45. Bao,W., Kojima,K.K. and Kohany,O. (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA*, **6**, 1–6.

46. Shumate,A. and Salzberg,S.L. (2021) Liftoff: accurate mapping of gene annotations. *Bioinformatics*, **37**, 1639–1643.

47. Jones,P., Binns,D., Chang,H.Y., Fraser,M., Li,W., McAnulla,C., McWilliam,H., Maslen,J., Mitchell,A., Nuka,G., *et al.* (2014) InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.

48. Götz,S., García-Gómez,J.M., Terol,J., Williams,T.D., Nagaraj,S.H., Nueda,M.J., Robles,M., Talón,M., Dopazo,J. and Conesa,A. (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.*, **36**, 3420–3435.

49. Matsumoto,T., Wu,J., Kanamori,H., Katayose,Y., Fujisawa,M., Namiki,N., Mizuno,H., Yamamoto,K., Antonio,B.A., Baba,T., *et al.* (2005) The map-based sequence of the rice genome. *Nat. 2005 4367052*, **436**, 793–800.

50. Zhao,Q., Feng,Q., Lu,H., Li,Y., Wang,A., Tian,Q., Zhan,Q., Lu,Y., Zhang,L., Huang,T., *et al.* (2018) Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet.*, 10.1038/s41588-018-0041-z.

51. Freschi,L., Vincent,A.T., Jeukens,J., Emond-Rheault,J.G., Kukavica-Ibrulj,I., Dupont,M.J., Charette,S.J., Boyle,B. and Levesque,R.C. (2019) The Pseudomonas aeruginosa Pan-Genome Provides New Insights on Its Population Structure, Horizontal Gene Transfer, and Pathogenicity. *Genome Biol. Evol.*, **11**, 109–120.

52. Davies,M.R., McIntyre,L., Mutreja,A., Lacey,J.A., Lees,J.A., Towers,R.J., Duchêne,S., Smeesters,P.R., Frost,H.R., Price,D.J., *et al.* (2019) Atlas of group A streptococcal vaccine candidates compiled using large-scale comparative genomics. *Nat. Genet. 2019 516*, **51**, 1035–1043.

53. Vernikos,G.S. (2020) A Review of Pangenome Tools and Recent Studies. *Pangenome Divers. Dyn. Evol. Genomes*, 10.1007/978-3-030-38281-0_4.

54. Bonnici,V., Maresi,E. and Giugno,R. (2021) Challenges in gene-oriented approaches for pangenome content discovery. *Brief. Bioinform.*, **22**.

55. Zhao,Y., Wu,J., Yang,J., Sun,S., Xiao,J. and Yu,J. (2012) PGAP: pan-genomes analysis pipeline. *Bioinformatics*, **28**, 416–418.

TABLE AND FIGURES LEGENDS

Table 1. List of Illumina sequencing data used from 248 African rice accessions. For each dataset, the species, the number of samples (#samples), the number of reads (#reads), the total number of gigabases (#Gb), the average depth sequencing (X) and the reference are provided.

Table 2. Assembly summary. This table provides statistics about the contigs (ctgs) assembled by abyss and the contigs keeped after filtering steps. The statistics include the contigs number (#raw ctgs, #filtered ctgs), the average number of contigs per sample (#raw ctgs per sample, #filtered ctgs per sample), the total length of sequence assembled, the average length of sequence assembled by sample and the average sequence size.

Figure 1. Summary of the approach ”Map-then-assemble” implemented in FrangiPANe. Raw pair-ended short reads are mapped to the reference genome, separately for each sample, and unmapped reads are assembled. Next, contigs from all individuals are pooled and clustered to reduce redundancy. Non-redundant contigs are finally anchored on the genome.

Figure 2. Contigs location on the 12 chromosomes of CG14. A total of 152,411 sequences were uniquely anchored, representing 31.5% of the total number of contigs.

TABLES

Species	#samples	#reads	#Gb	X	Ref
O. glaberrima	162	20,999,547,976	2,074	35	Cubry et al., 2018
O. barthii	84	9,940,764,450	982	28	Cubry et al., 2018
O. glaberrima	1 (CG14)	207,601,236	20.76	60	Monat et al., 2016
O. glaberrima	1 (TOG5681)	289,581,328	24.910	72	Monat et al., 2016

Table 1. List of Illumina sequencing data used from 248 African rice accessions. For each dataset, the species, the number of samples (#samples), the number of reads (#reads), the total number of gigabases (#Gb), the average depth sequencing (X) and the reference are provided.

Species	#raw ctgs	#raw ctgs per sample	#filtered ctgs	#filtered ctgs per sample	Total length (Mb)	Total length per sample	seq size (bp)
O. barthii	5,424,759	64,580	763,176	9,085	740	10.6	1,192
0. glaberrima	4,427,624	27,210	543,500	3,334	917	5.5	1,355
	9,887,127	39,867	1,306,676	5,290	1,657	8	

Table 2. Assembly summary. This table provides statistics about the contigs (ctgs) assembled by abyss and the contigs keeped after filtering steps. The statistics include the contigs number (#raw ctgs, #filtered ctgs), the average number of contigs per sample (#raw ctgs per sample, #filtered ctgs per sample), the total length of sequence assembled, the average length of sequence assembled by sample and the average sequence size.

FIGURES

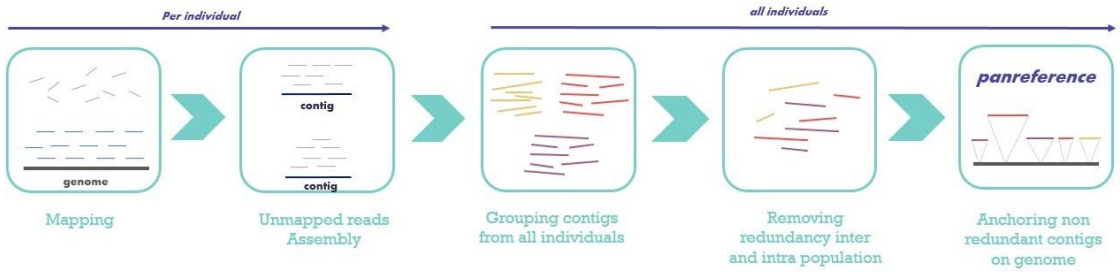

Figure 1. Summary of the approach ”Map-then-assemble” implemented in FrangiPANe. Raw pair-ended short reads are mapped to the reference genome, separately for each sample, and unmapped reads are assembled. Next, contigs from all individuals are pooled and clustered to reduce redundancy. Non-redundant contigs are finally anchored on the genome.

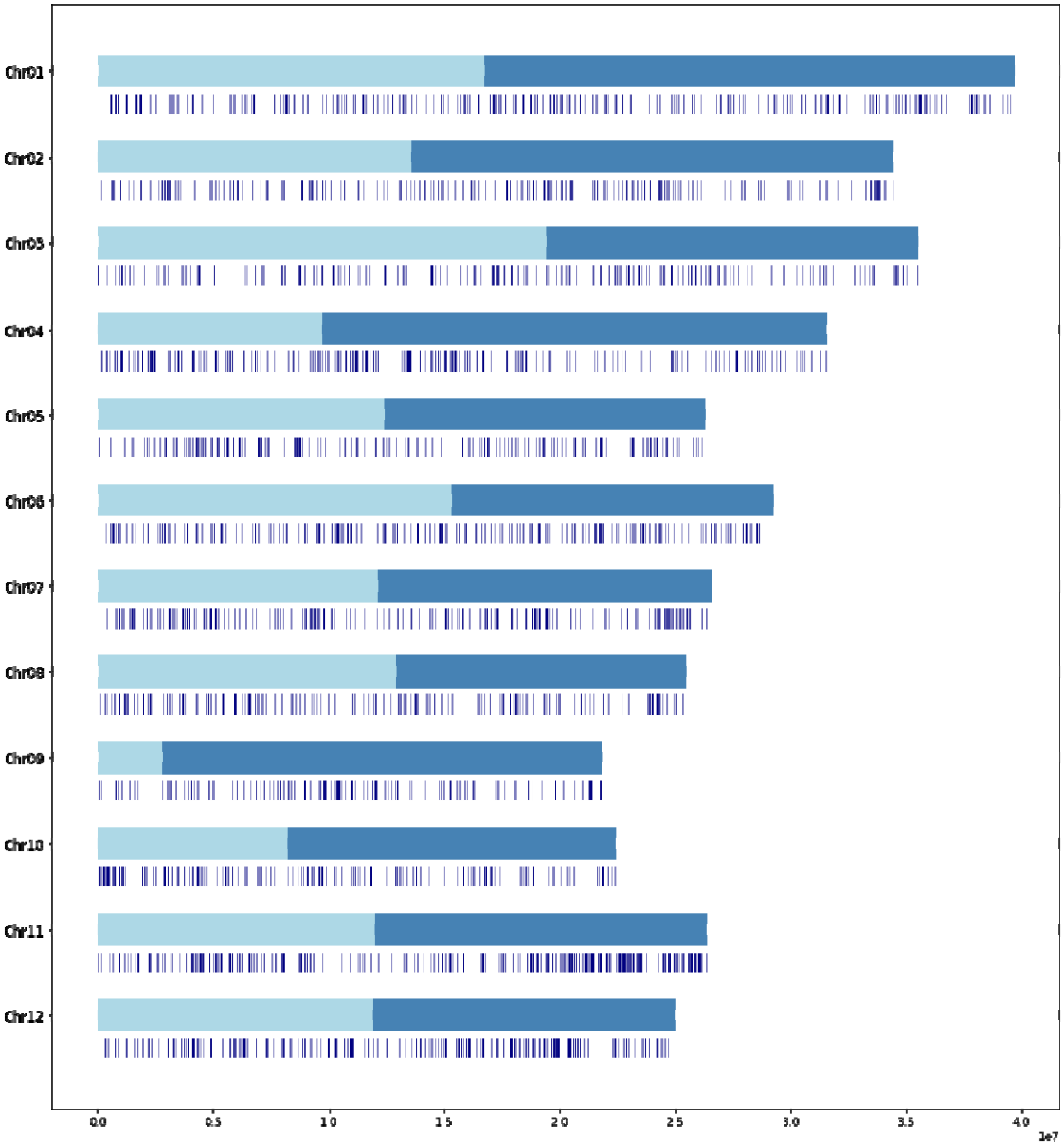

Figure 2. Contigs location on the 12 chromosomes of CG14. A total of 152,411 sequences were uniquely anchored, representing 31.5% of the total number of contigs.