Domestication lead to a drastic reduction of core-genome in two independent lineages of domesticated rices

Cécile Monat[1,2+†], Nguyet Dang[1,2†], Christine Tranchant-Dubreuil[1,2†], Stefan Engelen[3], Karine Labadie[3], Patrick Wincker[3], Ndomassi Tando[1,2], Emmanuel Paradis[4,5], and François Sabot[1,2*]

1 DIADE, Univ of Montpellier, IRD, Montpellier, France
2 South Green Bioinformatics Platform, IRD, BIOVERSITY,CIRAD,INRA, Montpellier, France
3 CEA, Genoscope, Evry, France
4 ISE-M, Univ of Montpellier, CNRS, IRD, EPHE, Montpellier, France

*** Corresponding author: francois.sabot@ird.fr**
+ Present address: Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Seeland, Germany
†Those three authors contribute equally to the work

Abstract

Introduction

Doemstication of rices occured independantly... Diversity loss/founder effect/genetic drift.... However, most previous analyses of domestication impact on diversity were limited to few markers or SNP-based. Recent papers [**?**] shown that the diversity is not only linked to the SNP and short InDels, but also to Presence/Absence Variations (PAV) and Copy Number Variation (CNV). This lead to the idea of Pangenome [Tettelin et al., 2005, **?**], composed of the core genome, genomic sequences shared between all individuals of the considered group, and of the dispensable genome, genomic sequences shared between some individuals (or even only one) but not all. Currently, only a few studies focused on the pangenome of plants, and only one, to our knowledge [**?**]....

In this paper, we benefit of the release of very large scale data on the two domesticated rice species, the Asian rice *Oryza sativa* [**?**] and the African one *O. glaberrima* [Cubry et al., 2018], and on their respective wild relatives, *O. rufipogon* (this paper) and *O. barthii* [Cubry et al., 2018], to explore the impact of domestication on the core genome of cultivated.

Materials and Methods

Plant material

One hundred and sixty-three accessions of *Oryza glaberrima* and 83 accessions of *O. barthii* were obtained from Genoscope and described in [Cubry et al., 2018] (Supplementary Data S1). For Asian rices, we selected 450 samples from the 3,000 Rice Genome Project (3kRGP, [**?**]) based on their location on the global diversity and their high-level of data (min 20x, Supplementary Data S1), and obtained 45 *O. rufipogon*

sequences in Illumina, high-depth from Genoscope through the IRIGIN project (`http://irigin.org`).

DNA Sequencing for *O. rufipogon*

DNA was sequenced at the Genoscope (Evry, France) on Illumina HiSeq 2000/2500/400 with paired-end reads, 100-150 bases long, and cleaned as described in [Cubry et al., 2018, **?**]. Briefly, the mean sequencing depth is of 35x with a maximum of 60x, representing a total of XXXX,XXXX and XXXX reads and YYYYY bases. The data were cleaned as described in [Cubry et al., 2018, **?**].

Short reads mapping

Cleaned reads were then mapped against the Asian reference genome (Os-Nipponbare-Reference-IRGSP-1.0/MSU7.0 [Mcnally et al., 2009, Kawahara et al., 2013]) through a dedicated TOGGLe pipeline [Monat et al., 2015, **?**], as follows: mapping with BWA aln/sampe legacy [Li and Durbin, 2009] (edit distance of 5 bases, insert size of 500 bases), conversion into BAM file and sorting by coordinates with PicardTools (!!REF!!) (for details information about mapping see Supplementary Data S3).

Reads count and normalization

The reads count for each individual on 10kb bin (38,000 for a roughly 380Mb genome) were generated with the multicov tool from BEDTOOLS [Quinlan, 2014]. Counts were normalized to 1 million total reads according to the initial individual sequencing coverage to obtain FP10KM values (Features per 10 kilobase per millions of reads).

Bootstrapping and subsampling

All subsamplings and bootstrapping to were performed from the read count table, with random choice, using truely random selection scripts.

Pnorm and FDR analysis

Pnorm analysis and FDR approach at 5% were performed using R and specific packages pnorm and qvalue [with contributions from Andrew J. Bass et al., 2015] (VERSION). Bins for which the FP10KM mean were lower than 2 were arbitrarily previously ride out the analysis and considered as absent for all individuals. Pan-matrix obtained shown '0' when bin was considered as absent, '1' if the bin was considered as present and 'Uk' (Unknown) if feature does not pass the initial pnorm test.

GO enrichment analyses

Genes underlying the PAV were recovered using the MSUv7 annotation of the IRGSP1.0 assembly (`http://rice.plantbiology.msu.edu/`)). Gene ontology enrichment analyses were performed using a home-made R script and the topGO package [Alexa and Rahnenfuhrer, 2016], with a display of the 5 top most significant GO retrieved using the Fisher classic and weight01 algorithms.

Identification of domesticated related gene families

GenFam was used... Data from gene families were obtained from Green Phyl V.4 [Conte et al., 2008a, Conte et al., 2008b, Rouard et al., 2011], selecting sequences without spliced forms.

Availability of scripts

The whole bash, Perl and R scripts are available on GitHub under Cecill-B/GPLv3 double licenses on the GitHub of the project: `HTTP`

Results

Mapping and Normalization

After cleaning, the raws reads were mapped against the Asian reference genome (Os-Nipponbare-Reference-IRGSP-1.0 [Mcnally et al., 2009, Kawahara et al., 2013]). For African rices, we used 163 accessions of the cultivated African rice *O. glaberrima* and 83 accessions of its wild relative *O. barthii*, with a mean percentage of mapping of 86.49% and 86.44%, respectively [Cubry et al., 2018]. For Asian rice, we used 3,000 library representing 450 samples representing rice diversity for the cultivated, with a mean value of xx% of mapping, and 45 *O. rufipogon* whole genome sequences. From these last samples, 3 of them harboured a very lower level of mapping compared to the other samples (less than 80% compared to a minimum of 85%); after careful check, they were removed as being not *O.rufipogon* but *O. meridionalis* and selected by error, ending up with 42 samples. We then performed reads count normalization according to the initial reads number specific to each individual reported to 1 million and the bin length, obtaining numerical F10KPM data. Distribution of F10PKM density mainly follows a normal law (see `Fig.??` for an example), thus we successively applied pnorm and FDR statisticals test to define if a given bin was present or not for each individual.

Chromosomal, individual and population effects

We checked whether the global FP10KM profile were the same between the different chromosomes, between sequence type (gene, CDS, UTR or TE features) and if there were or not individual(s) effects. We seen that there are no significant differences between the 12 chromosomes on all the four groups analyzed (Supplementary Data S4).We then performed the same analysis only on exons and none exons (Supplementary Data 4), and confirmed that profils are the same no matter the type of features we were looking at. As a proxy for the rest of the chromosome, we choosed the chromosome 10 for the next validations.

In order to test potential individual effect, we performed a hundred bootstrap-like analysis by resampling randomly 76 individuals for each group. No individual effect were detectable but in *O. barthii*, where we found two classes of profils depending on the individuals in the group (Supplementary Data S4). As described in Orjuela et al [Orjuela et al., 2014], this species can be divided in two populations respectively with 23 and 51 individuals distributed. To check if the two profiles were due to population split, we tested the hundred bootstrap analysis on *O. barthii* with only 23 individuals of the first population and 51 of the second (Supplementary Data 4). Two profiles were still detectable on the population 1, so we checked if 23 individuals were enough with 23 individuals of population 2. This time the curves shown more discrepancy suggesting that 23 individuals is not a sufficient number of individuals to see a robust profile. By testing lower subsampling we were able to determine that 40 individuals is a minimal

number to avoid masked population structure effect. In addition, the subsampling analyses were used to confirm that the difference in number of samples between wild and cultivated will not impact the main results (Supplementary Data 4).

Core-genome vs partial dispensable

Identification of domestication-related effects on the pangenome structure

Discussion

Acknowledgments

Authors wants to thanks Bénédicte Rhoné for his R script to analyse the GO.
FINANCIALS!

Authors contributions

FS supervized the whole study; KL, SE and PW performed the sequencing and initial QC analyses; CM, ND, CTD, NT, and FS performed the initial bioinformatics treatments; CM, FS and EP developed and implemented the statistical methods; CM, FS and CTD performed analyses on African rices and ND and FS on Asian rices; FS and CM wrote the manuscript; all authors read and validate the current version of the manuscript.

References

Alexa and Rahnenfuhrer, 2016. Alexa, A. and Rahnenfuhrer, J. (2016). *topGO: Enrichment Analysis for Gene Ontology*. R package version 2.24.0.

Conte et al., 2008a. Conte, M. G., Gaillard, S., Droc, G., and Perin, C. (2008a). Phylogenomics of plant genomes: a methodology for genome-wide searches for orthologs in plants. *BMC genomics*, 9:183.

Conte et al., 2008b. Conte, M. G., Gaillard, S., Lanau, N., Rouard, M., and Périn, C. (2008b). GreenPhylDB: A database for plant comparative genomics. *Nucleic Acids Research*, 36(SUPPL. 1):991–998.

Cubry et al., 2018. Cubry, P., Tranchant-Dubreuil, C., Thuillet, A.-C., Monat, C., Ndjiondjop, M.-N., Labadie, K., Cruaud, C., Engelen, S., Scarcelli, N., Rhoné, B., Burgarella, C., Dupuy, C., Larmande, P., Wincker, P., François, O., Sabot, F., and Vigouroux, Y. (2018). The Rise and Fall of African Rice Cultivation Revealed by Analysis of 246 New Genomes. *Current Biology*.

Kawahara et al., 2013. Kawahara, Y., de la Bastide, M., Hamilton, J. P., Kanamori, H., McCombie, W. R., Ouyang, S., Schwartz, D. C., Tanaka, T., Wu, J., Zhou, S., Childs, K. L., Davidson, R. M., Lin, H., Quesada-Ocampo, L., Vaillancourt, B., Sakai, H., Lee, S. S., Kim, J., Numa, H., Itoh, T., Buell, C. R., and Matsumoto, T. (2013). Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. *Rice (New York, N.Y.)*, 6(1):4.

Li and Durbin, 2009. Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–60.

Mcnally et al., 2009. Mcnally, K. L., Childs, K. L., Bohnert, R., Davidson, R. M., Zhao, K., Ulat, V. J., Zeller, G., Clark, R. M., Hoen, D. R., Bureau, T. E., Stokowski, R., Ballinger, D. G., Frazer, K. A., Cox, D. R., Padhukasahasram, B., Bustamante, C. D., Weigel, D., Mackill, D. J., Buell, C. R., Leung, H., Leach, J. E., Bruskiewich, R. M., and Ra, G. (2009). Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. pages 1–6.

Monat et al., 2015. Monat, C., Tranchant-Dubreuil, C., Kougbeadjo, A., Farcy, C., Ortega-Abboud, E., Amanzougarene, S., Ravel, S., Agbessi, M., Orjuela-Bouniol, J., Summo, M., and Sabot, F. (2015). TOGGLE: toolbox for generic NGS analyses. *BMC Bioinformatics*, 16.

Orjuela et al., 2014. 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. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, 127(10):2211–23.

Quinlan, 2014. Quinlan, A. R. (2014). *BEDTools: The Swiss-Army Tool for Genome Feature Analysis.*, volume 47.

Rouard et al., 2011. Rouard, M., Guignon, V., Aluome, C., Laporte, M. A., Droc, G., Walde, C., Zmasek, C. M., Périn, C., and Conte, M. G. (2011). GreenPhylDB v2.0: Comparative and functional genomics in plants. *Nucleic Acids Research*, 39(SUPPL. 1):1095–1102.

Tettelin et al., 2005. 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., Deboy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., Madupu, R., Brinkac, L. M., Dodson, R. J., Rosovitz, M. J., Sullivan, S. a., Daugherty, S. C., Haft, D. H., Selengut, J., Gwinn, M. L., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'Connor, K. J. B., Smith, S., Utterback, T. R., White, O., Rubens, C. E., Grandi, G., Madoff, L. C., Kasper, D. L., Telford, J. L., Wessels, M. R., Rappuoli, R., and Fraser, C. M. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences of the United States of America*, 102(39):13950–5.

with contributions from Andrew J. Bass et al., 2015. with contributions from Andrew J. Bass, J. D. S., Dabney, A., and Robinson, D. (2015). *qvalue: Q-value estimation for false discovery rate control*. R package version 2.4.2.