Transcriptome population genomics reveals severe bottleneck and domestication cost in the African rice (Oryza glaberrima)

Benoit Nabholz, Gautier Sarah, François Sabot, Manuel Ruiz, Hélène Adam, Sabine Nidelet, Alain Ghesquière, Sylvain S. Santoni, Jacques David, Sylvain Glémin

To cite this version:

HAL Id: hal-01919691
https://hal.archives-ouvertes.fr/hal-01919691
Submitted on 12 Nov 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Transcriptome population genomics reveals severe bottleneck and domestication cost in the African rice (Oryza glaberrima)

BENOIT NABHOLZ,*† GAUTIER SARAH,†‡ FRANÇOIS SABOT,† MANUEL RUIZ,† HÉLÈNE ADAM,† SABINE NIDELET,** ALAIN GHSQUIÈRE,† SYLVAIN SANTONI,§ JACQUES DAVID† and SYLVAIN GLEMIN*

*Institut des Sciences de l’Évolution-Montpellier, UMR CNRS-UM2 5554, University Montpellier II, Montpellier, France, †UMR AGAP 1334, Montpellier SupAgro, Montpellier, France, ‡UMR AGAP 1334, CIRAD, Montpellier, France, §UMR AGAP 1334, INRA, Montpellier, France, **UMR IRD-UM2-CIRAD DIADE, IRD, Montpellier, France, MGX-Montpellier GenomiX, c/o Institut de Génomique Fonctionnelle, Montpellier, France

Abstract

The African cultivated rice (Oryza glaberrima) was domesticated in West Africa 3000 years ago. Although less cultivated than the Asian rice (O. sativa), O. glaberrima landraces often display interesting adaptation to rustic environment (e.g. drought). Here, using RNA-seq technology, we were able to compare more than 12 000 transcripts between 9 O. glaberrima, 10 wild O. barthii and one O. meridionalis individuals. With a synonymous nucleotide diversity \( \pi_s = 0.0006 \) per site, O. glaberrima appears as the least genetically diverse crop grass ever documented. Using approximate Bayesian computation, we estimated that O. glaberrima experienced a severe bottleneck during domestication. This demographic scenario almost fully accounts for the pattern of genetic diversity across O. glaberrima genome as we detected very few outliers regions where positive selection may have further impacted genetic diversity. Moreover, the large excess of derived nonsynonymous substitution that we detected suggests that the O. glaberrima population suffered from the ‘cost of domestication’. In addition, we used this genome-scale data set to demonstrate that (i) O. barthii genetic diversity is positively correlated with recombination rate and negatively with gene density, (ii) expression level is negatively correlated with evolutionary constraint, and (iii) one region on chromosome 5 (position 4–6 Mb) exhibits a clear signature of introgression with a yet unidentified Oryza species. This work represents the first genome-wide survey of the African rice genetic diversity and paves the way for further comparison between the African and the Asian rice, notably regarding the genetics underlying domestication traits.

Keywords: bottleneck, cost of domestication, domestication, nucleotide diversity, Oryza glaberrima, RNA-seq

Received 25 November 2013; accepted 19 March 2014

Introduction

Domestication represents a unique opportunity to study the evolutionary process. In a few thousand years, domesticated populations have experienced major phenotypic changes under adaptation to the new agricultural environments and the pressure of human artificial selection (Diamond 2002; Gepts 2004). Although evolutionary rates under domestication may be slower than previously thought (Purugganan & Fuller 2009), drastic phenotypic changes in such a short timescale are rarely well documented for wild species, especially because the ancestral state is rarely available for comparative study. In this context, a major goal is to identify the genetics underlying the adaptation to domestication (Ross-Ibarra et al. 2007).

Correspondence: Benoit Nabholz, Fax: +33 4 67 14 36 22; E-mail: benoit.nabholz@univ-montp2.fr
One approach to fulfil this goal is to study the pattern of polymorphism across the genome to identify regions exhibiting signatures of adaptation. This so-called bottom-up approach (starting from the gene to the phenotype) assumes, often implicitly, that a low proportion of genes is affected by adaptive evolution. As a result, the average genomic pattern of polymorphism can be used to infer demographic history (Ross-Ibarra et al. 2007). Integrating demographic history is of importance as domestication often implies population bottleneck that affects genome-wide allele frequencies. This is particularly true in crop grasses (Poaceae) where domesticated populations are typically two to three times less variable at neutral markers than their wild relatives (Glémin & Bataillon 2009). In addition to the shuffling of allele frequencies, decrease in population size could lead to the accumulation of slightly deleterious mutations (Ohata 1992). Such an accumulation has been detected by an increase in nonsynonymous (amino acid changing) over synonymous (amino acid conservative) substitutions in dog (Björnerfeldt et al. 2006; Cruz et al. 2008), laboratory yeast (Gu et al. 2005) and yak (Wang et al. 2011) but also potentially in Asian rice (Lu et al. 2006). Beside demography, other genomic features including recombination and recombination rates and gene densities help better understand variable levels. However, these features positively selected sites are known to recombination rates and gene densities help better understand variable levels. However, these features were considered in domestication analyses so far.

of the major crops such as maize (Wright 2002). Its agricultural potential is however enhanced by the recent discovery that O. sativa x O. glaberrima hybrids could lead to high-yielding and stress-resistant varieties, the so-called NERICA (for NEw RICE for Africa). Characterizing the genetic diversity of the African rice could therefore be important in a broader context of conservation of genetic resources. Moreover, the phylogenetic proximity with the Asian rice (Wang et al. 1992; Zhu & Ge 2005) makes it an ideal candidate to test the hypothesis of parallel domestication (Lin et al. 2012). Compared with the Asian rice, the population genetic of the African rice was little studied (Bezaçon 1994). Li et al. (2011) estimated the genetic diversity of the African rice using 14 nuclear loci. This study pointed out a sharp loss of diversity (about 70%) in the domesticated population compared with the wild O. barthii. Semon et al. (2005) used O. sativa microsatellites to study the population structure of the African rice and revealed a weak population structure within O. glaberrima and elevated linkage disequilibrium expanding chromosome wide. The latter result has been interpreted as an effect of demography and population structure rather than of low recombination rate.

In this study, we analyse the transcriptome-wide polymorphism of 9 domesticated and 10 wild African rice individuals using RNA-seq technology (Wang et al. 2009). This technology allows us to characterize the genetic diversity of more than 12,000 transcripts in the African rice genome. Using the well-annotated Asian rice genome (International Rice Genome Sequencing Project 2005; Kawahara et al. 2013) as a reference and O. meridionalis as out-group, we tackle the following questions: (i) How polymorphism varies across the genome and which factors shape these variations? (ii) How strongly domestication has affected the genetic diversity of the domesticated compartment? and (iii) Can we identify potential genomic regions affected by selection during domestication?

Materials and methods

Plant material

Ten Oryza barthii and nine O. glaberrima varieties representing the species-wide distribution of the two species (Table S1, Supporting information) were grown in IRD greenhouse, for 7 weeks in short-day conditions. Five plants per accession were used, to obtain enough RNA from inflorescences. For O. glaberrima, all the plants were grown in greenhouse for two generations, with pollen bags to ensure a pure inbred. All five plants were derived from the same original seed. Panicles were collected 4–15 days after induction, to span the initials steps of development (stage 1 to early 8 as describe in Ikeda et al. 2004). Green leaves from the same plants were also collected.

Preparation of RNA samples

Samples were ground in liquid nitrogen, and total cellular RNA was extracted using a total RNA easy Plant minikit with RLT and RWT buffers (Qiagen, GmbH, Germany) with a DNase treatment (RNase-free DNase, Qiagen). RNA concentrations were first measured using a NanoDrop ND-1000 Spectrophotometer then using
the Quant-iT™ RiboGreen® (Invitrogen, USA) protocol on a Tecan Genius spectrofluorimeter. RNA quality was assessed by running 1 μL of each RNA sample on RNA 6000 Pico chip on a Bioanalyzer 2100 (Atablagilent Technologies, Inc., USA). Samples with an RNA integrity number (RIN) value greater than eight were deemed acceptable according to the Illumina TruSeq RNA protocol. For each genotype, 80% of RNA from the inflorescence and 20% from the leaf were mixed to obtain 2 μg of tissue-bulked RNA.

**Illumina library production**

The TruSeq RNA sample Preparation v2 kit (Illumina Inc., USA) was used according to the manufacturer’s protocol with the following modifications. In brief, poly-A-containing mRNA molecules were purified from 2 μg total RNA using poly-T oligo-attached magnetic beads. The purified mRNA was fragmented by addition of the fragmentation buffer and was heated at 94 °C in a thermocycler for 4 min. The fragmentation time of 4 min was used to yield library fragments of 250–300 bp. First-strand cDNA was synthesized using random primers to eliminate the general bias towards 3’ end of the transcript. Second-strand cDNA synthesis, end repair, A-tailing and adapter ligation were carried out in accordance with the manufacturer supplied protocols. Purified cDNA templates were enriched by 15 cycles of PCR for 10 s at 98 °C, 30 s at 65 °C and 30 s at 72 °C using PE1.0 and PE2.0 primers and with Phusion DNA polymerase (NEB, USA). Each indexed cDNA library was verified and quantified using a DNA 100 Chip on a Bioanalyzer 2100 then equally mixed by ten (from different genotypes). The final library was then quantified by real-time PCR with the KAPA Library Quantification Kit for Illumina Sequencing Platforms (Kapa Biosystems Ltd, SA) adjusted to 10 nM in water and provided to the Montpellier Genomix platform (http://www.mgx.cnrs.fr/) for sequencing.

**Illumina library clustering and sequencing conditions**

Final mixed cDNA library was sequenced using the Illumina mRNA-Seq, paired-end protocol on a HiSeq2000 sequencer, for 2 × 100 cycles. Library was diluted to 2 nM with NaOH and 2.5 μL transferred into 497.5 μL HT1 to give a final concentration of 10 pM. One hundred and twenty microlitres was then transferred into a 200-μL strip tube and placed on ice before loading onto the cBot, mixed library, from 10 individual indexed libraries, being run on a single lane. Flow cell was clustered using TruSeq PE Cluster Kit v3, following the Illumina PE_Amp_Lin_Block_V8.0 recipe. Following the clustering procedure, the flow cell was loaded onto the Illumina HiSeq 2000 instrument following the manufacturer’s instructions. The sequencing chemistry used was v3 (FC-401-3001, TruSeq SBS Kit) with the 2 × 100 cycles, paired-end, indexed protocol. Image analyses and basecalling were performed using the Hiseq CONTROL Software (HCS 1.5.15) and Real-Time Analysis component (RTA 1.13.48). Demultiplexing was performed using casava 1.8.1 (Illumina) to produce paired sequence files containing reads for each sample in Illumina FASTQ format. Raw reads are available at http://arcad-bioinformatics.southgreen.fr/african_rice.

**Mapping & SNP calling**

We used Oryza sativa as reference genome. We downloaded Oryza sativa transcriptome (version MSU6.1) from Ensembl plant (Kersey et al. 2009) using the BioMart Web interface portal (Guberman et al. 2011), http://plants.ensembl.org/biomart/martview/, downloaded in October 2012). O. sativa and glaberrima are very closely related (estimated divergence = 0.3%) therefore preventing a mapping biased towards conserved regions. We chose to download the coding sequence (CDS) plus 300 bp upstream and downstream in order to account for the presence of UTRs. Mapping was performed with the bwa software (Li & Durbin 2009) allowing at most five mismatches between a given read and the reference. Further cleaning involved excluding reads with more than two insertions/deletions (indels) or with indels larger than 5 bp. Pair-end reads mapped on different transcripts were also excluded from further analyses.

Genotyping involved two steps. First, genotypes were called using the method described in Tsagkogeorga et al. (2012). This method estimates the sequencing error rate from the data in a maximum-likelihood framework and computes the posterior probability of genotypes assuming Hardy–Weinberg equilibrium. Here, we included the parameter $F_{IS}$ in the computation of the expected heterozygous frequency to account for the high rate of selfing in Oryza barthii/glaberrima (See equations in Appendix S1, Supporting information). Because genotyping with a $F_{IS}$ of 0.95 or 0.5 led to very similar result, we therefore chose to only report the results obtained with $F_{IS}$ of 0.95. We kept genotypes with posterior probability higher than 0.95, provided that at least 10 reads were available for the considered position and individual. Otherwise, data were considered as missing. Second, dubious SNPs, potentially resulting from hidden paralogy, were cleaned using the paraclean method introduced by Gayral et al. (2013). This method used a likelihood-ratio test (LRT) for each SNP to compare the likelihood of a single locus model with the likelihood of a two-locus model (i.e. assuming that two
distinct genes have reads which were erroneously mapped on a single contig). Here too, the original model was modified to account for selfing by introducing a $F_{IS}$ parameter (See equations in Appendix S1, Supporting information). These methods are implemented in the software reads2SNP (http://kimura.univ-montp2.fr/PopPhyl/resources/tools/). Individuals RC4 and RC6 that issued from the same cultivated accession will be used to estimate the false positive rate. Following this control analysis, these individuals were merged in a consensus sequence as they were found to be genetically identical.

Population genetic analyses

Because of this high selfing rate, we treated the data as haploid by randomly drawing one haploid sequence per individuals. All analyses were restricted to coding sequences (i.e. excluding UTRs). Population genetic statistics including nucleotide diversity (i.e. $\pi$; Tajima 1983 and $\theta_W$ Watterson 1975), Tajima’s D (Tajima 1989) and $F_{ST}$ statistic (defined as $F_{ST} = 1 - \pi_w / \pi_B$; $\pi_w$ is the mean pairwise diversity within each population and $\pi_B$ mean pairwise diversity between populations; Hudson et al. 1992) were computed for each populations. The synonymous and nonsynonymous divergence (dS and dN, respectively) were computed using mean pairwise divergence with O. meridionalis. Because divergence was small (dS < 5%), no correction for multiple substitutions was applied. The computation of all above statistics was performed using a homemade program (available at http://arcad-bioinformatics.southgreen.fr/tools) build with the Bio++ library (Guéguen et al. 2013).

Population structure was inferred using the Bayesian software structure (version 2.3.4, Pritchard et al. 2000). For each number of clusters, from 2 o 5, we ran two MCMC chains of 60 000 steps including 30 000 steps of burn-in. The structure analysis was performed using 5648 SNP obtained by selecting randomly only one SNP per locus. We also reconstructed a phylogenetic tree using the concatenation of all sequences using the BIONJ clustering method (Gascuel 1997) where genetic distances were estimated using a TN93 model (Tamura & Nei 1993). The BIONJ phylogenetic reconstruction was performed using Bio++ library (Guéguen et al. 2013). We performed a rough approximation of linkage disequilibrium using the $r^2$ measure implemented in the ‘LDcorSV’ R package (Mangin et al. 2012).

Expression levels were estimated as the number of reads mapped on a transcript corrected by the transcript length and by the total number of reads sequenced per individual. Following Mortazavi et al. (2008), we computed the number of mapped reads per kilobase of coding sequence per million of mappable reads (RPKM):

$$\text{RPKM} = \frac{10^6 C}{NL},$$

where C is the number (count) of reads mapped on a transcript, N is the total number of mapped reads, and L is the length of the transcript.

Genomic determinants of polymorphism

We tested the effect of four genomic variables on the level of synonymous genetic diversity using multiple linear regressions. The variables are (i) synonymous divergence with O. meridionalis, (ii) recombination rates, (iii) GC content of third-codon position (GC3) and (iv) gene density (computed as the number of coding sites per Mb). For recombination rate, no estimate of recombination rate is currently available for the African rice. As a consequence, we use the same Marey map (i.e. genetic versus physical distance) as in Muyle et al. (2011). This map includes 1202 markers along the O. sativa genome. From this map, we estimated the recombination rate (in cM/Mb) using the slope of a local polynomial regression of degree two as proposed in the r package MAREYMAP (Rezvoj et al. 2007). The degree of smoothing of the polynomial regression (the loess function) was controlled by a span parameter. We tested two values of the span parameter, 0.2 and 0.4, which led to very similar results. Results are presented for a span parameter of 0.2. For GC content, mean GC3 was computed using the complete CDS of O. sativa. Gene density was estimated as the proportion of exonic sequences (including annotated UTRs) in a given genomic region of O. sativa. Exonic positions were downloaded from Ensembl plant (Kersey et al. 2009) using the BioMart Web interface portal (Guberan et al. 2011, downloaded in October 2012). Transcripts described as ‘transposon’ or ‘pseudogene’, potentially free from selective constraint, were excluded from the estimation of gene density.

All these explanatory variables and the synonymous genetic diversity (the response variable) were computed across windows of three different sizes (500 kb, 1 Mb or 2 Mb). For genetic diversity, we used the sum of diversity divided by the total number of genotyped sites (i.e. sites with reads2SNP posterior probability >0.95 and coverage >10x per individual) to avoid potential extreme values taken by small transcripts. We assumed that transcript positions were the same as in O. sativa. All variables were log-transformed, except GC3 that appeared to be almost normally distributed. We added the constant one to recombination rates and synonymous diversity in order to account for zero values. Spatial autocorrelation between windows was tested using
the Moran’s I index implemented in the ‘ape’ R package (Paradis et al. 2004). Normality, homoscedasticity and independence of the residuals were tested using Kolmogorov–Smirnov test (included in ‘stats’ R package), Harrison–McCabe test and Durbin–Watson test both from the ‘intest’ R package (Zeileis & Hothorn 2002). The effect of chromosome was tested using an ANOVA. All statistical analyses were performed using R version 2.15.3 (R Core Team 2013).

Approximate Bayesian Computation and outliers detection

We ran 2 000 000 coalescent simulations using the software ms (Hudson 2002) and following the simple demographic scenario presented in Fig. S1 (Supporting information). This scenario considers an ancestral population that split into a domesticated population and a wild population at time $T_{dom}$. Following this split, the domesticated population experienced a population bottleneck between $T_{dom}$ and $T_{bot}$ generations. Afterwards, the domesticated population was assumed to have the same size as the stable wild population ($N_0$). As it is impossible to estimate separately the strength and the duration of the bottleneck, we fixed $T_{dom}$ at 3000 generations before present (Linares 2002) and $T_{bot}$ at 2000 generations (i.e. leading to a bottleneck duration of 1000 generations). Using approximate Bayesian computation (ABC) method, we estimated $N_0$ (assuming a mutation rate $\mu$ of $10^{-8}$ mutation per site per generation) and a parameter $\alpha = N_{bot}/N_0$ where $N_{bot}$ is population size during the bottleneck. We generated uniform prior distribution for $\mu$ from 0.001 to 0.2, and $N_0$ from 20 000 to 120 000, using a Perl script. We used the number of SNP in both domesticated and wild populations and $F_{ST}$ (Hudson et al. 1992) and Tajima’s D (Tajima 1989) in both domesticated and wild populations as summary statistics. These statistics were computed across 1 Mb windows leading to 370 windows. This was done using a c++ program build with the bio++ library (Guéguen et al. 2013). Posterior distributions of $N_0$ and $\alpha$ parameters were estimated using the neural network method of Blum & François (2010) implemented in the ‘abc’ R package (Csilléry et al. 2012) using log transformation and a tolerance parameter of 0.002. Finally, we assessed the fit of our model by performing a posterior predictive checking (Gelman et al. 2013). To do that, we used the set of 4000 demographic parameters taken from the posterior distributions. Then, we obtained the distributions of the 5 summary statistics by simulating data sets using the 4000 sampled sets of parameters. Comparisons between the simulated and the observed statistics allow to check the fit of our model to the data. The source codes, executables and scripts used to perform the ABC analysis are available at http://arcad-bioinformatics. southgreen.fr/tools.

For the outlier detection analysis, we ran coalescent simulations using the demographic parameters inferred by ABC. 20 000 simulations were run for each of the 370 1 Mb window individually (taken into account the number of genotyped sites per Mb). To detect outlier windows, we computed $\Delta \pi$ (define as $\pi_s$ O. glaberrima/ $\pi_s$ O. barthii) and $F_{ST}$ statistics. These statistics have been proven powerful to detect selective sweep (Innan & Kim 2008). $P$-value of the null model (bottleneck without selection) was computed as the proportion of coalescent simulations producing $\Delta \pi$ and $F_{ST}$, respectively, below and above the observed values.

Finally, we also used GO-slim terms to sort the genes according to the ‘Biological process’ category. GO-slim terms were downloaded at the TIGR rice genome annotation resource (downloaded in April 2012, version 6.0, Ouyang et al. 2007). We compute the mean $\Delta \pi_s$ (defined as $\pi_s$ O. glaberrima/ $\pi_s$ O. barthii) and the mean $F_{ST}$ of the genes belonging to each of the 45 generic GO-slim terms (category ‘Biological process’). Then, we ask whether $\Delta \pi_s$ was lower or $F_{ST}$ higher in each category than in a random, similar-sized, subsets of genes within the genome (randomization test, $N = 10 000$ replicates). We applied the false discovery rate (FDR) method of Benjamini and Hochberg (1995) to adjust the $P$-value for multiple testing.

Cost of domestication

Using the two out-groups O. meridionalis and O. sativa to orientate SNPs, we counted the number of derived synonymous (Fs) and nonsynonymous (Fn) mutations which are fixed in O. glaberrima (resp. O. barthii) but polymorphic or fixed for the ancestral allele in O. barthii (resp. O. glaberrima). They correspond to fixed derived SNPs specific to one of the two populations. We performed this count on the whole data set, for half of the data set split according to the median of gene expression levels (lowly vs highly expressed genes), and for 2 Mb windows across genome to get enough counts per window. The contingency tables we obtained were tested by a Fisher’s exact test using R (R Core Team 2013). Counts were normalized by the number of synonymous (Ls) and nonsynonymous (Ln) positions to get ratios equivalent to the classical Dn/Ds ratios: Dn/Ds = Fn/Ln/Fs/Ls. Dn/Ds ratios were computed for windows with at least four fixed synonymous mutation, to limit very noisy estimates. At the 2 Mb window scale, we correlated Dn/Ds in the cultivated population with $\Delta \pi_s$ (defined as $\pi_s$ O. glaberrima/ $\pi_s$O. barthii) ratios. For this correlation, we excluded two windows over 131 with $\Delta \pi_s > 1$, but results remain unchanged if these two windows are included.
Results

Mapping, genotyping and paralogy detection

The transcriptomes of 10 O. barthii, 9 O. glaberrima (plus one technical replicate) and one O. meridionalis individuals were sequenced using RNA-seq Illumina technology. These individuals were chosen to represent the diversity of the cultivated (O. glaberrima) and wild populations (O. barthii) of the African rice (see Materials and Methods). O. meridionalis was used as an out-group in addition to O. sativa. The sequencing leads to a total of 16 to 46 millions of 75 and 100 bp reads per individuals (Table S2, Supporting information). After the cleaning process, between 55% and 73% of reads were successfully mapped to the O. sativa reference transcripts (see Methods), representing 10 to 29 millions reads per individuals (Table S2, Supporting information).

We performed genotyping and single nucleotide polymorphism (SNP) calling using an extended version of the reads2snp software, which includes the Wright’s fixation index, $F_{IS}$, in the genotype calling and paralog filtering procedure (paraclean) (Gayral et al. 2013) (see Supporting information). This allowed to take selfing rate into account. We set $F_{IS}$ to 0.95 but a lower value of $F_{IS}$ of 0.5 led to very similar results (see below). To be called, a genotype must have a posterior probability above 0.95 and a minimal coverage of 10× per individual. When using the paraclean procedure, of 38 620 SNPs called in O. barthii, 15 042 (39%) rejected the one-locus model (Likelihood-ratio test (LRT), $P < 0.05$). For O. glaberrima, 593 of 8193 SNPs called (7%) rejected the one-locus model (LRT, $P < 0.05$). All these sites were excluded from further analyses. Overall, 5685 transcripts contained at least one SNP that rejected the one-locus model in the process of paralogy detection. Two individuals, RC4 and RC6 issued from the same cultivated accession (i.e. expected to be genetically identical), were used to evaluate the accuracy of the cleaning procedure. Without paraclean, we detected 117 SNPs between RC4 and RC6, 4146 genes with an excess of heterozygotes (negative $F_{IS}$) and 232 SNPs with a premature stop codon. After using paraclean, we recorded only 17 SNPs between RC4 and RC6 (~0.001 per kb), representing less than one per cent of the expected diversity in O. glaberrima (~0.25 per kb, Table 1). Furthermore, we detected only 531 genes exhibiting an excess of heterozygotes and only 81 SNPs with a premature stop codon. Other genotype calling results with and without paraclean are provided in Table S2 (Supporting information). These results highlight the crucial need of removing paralogous sequences and inaccurate mapping, especially in highly selfing species with low diversity and heterozygosity.

Despite the cleaning procedure, three individuals (RS1, RC8 and RC9) appeared as clear outliers with more than 1500 heterozygous positions compared with an average of ~30 positions for other individuals (Fig. S2, Supporting information). Although we cannot exclude contamination during RNA extraction or library preparation, these individuals do not seem to correspond directly to a mix with any other individuals of the sample. This excess of heterozygous positions could be due to a recent outcrossing event that occurred in the wild or during seed production in the genetic resource centre. We thus chose to discard these individuals and reperformed the SNPs calling/paralogy filtering. However, results were mostly unchanged after removal (results not shown).

For the final data set, we only considered sites with at least 7 individuals per population, representing a maximum of one missing individual for O. glaberrima and two missing individuals for O. barthii. Transcripts with less than 30 bp were excluded. We obtained 12 169 transcripts for a total of 11 987 421 aligned bp. This represent 29% of the 41 678 annotated CDS in the O. sativa genome (version MSU6.1, excluding genes annotated as ‘transposon’ and mitochondrial proteins). Our data set contain proportionally less transcripts described as ‘transposon’ or ‘retrotransposon’ (according to Ensembl gene description) than in the complete genome (1.6%, 194 out of 12 169, in the present data set vs. 29%, 16 564 out of 58 058, in the complete genome). This result indicates that our data set is relatively free of transposable elements including young paralogous sequences where polymorphism and divergence are probably difficult to disentangle.

In the final data set, we estimated a $F_{IS}$ of 0.87 and 0.95 for O. barthii and O. glaberrima populations, respectively, values which are, respectively, lower and equal to the parameter used in reads2snp (see above). Nevertheless, using a lower $F_{IS}$ parameter in reads2snp (e.g. 0.5) only weakly affects these estimates ($F_{IS} = 0.85$ and

<table>
<thead>
<tr>
<th>Population</th>
<th>No. of genes</th>
<th>Size (Mb)</th>
<th>No. of SNPs</th>
<th>$\pi$ total per kb</th>
<th>$\pi_s$ per kb</th>
<th>$\pi_s/\pi_a$</th>
<th>Tajima’s D</th>
</tr>
</thead>
<tbody>
<tr>
<td>O. barthii</td>
<td>12 169</td>
<td>11.987</td>
<td>23 578</td>
<td>0.684</td>
<td>1.396</td>
<td>0.284</td>
<td>−0.221</td>
</tr>
<tr>
<td>O. glaberrima</td>
<td>7597</td>
<td></td>
<td>593</td>
<td>0.255</td>
<td>0.557</td>
<td>0.270</td>
<td>−0.044</td>
</tr>
</tbody>
</table>
0.94 for O. barthii and O. glaberrima populations, respectively). It also does not change any of the global statistics (Table S3, Supporting information). Therefore, we hereafter consider only the results obtained with the parameter $F_{IS}$ set to 0.95. Because of this high rate of inbreeding, we treated individuals as haploid by randomly selecting one haploid sequence per individual. It should be noted that these $F_{IS}$ could not correspond to the $F_{IS}$ of natural populations as our individuals were isolated for several generations in greenhouse.

Global patterns of genetic diversity

Basic population genetic statistics are summarized in Table 1. Genetic diversity is low in the wild O. barthii population (nucleotide diversity $\pi = 0.68$ per kb) and is drastically reduced in the domesticated O. glaberrima population ($\pi = 0.25$ per kb, Table 1). Our estimates are comparable with the results of Li et al. (2011) for the domesticated population ($\pi_{silent} = 0.61$ per kb in Li et al. (2011) vs. $\pi_{synonymous} = 0.57$ per kb in the present study), whereas for the wild population, Li et al. (2011) obtained a seemingly higher genetic diversity ($\pi_{silent} = 2.50$ per kb), but not statistically different from our estimate ($\pi = 1.40$ per kb, t-test, $P = 0.11$). This difference appears to be mainly driven by two loci with high diversity (Adh1 and CatA) in Li et al. (2011). This estimation appears robust to individuals sampling as we obtain the same value excluding randomly two individuals in each population from our data sets (results not shown).

Table 2 Genetic diversity in some wild and cultivated crops. $\pi_s$ and $\pi_{total}$ are nucleotide diversity of synonymous (or silent) and all the coding sites, respectively. Ratio cultivated/wild is the ratio of $\pi_s$ when available and $\pi_{total}$ otherwise

<table>
<thead>
<tr>
<th>Population</th>
<th>No. of genes</th>
<th>No. of sites</th>
<th>$\pi_s$ Wild</th>
<th>$\pi_s$ Cultivated</th>
<th>$\pi_{total}$ Wild</th>
<th>$\pi_{total}$ Cultivated</th>
<th>Ratio Cultivated/Wild</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oryza barthii/O. glaberrima</td>
<td>12 169</td>
<td>987 421</td>
<td>1.40</td>
<td>0.56</td>
<td>0.59</td>
<td>0.69</td>
<td>0.26</td>
<td>0.40</td>
</tr>
<tr>
<td>O. barthii/O. glaberrima</td>
<td>14</td>
<td>11 000</td>
<td>2.50</td>
<td>0.61</td>
<td></td>
<td></td>
<td></td>
<td>0.24</td>
</tr>
<tr>
<td>Triticum turgidum dicoccoides/T. t. dicoccum</td>
<td>21</td>
<td>21 720</td>
<td>3.60</td>
<td>1.20</td>
<td>2.70</td>
<td>0.80</td>
<td></td>
<td>0.33</td>
</tr>
<tr>
<td>O. rufipogon/O. s. japonica + O. s. indica combined</td>
<td>111</td>
<td>54 541</td>
<td>5.19</td>
<td>3.20</td>
<td>3.57</td>
<td>2.29</td>
<td></td>
<td>0.61</td>
</tr>
<tr>
<td>O. rufipogon/O. s. japonica + O. s. indica combined</td>
<td>G.W</td>
<td></td>
<td>7.2</td>
<td>5.4</td>
<td></td>
<td></td>
<td></td>
<td>0.75</td>
</tr>
<tr>
<td>Zea mays mays/Z. m. parviglumis</td>
<td>774</td>
<td>230 638</td>
<td>9.74</td>
<td>6.51</td>
<td></td>
<td></td>
<td></td>
<td>0.67</td>
</tr>
<tr>
<td>Z. m. mays/Z. m. parviglumis</td>
<td>G.W</td>
<td></td>
<td>7.4</td>
<td>5.9</td>
<td>4.9</td>
<td></td>
<td></td>
<td>0.83</td>
</tr>
<tr>
<td>Z. m. mays/Z. m. parviglumis</td>
<td>12</td>
<td>11 301</td>
<td>21.1</td>
<td>13.1</td>
<td></td>
<td></td>
<td></td>
<td>0.62</td>
</tr>
<tr>
<td>Pennisetum glaucum glaucum/P. g. monodii</td>
<td>20</td>
<td>9649</td>
<td>7.44</td>
<td>3.13</td>
<td>6.04</td>
<td>4.11</td>
<td></td>
<td>0.42</td>
</tr>
</tbody>
</table>

*Study using next-generation sequencing technology. G.W, Genome-Wide estimate.

With synonymous polymorphism levels approximately three times lower than in human (Bustamante et al. 2005), O. barthii is the less polymorphic wild progenitor ever recorded in crop grasses (Table 2). It is, for example, 15 times less diverse than teosinte (Zea mays; Wright et al. 2005) and nearly four times less diverse than the wild Asian rice (Oryza rufipogon; Caicedo et al. 2007).

Population structure

In the STRUCTURE analysis, we found a likelihood plateau at $K = 3$ populations when O. sativa individual is included (Fig. S3, Supporting information). The domesticated population (O. glaberrima) form a homogeneous population, whereas the wild O. barthii appears less homogeneous with RS8 and, to a lower extent, RS2 possibly sharing common ancestry with O. sativa (Fig. 1A). No wild individual appears closer to the domesticated population, and no clear population structure could be identified among wild individuals. This result is confirmed by neighbour-joining phylogenetic tree inferred from the concatenation of all transcripts (Fig. 1B).

Introgression

A visual comparison of the genetic diversity between O. barthii and O. glaberrima populations revealed a region of chromosome 5 with an exceptionally high genetic diversity ($\pi_{glaberrima} = 2.59$ per kb vs $\pi_{barthii} = 1.03$, Fig. 2). Intriguingly, close inspection of these
genomic regions revealed that most of the variation comes from one cultivated individual (RC3). Excluding RC3 leads to a severe reduction in *O. glaberrima*'s genetic diversity is this region (from $\pi = 8.01$ per kb with RC3 to $\pi = 0.26$ per kb without RC3). A phylogeny reconstructed using the BIONJ clustering method and TN93 substitution model. RS represent *O. barthii* individuals and RC, *O. glaberrima*. The branch leading to *O. sativa* is artificially reduced for clarity.

**Fig. 2** Nucleotide diversity computed over window of 1 Mb along the 12 chromosomes in wild (*O. barthii*) and domesticated (*O. glaberrima*) African rice populations. Grey and black lines indicate lowess fits of *O. glaberrima* and *O. barthii* data, respectively, compute using the $\text{loess}$ function. Vertical dotted lines indicate centromere positions obtained from http://rice.plantbiology.msu.edu/annotation_pseudo_centromeres.shtml. Y-axis is in logarithmic scale of base 10.

is most likely explained by an introgression from another *Oryza* species or possibly with another subspecies. More (sub-)species should be included in phylogenetic analyses to clarify this point. Here, we only excluded that *O. s. ssp indica* was the subspecies involved in the introgression (See Supporting information and Fig. S4). Polymorphism levels of *O. glaberrima* also exceed polymorphism levels of *O. barthii* on 4 additional regions (chromosome 4, 10, 11 and 12 positions 13.5, 0.5, 22.5 and 15–17 Mb, see section ‘Possible genetic targets of the domestication process’). Here,
however, phylogenetic analyses did not reveal any particular pattern in the gene genealogies.

**Determinant of gene diversity across the African rice genome**

To study variations of polymorphism across the Asian rice genome, we chose to average genetic diversity over relatively large windows (500 kb to 2 Mb), assuming the same gene positions as in the Asian rice genome. This approach reduces the background noise created by individual gene variations and is also justified by the extensive linkage-disequilibrium (LD) detected in our data. A visual inspection of the decay of LD estimates using correlation between genotype indicates that LD extends up to 1 Mb in both *O. barthii* and *O. glaberrima* (Fig. S5, Supporting information).

When averaged over 1 Mb windows, synonymous genetic diversity varies extensively from 0.13 to 3.56 per kb in the wild *O. barthii* (excluding windows with less than 1 kb of coding sequence, Fig. 2). Overall, levels of local polymorphism correlate well between *O. barthii* and *O. glaberrima* populations ($R^2 = 0.49$, $P < 10^{-16}$). Effects of natural selection through genetic linkage are frequently identified as an important factor affecting neutral genetic diversity (see Cutter & Payseur 2013 for a review). Natural selection is expected to reduce polymorphism levels at linked sites, in the case of selective sweeps of beneficial mutations (Smith & Haigh 1974) or purifying selection against deleterious mutations (i.e. background selection, Charlesworth et al. 1993). In both cases, a positive correlation with recombination rates (Begun & Aquadro 1992) and a negative correlation with coding-site density (Payseur & Nachman 2002) are expected. Additionally, local variations in mutation rates can also impact the level of neutral genetic diversity (Hellmann et al. 2005).

To test for the effect of linked selection on genetic diversity in African rice, we correlated synonymous polymorphism levels with recombination rates and gene density using the *O. sativa ssp japonica* cv NipponBare genome as a reference (see Method for details). We also tested the influence of two additional explanatory variables, namely synonymous divergence with *O. meridionalis* (dS) and GC content (computed on third-codon positions, GC3). Synonymous divergence serves as a proxy of mutation rates, whereas GC content was shown to influence genetic diversity in mammals and birds (Hellmann et al. 2005; Mugal et al. 2013a). Genome-wide estimates of these variables and of synonymous genetic diversity were computed over nonoverlapping windows of three different sizes: 500 kb, 1 Mb or 2 Mb. The average number of genes per window was of 16, 32 and 64 for the 500 kb, 1 Mb and 2 Mb windows, respectively. Spatial autocorrelation was generally high for all the explanatory variables and especially so for recombination rates and gene density (mean Morans’ I index per chromosome of 0.31 for recombination rates, 0.39 for gene density, 0.10 for synonymous divergence and 0.04 for GC3, estimated using 1 Mb window). To limit the impact of autocorrelation, we selected only odd windows for multiple regression analyses (therefore, excluding one of two windows, Morans’ I index dropped to 0.14, 0.25, 0.04 and 0.02, respectively). The results obtained using even windows were similar to those obtained using odd windows. We thus present only results using odd windows, except when stated otherwise.

Among the four explanatory variables, recombination rates and coding sequence density have the strongest influence on synonymous nucleotide diversity (Table 3, Fig. 4). The effect of recombination is always positive and that of gene density is always negative, regardless of window size or population (cultivated and wild, Table 3). These effects are, however, statistically significant in only four of six combinations (three sizes of windows and two species) for recombination and in three of six combinations for gene density (Table 3). These results are in agreement with a predominant effect of selection reducing genetic diversity in lowly recombining regions and/or genomic regions where coding sites (i.e. potential targets to selection) are abundant. Correlations appear less pronounced in the
domesticated populations, indicating that potential adaptive events during domestication do not strengthen the relationship between genetic diversity and recombination rates/gene density. However, it should be noted that correlations are expected to be weaker in the domesticated population because of the higher variance in polymorphism created by the population bottleneck during domestication and because the selection efficiency against deleterious mutation is expected to be weaker (see below).

Additionally, GC3 has a weak negative effect on nucleotide diversity. This effect is never significant (Table 3). Finally, synonymous divergence has a weak positive effect that decreases in intensity from 500 kb to 2 Mb window size. This result suggests that local variations in mutation rates do exist across the rice genome but likely at a small scale only (significant effect only for windows <1 Mb). Similarly to GC content, synonymous divergence appears to be strongly correlated with recombination rate (Spearman’s $p = 0.35$, $P < 0.001$ at 1 Mb windows). This correlation could be a consequence of a mutagenic effect of recombination (Hellmann et al. 2005; Flowers et al. 2012) maybe linked to open chromatin (Thurman et al. 2012) but could also be the consequence of ancestral polymorphism (Li 1977) that could substantially contribute to the divergence between *O. barthii* and *O. meridionalis* as they appear closely related (mean synonymous divergence: $dS = 2.7\%$).

Finally, we also tested for a potential chromosome-wide effect on *O. barthii* synonymous polymorphism. All chromosomes appear statistically similar in terms of neutral genetic diversity except chromosome 11 which exhibits a higher diversity (Tukey’s HSD test, $P < 0.05$ for all the pairwise comparison including the chromosome 11). The diversity of chromosome 11 is particularly high close to the end between position 15 and 25 Mb (Fig. 2). This strong genetic diversity is not explained by any of the four genomic features tested above ($dS$ with *O. meridionalis*, recombination rates, gene density and GC content). In addition, population genetic statistics, such as Tajima’s $D$, do not indicate that balancing selection is acting in that particular region. Excluding chromosome 11 from multiple regression analyses shows very similar results (Table S4, Supporting information).

Gene expression and nonsynonymous variations

Recently, expression level was proposed as an important determinant of protein evolutionary constraint (Drummond et al. 2005; Drummond & Wilke 2008; Gout et al. 2010). We tested for the effect of gene expression, measured as the number of mapped reads standardized by transcript length and sequencing effort (RPKM), on the ratio of nonsynonymous ($p_n$) over synonymous ($p_s$) nucleotide diversity in *O. barthii*. We divided genes into 20 groups of identical size based on their gene expression (RPKM). We found a strong negative correlation

![Graph](https://example.com/graph.png)  
**Fig. 4** Relationship between *O. barthii* synonymous diversity and recombination rate divided by the density of coding sequenced computed over 2 Mb windows. Black line is the linear regression line.
between expression level and $\pi_n/\pi_s$ (Spearman’s $r = -0.55 \ P = 0.002$, Fig. 5). This result agrees with previous studies linking protein expression and evolutionary constraint. However, it appears that expression level is negatively correlated to $\pi_s$ but also positively related to $\pi_n$, indicating that the relationship is not fully explained by a direct link between transcription level and evolutionary constraint. It should be noted that the positive correlation between $\pi_n$ and expression level is not explained by recombination and that expression level is not correlated with $\pi_s$ if included as explanatory variable in the multiple correlation analyses performed above (result not show).

**Estimation of the domestication bottleneck intensity by Approximate Bayesian Computation**

The first objective of this analysis was to estimate the intensity of the bottleneck experienced by the African rice during domestication, primarily to make it comparable with others plant species. The second was to obtain a demographic scenario serving as a neutral reference to detect outliers regions potentially under selection (Wright et al. 2005; Innan & Kim 2008). We used a very simple demographic model (Fig. S1, Supporting information) represented by an ancestral population that split into a domesticated population and a wild population at time $T_{dom}$. Whereas the wild population was assumed to be stable, the domesticated population experienced a bottleneck from $T_{dom}$ to $T_{bot}$ generation.

Between $T_{bot}$ and time $t = 0$, we assumed that the population size of the domesticated population equaled the wild population. This model has only two free parameters: namely, the wild $θ$ (which is the product of ancestral population size, $N_0$, and mutation rate, $μ$) and the bottleneck intensity (which is the ratio of the population size during domestication, $N_{bot}$, and the duration of the domestication, that is, $D_{dom} = T_{bot} - T_{dom}$). In the simulations, we did not estimate directly $N_{bot}$ but instead we assumed a parameter $\alpha = N_{bot}/N_0$. We fixed $T_{dom}$ at 3000 generations before present (Linares 2002) and $T_{bot}$ at 2000 generations ($D_{dom} = 1000$ generations). We also assumed a mutation rate of $10^{-8}$ per year and per site (Caicedo et al. 2007). We estimated $α$ and $N_0$ using an approximate Bayesian computation (ABC) method (Bertorelle et al. 2010; Csilléry et al. 2012). We run 2 millions coalescent simulations using uniform prior distribution for $α$ and $N_0$ and estimated posterior distributions using a neural networks method (Blum & François 2010). We obtained estimates of $N_0$ and $α$ from the mean and the credibility interval (CI) of the posterior distributions with $N_0 = 71\ 600$ (95% CI = 67 400–75 600) and $α = 0.017$ (95% CI = 0.015–0.020), which corresponds to $N_{bot}$ $≈$ 1200 during 1000 generations. The extreme bottleneck experienced by *O. glaberrima* during domestication corresponds to a reduction in approximately 98% of the effective population size during 1000 generations.

Posterior predictive checking indicates that our simple demographic model has a moderate fit to the data (Fig. S6, Supporting information). First and not surprisingly, the model did not fit the negative Tajima’s $D$ observed in the wild population (mean Tajima’s $D = −0.21$). Only 0.5% of the posterior predictive simulations have a Tajima’s $D$ lower than the observed value. Excluding Tajima’s $D$ from the summary statistics, however, does not change the estimated parameters ($N_0 = 70\ 800$, 95% CI = 67 100–74 900 and $α = 0.017$, 95% CI = 0.015–0.020). Additionally, the number of SNP in the domesticated population and $F_{ST}$ are both not very well recovered by the posterior predictive simulations [in both cases, the observed statistics were among the 3% most extreme values, Fig. S6 (Supporting information)]. Excluding the number of SNP in the domesticated population from the summary statistics does not affect the estimation of $N_0$ but leads to a slightly higher $α$ ($α = 0.020$, 95% CI = 0.018–0.024) and, in contrast, excluding $F_{ST}$ leads to a slightly lower $α$ ($α = 0.013$, 95% CI = 0.011–0.016). We choose to take into account this variation in the next section.

**Possible genetic targets of the domestication process**

Following the recommendations of Innan & Kim (2008), we used the statistics $Δπ$ ($Δπ = π_{O. glaberrima}/π_{O. barthii}$).
O. barthii) and \( F_{ST} \) (Hudson et al. 1992) computed for 370 1 Mb windows. For each window, a neutral distribution of these statistics was computed from coalescent simulations using the demographic parameters inferred by ABC (using one value for \( N_0 \) but three values for \( \alpha = 0.013, 0.017 \) and 0.020, see above). A P-value for selection was computed as the proportion of coalescent simulations producing a \( \Delta \pi \) and a \( F_{ST} \), respectively, below and above the observed values.

Using \( \alpha = 0.017 \), only two windows of 370 have a \( \Delta \pi \) significant at a 5% threshold (Table 4), one on chromosome 6, position 7.5 Mb and one on chromosome 10, position 13.5 Mb. Using \( F_{ST} \), three windows were detected as outliers at a 5% threshold. All are on chromosome 6 between positions 14 and 17 Mb, and one of these windows (position 16.5 Mb) is also marginally significant for the \( \Delta \pi \) statistics (Table 4, P = 0.051). Overall, the number of windows detected as outliers is very low, lower than expected solely by chance (i.e. 18.5 windows are expected at the 5% level) although this expectation relies on the assumptions that windows are independent which is probably not the case as suggested by the high levels of LD (Fig. S6, Supporting information) and by the spatial autocorrelation previously detected (see section ‘Recombination and gene density affect synonymous diversity’). Looking for outliers in the upper value of the \( \Delta \pi \) statistic (i.e. excess of diversity in O. glaberrima compared with O. barthii) allowed us to identify the possibly introgressed window detected earlier (chromosome 5, positions 4-5 Mb, \( P < 0.01 \)). We also identified five other windows on chromosomes 4, 10, 11 and 12 positions 13.5, 0.5, 22.5 and 15-17 Mb, respectively. These windows represent good candidates for introgressed genomic regions although phylogenetic analysis did not reveal any clear pattern as opposed to the region on chromosome 5, positions 4-5 Mb (Fig. 3). Using \( \alpha = 0.013 \) has only the effect to move the P-value slightly higher (for example, the P-value of the most extreme windows in terms of \( \Delta \pi \) move from 0.024 to 0.039). Similarly, using \( \alpha = 0.013 \) only slightly reduces P-values (the P-value of the most extreme windows in terms of \( \Delta \pi \) move from 0.024 to 0.017).

In a second approach, we categorized the genes according to the 45 GO-slims terms of ‘Biological process’ category. For each term, we tested whether the mean \( \Delta \pi \) was lower or the mean \( F_{ST} \) higher than a random, similar-sized subsets of genes (randomization test). Interestingly, the three categories with the most significant \( F_{ST} \) were related to development (Table S5, Supporting information): namely ‘anatomical structure morphogenesis’ (\( N = 396, F_{ST} = 0.28, P\)-value = 0.004), ‘post-embryonic development’ (\( N = 625, F_{ST} = 0.26, P\)-value = 0.006) and ‘cell differentiation’ (\( N = 251, F_{ST} = 0.28, P\)-value = 0.008). However, none of these GO categories had a significant P-value after correction for multiple testing (FDR adjusted P-value > 0.10).

Finally, we followed the method of Wright et al. (2005) to estimate the proportion of windows potentially affected by selection. This method assumes a proportion \( f \) of windows that underwent a severe bottleneck (mimicking positive selection), whereas the remaining 1–\( f \) windows underwent the domestication bottleneck only. As in Wright et al. (2005), the severe bottleneck is chosen to be 10 times stronger that the genome-wide estimate obtained by ABC. The proportion \( f \) is the estimated by a likelihood approach (see Method section and Wright et al. 2005). In contrast to Wright et al. (2005), we found a maximum-likelihood estimation of \( f = 0 \) (Fig. S7, Supporting information), demonstrating that a model with a single bottleneck adequately fits the domestication history of the whole genome of O. glaberrima: variations in the loss of polymorphism we observed across the genome (Fig. 2).

---

**Table 4 The five most extremes outlier windows using \( \Delta \pi, F_{ST} \) statistics and parameter \( \alpha = 0.017 \)**

<table>
<thead>
<tr>
<th>Statistic</th>
<th>No. of genes</th>
<th>Size (kb)</th>
<th>No. of SNPs Wild</th>
<th>No. of SNPs Cultivated</th>
<th>( \pi ) Wild</th>
<th>( \pi ) Cultivated</th>
<th>( F_{ST} )</th>
<th>Chromosome</th>
<th>Location (Mb)</th>
<th>( P^* )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Delta \pi )</td>
<td>28</td>
<td>27.2</td>
<td>61</td>
<td>0</td>
<td>0.805</td>
<td>0</td>
<td>0.192</td>
<td>6</td>
<td>7.5</td>
<td>0.024</td>
</tr>
<tr>
<td></td>
<td>27</td>
<td>20</td>
<td>31</td>
<td>0</td>
<td>0.462</td>
<td>0</td>
<td>0.089</td>
<td>10</td>
<td>13.5</td>
<td>0.043</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>18.4</td>
<td>33</td>
<td>0</td>
<td>0.645</td>
<td>0</td>
<td>0.652</td>
<td>6</td>
<td>16.5</td>
<td>0.051</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>14.1</td>
<td>25</td>
<td>0</td>
<td>0.686</td>
<td>0</td>
<td>0.601</td>
<td>12</td>
<td>5.5</td>
<td>0.077</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>12.8</td>
<td>17</td>
<td>0</td>
<td>0.476</td>
<td>0</td>
<td>0.134</td>
<td>7</td>
<td>7.5</td>
<td>0.086</td>
</tr>
<tr>
<td>( F_{ST} )</td>
<td>6</td>
<td>5</td>
<td>11</td>
<td>1</td>
<td>0.821</td>
<td>0.057</td>
<td>0.727</td>
<td>6</td>
<td>14.5</td>
<td>0.023</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>13.5</td>
<td>39</td>
<td>1</td>
<td>1.169</td>
<td>0.021</td>
<td>0.679</td>
<td>6</td>
<td>15.5</td>
<td>0.033</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>18.4</td>
<td>33</td>
<td>0</td>
<td>0.645</td>
<td>0</td>
<td>0.652</td>
<td>6</td>
<td>16.5</td>
<td>0.042</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>23.7</td>
<td>59</td>
<td>2</td>
<td>0.794</td>
<td>0.048</td>
<td>0.63</td>
<td>2</td>
<td>17.5</td>
<td>0.051</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>14.1</td>
<td>25</td>
<td>0</td>
<td>0.686</td>
<td>0</td>
<td>0.601</td>
<td>12</td>
<td>5.5</td>
<td>0.066</td>
</tr>
</tbody>
</table>

*P-value is defined as the proportion of simulation showing a lower \( \Delta \pi \) or a higher \( F_{ST} \) than the observed value.
could thus simply correspond to stochastic variations of the bottleneck process (Thornton et al. 2007). This result is congruent with the very low number of windows detected by the outliers approach. Finally, power analyses (see Appendix S1 Supporting information) revealed that the strong bottleneck experience by the African rice might limit our power to detect outlier windows. Power analyses indicate that the power drops rapidly below \( \alpha = 0.05 \) and is 55% at \( \alpha = 0.017 \) (i.e. the estimated value).

The cost of domestication: fixation of slightly deleterious alleles in *O. glaberrima*

A potential consequence of the strong genome-wide bottleneck experienced by *O. glaberrima* during domestication is a decrease in the efficacy of selection genomewide. This could lead to higher rates of fixation of slightly deleterious mutations. To test this hypothesis, we computed the number of derived synonymous (Fs) and nonsynonymous (Fn) SNPs that are fixed in one population but polymorphic or fixed for the ancestral state (defined as the *O. sativa* and *O. meridionalis* state see Method) in the other population (Fig. 6). These absolute numbers are divided by the number of synonymous and nonsynonymous sites, respectively, to obtain a ratio equivalent to Dn/Ds. We found significantly more Fn relative to Fs in the domesticated population (Dn/Ds = 0.19) compared with the wild population (Dn/Ds = 0.12, Fisher’s exact test, \( P < 2.10^{-5} \)). Interestingly, this difference in Dn/Ds between populations is noticeably higher in transcripts with low expression level (lower than the median RPKM) compared with transcripts with high expression level (higher than the median RPKM) (lower than the median RPKM).

**Fig. 6** The ratio of nonsynonymous (Dn) over synonymous (Ds) divergence in *O. glaberrima* and *O. barthii* in highly and lowly expressed genes (genes categorized according to the median of the expression level).

...
to this low diversity. However, as seeds of wild individuals come from isolated lines kept in greenhouse, our estimated \( F_{\text{IS}} \) of 0.87 is likely an overestimation of the \( F_{\text{IS}} \) of natural populations and should be taken with caution.

Across the genome, we found recombination rates and gene density to be the best predictors of local genetic diversity. A positive effect of recombination rates and a negative effect of gene density are clearly in line with an effect of natural selection (Cutter & Payseur 2013) that reduces genetic diversity through hitchhiking with beneficial (Smith & Haigh 1974; Wiehe & Stephan 1993) or deleterious mutations (Charlesworth et al. 1993; Hudson & Kaplan 1995). Correlations between recombination and polymorphism are usually weak because low recombination rates are sufficient to break down selective interferences. In the African rice, high selfing rate generates high LD, which extends the effect of selection over hundreds or thousands of kb and likely contributes to the rather strong correlation we detected. Similarly, a strong correlation between recombination and polymorphism was also found in the highly selfing nematode, Caenorhabditis elegans (Cutter & Payseur 2003). In a recent study, Flowers et al. (2012) found an intriguing pattern where both gene density and recombination rates had a negative effect on nucleotide diversity in the Asian rice genome – a result not expected by theoretical predictions. Flowers et al. (2012) interpreted this negative correlation as a by-product (statistical artefact) of the positive correlation between gene density and recombination rates. Not surprisingly, the coding-site density is also strongly positively correlated with recombination rates in our data set (Spearman’s \( r = 0.58, P < 10^{-16} \)). We manage to found both a positive effect of recombination rates and a negative effect of gene density probably because we analysed a considerably larger data set than Flowers et al. (2012) (552 loci in Flowers et al. (2012) vs more than 12,000 in the present data set) therefore increasing the power to disentangle the effect of each explanatory variable.

**Severe bottleneck and high cost of domestication in the African rice**

We found that African rice experienced a severe bottleneck during domestication. Assuming \( D_{\text{dom}} = 1000 \) generations for the duration of the bottleneck, the population size was reduced to less than 2% of the initial population size, that is, \( N_{\text{bot}} \approx 1200 \) individuals. This corresponds to a bottleneck intensity of \( k = N_{\text{bot}} / D_{\text{dom}} \approx 1.2 \), which is twice stronger than the one estimated in maize, \( k = 2.45 \) (Wright et al. 2005). Several reasons can explain such a strong bottleneck. The area of domestication could have been very limited or the selection initially strong. However, archaeological records suggest rather slow evolutionary process, at least for initial steps (Tanno & Willcox 2006; Purugganan & Fuller 2009). High selfing could also amplify the bottleneck strength as selection can extend genome-wide, as discussed above (see also Caicedo et al. 2007; Glémin & Bataillon 2009). Finally, \( O. glaberrima \) was largely replaced by \( O. sativa \) after the latter was introduced in West Africa during the 16th century (Bezancon 1993). This may also have contributed to the genetic impoverishment of the African rice.

A potential consequence of a strong domestication bottleneck is the accumulation of deleterious mutations as already observed in dog (Björnerfeldt et al. 2006; Cruz et al. 2008), yeast (Gu et al. 2005) and yak (Wang et al. 2011). The effect could be increased by hitchhiking of deleterious alleles especially in low recombining selfing species (Hartfield & Otto 2011; Hartfield & Glémin 2014). In agreement with this hypothesis of a cost of domestication, we observed an excess of fixation of derived nonsynonymous mutations in \( O. glaberrima \) compared with \( O. barthii \), especially for weakly expressed genes and in genomic regions that suffered from stronger genetic drift (Fig. 6). Increased fixation of nonsynonymous mutations could also be explained by relaxed selection on certain genes due to friendlier environmental conditions in agro-systems. However, this hypothesis does not explain the stronger accumulation in genomic regions that experienced higher drift and would require that relaxed selection had only affected lowly expressed genes. Accumulation of weakly deleterious mutations is thus the most likely hypothesis to explain our results. Lu et al. (2006) reported a similar result on the Asian rice. However, the authors compared \( D_n/D_s \) ratios between the two Asian rice groups, \( O. sativa \) ssp japonica and \( O. sativa \) ssp indica, with \( D_n/D_s \) ratios computed with a much more distant species, \( O. brachyangta \), thus mixing polymorphism, for the domesticated group, and divergence, for the wild one. Therefore, they did not take into account the confounding effect of divergence time on \( D_n/D_s \), as \( D_n/D_s \) is expected to decrease with divergence time (Rocha et al. 2006; Peterson & Masel 2009; Wolf et al. 2009; Dos Reis & Yang 2013; Mugal et al. 2013b). Here, we directly compared the domesticated and the wild populations and clearly separated polymorphic from fixed mutations. We thus provide here a clear evidence for a high cost of domestication in the African rice, which still needs to be confirmed in the Asian rice.

**Which targets of selection during the domestication process?**

In our analysis, we identified only very few genomic regions as candidate targets of positive selection during
domestication (e.g. only two windows using the Δπ statistics). As an alternative, we also focused on loci that have been identified as important determinant of agronomic traits in the Asian rice such as shk, qSH1, qSW5 or SD1 (Sang & Ge 2013). A close inspection of these loci in our data sets did not reveal any particular pattern of genetic diversity. This lack of positive result surely does not indicate an absence of selection during the African rice domestication. Rather, it more likely reflects the difficulty to identify selection after a very strong population bottleneck. The pattern of molecular variation created by population bottlenecks can be surprisingly similar to positive selection, and variations in allele frequencies induced by genetic drift can mask the effect of selection (Thornton et al. 2007). This difficulty has already been pointed out in other cultivated plants (e.g. Sorghum bicolor; Hamblin et al. 2006). Our simulations demonstrated that when the bottleneck intensity is moderate (e.g. α = 0.1 over 1000 generations), FST and especially Δπ appear powerful to detect positively selected genomic regions (Innan & Kim 2008). In our case, however, we estimated the bottleneck to be much more intense (α ~ 0.02), which decreases the statistical power. An alternative explanation might be that selection has affected very large regions throughout the genome of O. glaberrima. Here, the high selfing rate of the African rice could broaden the impact of any selective sweep, as also suggested for the Asian rice (Caicedo et al. 2007). This would lead to an overestimation of the bottleneck intensity and, subsequently, an underestimation of the number of outlier regions. Interestingly, Andersen et al. (2012) provided convincing evidences of selective sweeps occurring at the scale of a chromosome in the selfing species Caenorhabditis elegans.

**Implications for breeding**

In addition to providing the first genome-wide characterization of genetic diversity evolution in the African rice domestication, our results also bear implications for rice breeding. As previously noted (e.g. Li et al. 2011), genetic diversity is very low in the cultivated compartment, likely the lowest value reported in plants so far (compared with Jeffers et al. 2012 and Glémin et al. 2006). Such very low genetic diversity may challenge breeding programmes. Beyond introducing interesting agronomic traits from Asian rice, as in NERICA varieties, broadening the usable genetic diversity of O. glaberrima should be considered as a priority goal for breeding programmes, for instance using O. barthii material. Another, overlooked, strategy would be not to focus on improving specific traits but more globally on purging the genetic load accumulated during the domestication process. Again, the use of O. barthii could be useful in breeding programme to re-introduce fitter alleles lost during the domestication bottleneck. Our results show that population genomic approaches may help targeting genes or genomic regions for which purging would be necessary. Finally, our results suggest that high selfing rate has likely contributed to the reduced diversity and load in O. glaberrima. Increasing outcrossing and recombination in this species should also improve the efficiency of breeding strategies.

**Acknowledgements**

This work was supported by Agropolis Fondation under the ARCAD projet No 0900-001 (http://www.arcad-project.org/) which funded BN and GS salaries. We are grateful to Laure Sauné for experimental help, Nicolas Galtier and Vincent Caïs for helpful discussions and help with the REAS2SNP software and Yves Clément for help with manuscript redaction. We also thank Jeffrey Ross-Ibarra and two anonymous reviewers for helpful comments on the manuscript. This is ISEM publication number ISEM 2014-034.

**References**


Li WH (1977) Distribution of nucleotide differences between two randomly chosen cistrons in a finite population. Genetics, 85, 331–337.
Mugal CF, Nabholz B, Ellegren H (2013a) Genome-wide analysis in chicken reveals that local levels of genetic diversity are mainly governed by the rate of recombination. BMC Genomics, 14, 86.
R Core Team (2013) R: A Language and Environment for Statistical Computing. Austria, Vienna.
Wolf JBW, Künstner A, Nam K, Jakobsson M, Ellegren H (2009) Nonlinear dynamics of nonsynonymous (dN) and
synonymous (ds) substitution rates affects inference of selection. Genome Biology and Evolution, 1, 308–319.


S.G. and J.D. designed research. B.N. performed the research. B.N. and S.G. analysed the data and wrote the paper. G.S., F.S., B.N. and M.R. handled the Illumina raw data and performed the mapping. S.S., F.S., H.A., S.N. and A.G. selected and collected the sampled, performed RNA extraction and the Illumina library production and contributed new reagents.

Data accessibility


Supporting information

Additional supporting information may be found in the online version of this article.

Fig. S1 Demographic scenario used in the coalescent simulation. Parameters in bold are estimated, the others are fixed in the ABC analysis. \( N_{\text{pop}} = \alpha N_0 \), see text for detail.

Fig. S2 Number of heterozygous sites between individuals of O. barthii (RS) and O. glaberrima (RC)

Fig. S3 Maximum-likelihood estimation according to the number of populations (K) in STRUCTURE analyses. Red and black lines correspond to the two MCMC chains analysed.

Fig. S4 Phylogenetic tree of the 60 loci positioned between the position 4 Mb and 6 Mb of the chromosome 5 including Oryza sativa indica sequences. Phylogenetic tree reconstructed using BIONJ clustering method and TN93 substitution model. RS represent O. barthii individuals and RC, O. glaberrima. Nodes with a bootstrap supports above 95% (100 replicates) are indicated with a asterisks.

Fig. S5 LD decay (\( r^2 \)) in a sample of 9 O. barthii and 7 O. glaberrima, plotted as a function of the genetic distance between SNP. Blue and green lines indicate lowest fits of O. glaberrima and O. barthii data, respectively compute using the \( R \) ’loess’ function.

Fig. S6 Distribution of summary statistics obtained by posterior predictive simulations (simulations under estimated parameters). On each panels, vertical red line indicate the observed value of the statistic.

Fig. S7 The likelihood surface fitting \( f \) the proportion of genes in the severe bottleneck (selected) class (\( \alpha = 0.0017 \)).

Fig. S8 Relation between the ratio of nonsynonymous (Dn) over synonymous (Ds) divergence and the ratio of O. glaberrima nucleotide diversity (\( \pi_{g\text{glab}} \)) over O. barthii nucleotide diversity (\( \pi_{\text{bot}} \)) computed over 2 Mb windows. Black line is the linear regression line.

Fig. S9 Power analysis of \( \Delta \pi \) and \( F_{ST} \) according to the bottleneck intensity \( \alpha = N_{\text{pop}}/N_0 \). Red dots correspond to the estimated value of \( \alpha = 0.017 \). Sample size for the wild and the domesticated population is \( n = 9 \) for the wild and \( n = 8 \) for the domesticated population in both cases.

Fig. S10 Power analysis of \( \Delta \pi \) and \( F_{ST} \) according to the bottleneck intensity \( \alpha = N_{\text{pop}}/N_0 \). Red dots correspond to the estimated value of \( \alpha = 0.017 \). Sample size for the wild and the domesticated population is \( n = 20 \) in both cases.

Appendix S1 Supplementary methods and analyses.

Table S1 Details on the individuals sampled.

Table S2 Sequencing and mapping performance of cultivated (RC) and wild (RS) African rice.

Table S3 Basic population genetic statistic with various parameters and option used in reads2snp.

Table S4 \( R^2 \) Estimates and \( P \)-values in a multi-linear regression analysis for potential genomic explanatory variables of synonymous nucleotide diversity (\( \pi_s \)) of domesticated (O. glaberrima) and wild (O. barthii) population excluding the chromosome 11.

Table S5 Mean Dp mean \( F_{ST} \) and \( P \)-val associated with randomization test for the 45 GO-slims terms of category ‘Biological process’. 