Theor Appl Genet (2014) 127:2211–2223
DOI 10.1007/s00122-014-2374-z

ORIGINAL PAPER

An extensive analysis of the African rice genetic diversity through a global genotyping

Julie Orjuela · François Sabot · Sophie Chéron · Yves Vigouroux · Hélène Adam · Harold Chrestin · Kayode Sanni · Mathias Lorieux · Alain Ghesquière

Received: 16 December 2013 / Accepted: 29 July 2014 / Published online: 15 August 2014

Abstract

Key message **We present here the first curated collection of wild and cultivated African rice species. For that, we designed specific SNPs and were able to structure these very low diverse species.**

Abstract *Oryza glaberrima*, the cultivated African rice, is endemic from Africa. This species and its direct ancestor, *O. barthii*, are valuable tool for improvement of Asian rice *O. sativa* in terms of abiotic and biotic stress resistance. However, only a few limited studies about the genetic diversity of these species were performed. In the present paper, and for the first time at such extend, we genotyped 279 *O. glaberrima*, selected both for their impact in current breeding and for their geographical distribution, and 101 *O. barthii*, chosen based on their geographic origin, using a set of 235 SNPs specifically designed for African rice diversity. Using those data, we were able to structure the individuals from our sample in three populations for *O. barthii*, related to geography, and two populations in *O. glaberrima*; these two last populations cannot be linked however to any currently phenotyped trait. Moreover, we were also able to identify misclassification in *O. glaberrima* as well as in *O. barthii* and identified new form of *O. sativa* from the set of African varieties.

Introduction

In the modern agriculture, two species from the *Oryza* genus were independently domesticated: the Asian rice *Oryza sativa* L., and the African rice *Oryza glaberrima* Steud. (Vaughan et al. 2008). The Asian rice was domesticated from the wild rice *O. rufipogon* more than 10, 000 years ago, in Southern Asia (Huang et al. 2012). The cultivated African rice derived from the wild species *O. barthii* probably originated from the Niger River delta, around 1,000 BC (Portères 1962; Linares 2002; Murray 2004). The two ancestors of Asian and African cultivated species probably diverged between one and two million years ago (Ma and Bennetzen 2004; Vaughan et al. 2008). This divergence is associated with an important reduction of the genetic diversity of the African wild species compared to the Asian one (Vaughan et al. 2008).

O. sativa was introduced in West Africa by European navigators in sixteenth century (Bezançon 1993), and this introduction endangered the conservation of local genetic resources (Barry et al. 2007). Despite its status of endemic and cultivated species, *O. glaberrima* was progressively replaced by modern, more yield-efficient *O. sativa* varieties. In the second half of the twentieth century, African rice culture was based in a parceled system where some varieties of *O. sativa* and *O. glaberrima* were cultivated together

J. Orjuela and F. Sabot contributed equally to the work.

Communicated by Takuji Sasaki.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-014-2374-z) contains supplementary material, which is available to authorized users.

J. Orjuela · F. Sabot (✉) · S. Chéron · Y. Vigouroux · H. Adam · H. Chrestin · M. Lorieux · A. Ghesquière
DIADE UMR IRD/UM2, 911 Avenue Agropolis, BP 64501,
34394 Montpellier Cedex 5, France
e-mail: francois.sabot@ird.fr

K. Sanni
AfricaRice Center, 01 BP 2031, Cotonou, Benin

Present Address:
K. Sanni
AATF, 30709-00100 Nairobi, Kenya

in mixtures by farmers in upland and rainfed lowland environments (Linares 2002). Moreover, *O. sativa* is sometimes cultivated close to the wild African rice *O. barthii*.

O. glaberrima is nowadays generally grown in West Africa, in small surfaces, because of its lower yield compared to *O. sativa*. Yet, it is more adapted to rustic African conditions than *O. sativa* (Linares 2002) and presents resistance or tolerance to viruses (Albar et al. 2003), nematodes, bacteria, drought, iron toxicity and high salinity (Linares 2002). Thus, despite its supposed low level of genetic diversity, *O. glaberrima* is a very valuable tool for *O. sativa* improvement, through interspecific hybridization. However, a strong sterility barrier, mainly driven by the S_1 locus (Sano, 1990; Garavito et al. 2010), exists between these two recently diverged species, which complicates the use of *O. glaberrima* in *O. sativa* breeding. Still, it is possible to obtain fertile "hybrid" varieties through embryo rescue or manual fertilization and backcrossing, as it was performed in production of NERICA varieties (NEw RICe for Africa, Gridley et al. 2002). Despite this well-known reproductive barrier, several unexpected natural hybrid forms were reported in collections obtained through systematic prospections in Africa, with intermediate genotypes between *O. sativa* and *O. glaberrima* (Semon et al. 2005; Barry et al. 2007; Nuijten et al. 2009; Dramé et al. 2011).

One other factor, which complicates the introgression of African traits into cultivated Asian rice, is the lack of studies about African local genetic resources. One of the most extensive studies so far (Semon et al. 2005) found a genetic structure of *O. glaberrima* accessions in five different groups using *O. sativa* SSR markers. Two of these groups clustered with *O. sativa*, suggesting that some *O. glaberrima* accessions represent some degree of admixtures. The remaining three *O. glaberrima* subpopulations were associated with phenotypic traits, which might reflect some form of ecological adaptation. A more recent study, using 14 nuclear loci (Li et al. 2011), showed that *O. glaberrima* accessions do not exhibit such a clear structure. This last study also suggest a single domestication origin of African rice in areas of the Upper Niger and Sahelian Rivers (Li et al. 2011). Results of these two studies must be however considered with caution as the wild species sample number is very limited.

The objective of the current study is to understand the genetic diversity of a large sample of the two African species, *O. glaberrima* and *O. barthii*, using a unique collection covering the whole West African range of *O. glaberrima* and *O. barthii*. To evaluate this diversity, we designed a set of SNPs specific for wild and cultivated African rice species. We also added to our samples Asian rice varieties and lab-made hybrids between African and Asian rices.

Our study clearly separates *O. sativa* from the African rice species (*O. barthii* and *O. glaberrima*). We clarify the status of various accessions previously identified as hybrids between *O. sativa* and *O. glaberrima*. More importantly, we find clear differentiation between wild and cultivated rice in Africa and evidence a geographical population structure in the wild African species only. Finally, we provide highly valuable information on the most genotyped and curated African rice collection available so far.

Materials and methods

SNP design and selection

Four sets of sequencing data were used for the Illumina VeraCode chip design. First, we sequenced two *O. glaberrima* individual plants using 76 bases pair-ended (PE) Illumina sequences (Tog5307 and Tog5681, 18 × and 25 × depth, respectively) on a *GaIIx* machine (*Integragen*, Evry, France). Second, we obtained two bulks of 100 bases PE Illumina sequences at 37× depth each (*HiSeq 2000* machine, *Fasteris*, Geneva, Switzerland) using nine individuals from *O. glaberrima* and nine from *O. barthii* (the same as in Nabholz et al. 2014; see Supplemental Table 1). The quality and potential contamination for each dataset were checked using *FASTQC* (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were cleaned using *CutAdapt* (Martin 2011) with the appropriate adapter sequences (QUAL threshold = 20, minimal length = 20, overlap = 7), and were mapped using *BWA* (Li and Durbin 2009). We used as reference the MSU 6.1 assembly of *O. sativa ssp japonica* cv Nipponbare (http://rice.plantbiology.msu.edu/index.shtml), allowing three mismatches and one gap ($n = 3$ and $e = 1$, respectively). The resulting SAM files were cleaned to conserve only correctly mapped pairs (samtools view filter $f = 0 \times 02$), and optical duplicates removed, using *SAMtools* (Li et al. 2009).

Using the *GATK* framework (McKenna et al. 2010), after realignment of BAM files as recommended, the polymorphic positions between NipponBare and at least one of the four African datasets were called (through *UnifiedGenotyper*), using standard criteria: min QUAL = 20, max cov = 250, min cov = 10. More than 9 millions of raw SNPs and InDels were recovered at this step. We performed a second filter using *VariantFiltrator*, with classical conditions (QUAL = 50; SNP only; MQ0 < 10 % or MQ0 < 4; clusterWindowsize = 60; clusterSize = 2); around 365 000 SNPs succeeded this initial filtering.

We then use a set of home-made *Perl* scripts (available on request) to continue to filter the VCF file using as main criteria a 10× coverage for each of the four African datasets, and a polymorphism in at least one of the bulk sample (homozygous reference or alternate for the individual sequences, heterozygous for the bulk sequences). 23,307

SNPs were still valid at this step, being polymorphic within the African rice species. We then split the reference genome in 1 Mbases bin, obtaining 384 bins, in which we choose 10 SNPs centered around the middle 500 kb position (−250/+ 250 kb) in each bin.

We then added 153 SNPs from the 44 k Chipset (Zhao et al. 2011) that are potentially polymorphic in African rice species (based on our sequencing data, as described earlier) and that are not clustered (60bases window), to obtain a set of almost 4,000 SNPs of high quality. Those 4000 SNPs were tested for *VeraCode* efficiency, and a final set of 384 SNPs was selected for the analysis, including 49 SNPs from the 44 k chip (Supplemental Table 2). All the SNPs we selected here are not only polymorphic between NipponBare and the African rice species, but also polymorphic within the African rice species, and thus cannot be used as markers for introgression of one compartment within the other (Asian vs African); as for each point the two alleles are present in the African rice pool.

Plant materials

A reference set of *O. glaberrima* was established jointly between IRD and Africa Rice, to gather all important accessions identified during previous field and greenhouse evaluations: parents of the NERICA varieties, important reference populations, donors of important traits for adaptation to biotic and abiotic stresses (Linares 2002; Albar et al. 2003; Vaughan et al. 2008; Thiémélé et al. 2010), and other cultivars. They represent as much as possible the geographical distribution of the species in West Africa before 1950, and before its progressive replacement by *O. sativa* as supposed by Portères (1962).

Regarding *O. barthii*, a special effort was made to represent actual, true wild forms of this species, which are poorly represented in current rice collections and documentations. The accessions correspond to population samples collected in remote areas and characterized, within other traits, by large coarse grains, as described by Second (1982). As a result, 102 accessions were selected from the complete set of IRD collection (formerly known as ORSTOM collection), built between 1974 and 1982 during prospection missions supported by IBPGR (International Board of Plant Genetic Resources, now Biodiversity) and the CGIAR Generation Challenge Program (Consultative Group on International Agricultural Research).

In total, our African Rice collection is made of 86 wild and 16 weedy *O. barthii* and of 279 *O. glaberrima* accessions from 18 different recorded origins (see Table 1). Geographical distribution is shown in Fig. 1, and correspondences to others collection names are given in the Supplemental Table 3. Fourteen of those accessions are common with the ones used in Li et al. (2011).

Table 1 Type and geographical origin of African rice accessions used in the current study

Origin	*O. barthii*		*O. glaberrima*
	Weedy	Wild	
Botswana		2	
Burkina Faso			7
Cameroon	2	15	15
Yvory Coast			28
Gambia			4
Ghana			9
Guinea	1		41
Guinea Bissau	1	1	1
Chad			8
Lake Chad		44	
Liberia			18
Mali	12	17	34
Nigeria		3	55
Senegal		1	30
Sierra Leone			8
Tanzania		1	5
Zambia		1	
Zimbabwe			3
Total	16	85	266[a]

[a] Total not including 9 *O. glaberrima* and 4 off-type accessions reclassified in this study

We completed this set with 53 accessions that well represent the *O. sativa* diversity (from *HaplOryza* project; Billot et al. 2007), also called "mini-core collection", 4 *O. longistaminata* as outgroup, 24 CSSLs (Chromosome Segment Substitution Lines), BC3DH plants between *O. glaberrima* and *O. sativa* as recurrent parent from CIAT (Gutiérrez et al. 2010), and five laboratory-created hybrids in BC1 (with *O. sativa* as recurrent parent) stages. The total number of studied materials is 467.

All the descent from the here-tested African rice plants are available on request at the Genetic Resources Unit from the AfricaRice Center.

DNA extraction, target probe preparation and array hybridization

Total genomic DNA from a single plant per accession was extracted for all the genotypes using standard CTAB method, diluted in pure water and standardized at 50 ng/µl at *ADNid* (Clapiers, France). Genotyping assays were carried out following the manufacturer's instructions and protocols for the GoldenGate Genotyping Assay for VeraCode Manual Protocol (Illumina Part # 11275211) at *ADNid* (Clapiers, France). Ten artificial DNA mix (*O sativa* ssp *japonica* cv NipponBare and *O. sativa* ssp *indica* cv

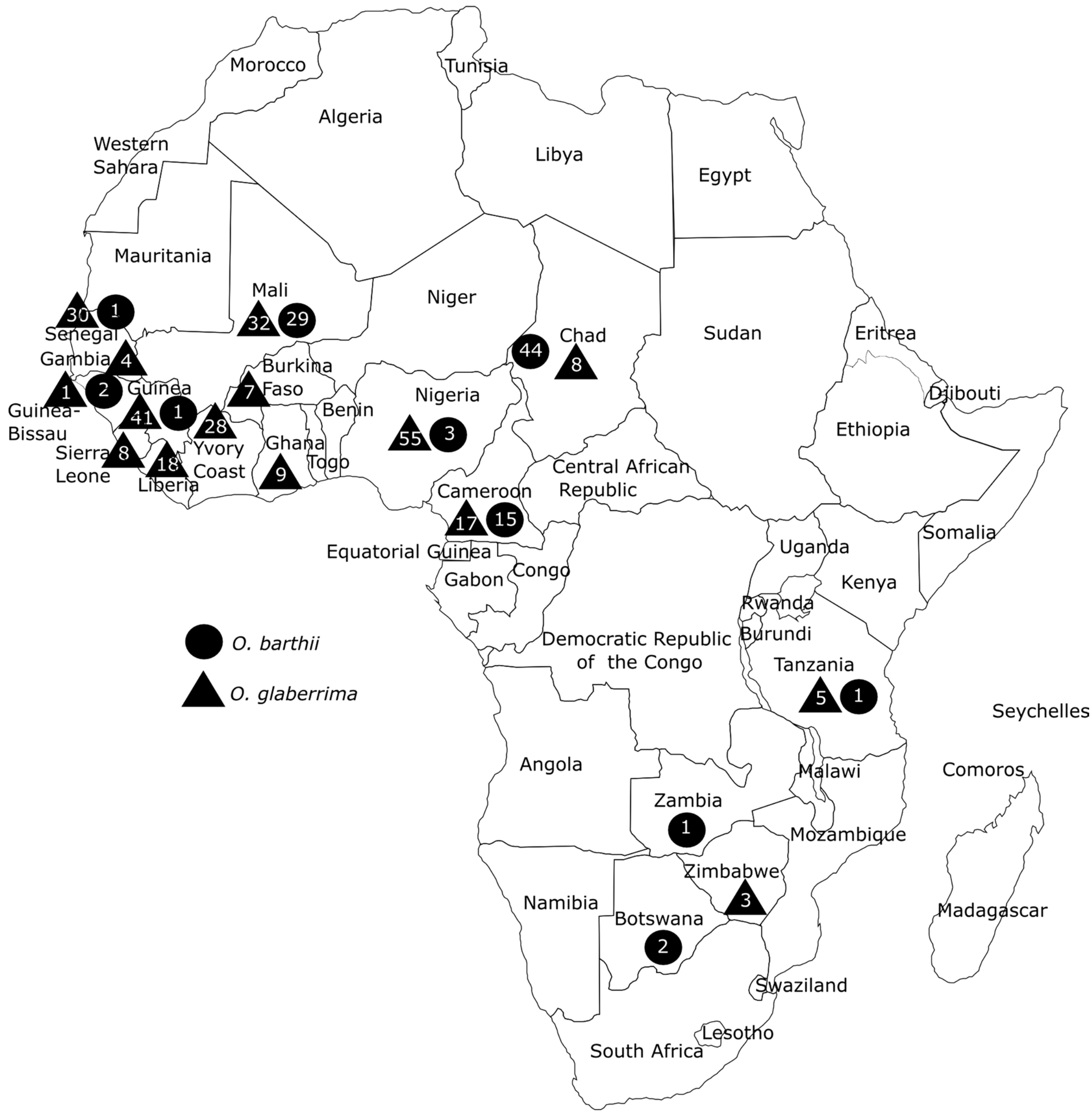


Fig. 1 Geographical distribution of the African Rice accessions used in this study. The number of accessions of *O. glaberrima* (*triangles*) and of *O. barthii* (*circles*) are shown for each country

IR64) were introduced in the plates as positive controls for hybridization detection, as well as five pure-water negative controls.

SNP allele calling

The SNP data from each plate were recorded, and allele calls manually curated using the Genotyping module (v 1.6.3) of the Illumina *GenomeStudio* software package v 2010.1 (more details can be found in Thomson et al. 2011). Allele recoding for SNP visualization was also performed using *FLAPJACK* graphical genotyping software (Milne et al. 2010). The SNP data were loaded with an initial GenCall Threshold of 0.25. *GenomeStudio* was able to identify the heterozygous clusters (CSSL and artificial heterozygous mix controls).

Analysis of genetic relatedness between accessions

As part of the diversity analysis, *STRUCTURE* v 2 software (Pritchard et al. 2000) was used to determine genotype assignments to subpopulations, with increasing *K* (population number) values from 2 to 14 and running with 100,000 burn-ins and 1,000,000 iterations. The *K* value with the highest *Ln* P(D) values was retained, as described by Evanno et al. (2005). Plants with an arbitrary ancestry percentage of min. 70 % were assigned to each cluster. In addition, PCA analyses and Neighbor-Joining trees were constructed for population sets with the software *Darwin* v 5.0 software (Perrier and Flori 2003). All graphics and trees were edited using the *MEGA* v 5.0 software (Tamura et al. 2011).

Genetic diversity analysis

Genetic diversity indexes were compared among the clusters identified by *STRUCTURE*, using only those plants that were assigned to each cluster. Observed heterozygosity (*Ho*), expected heterozygosity (*He),* genetic differentiation between populations (F_{ST}) and molecular variance analysis (AMOVA; Excoffier et al. 2005) including fixation indexes such as F_{IS} and F_{IT} were performed using the program *Arlequin* (Excoffier et al. 2005) and *Genepop* v 4.2 (Rousset and Raymond 1995); statistical significance of each variance component was assessed based upon 1,023 permutations of the data and a significance level of 0.05.

Results

SNP set validation

Of the 384 SNPs designed using the in silico approach, 235 were polymorphic and retained for used in subsequent analysis. Of the 149 negative points, three were negative (no results), 18 were of low quality (high levels of missing data—20 % or more—*GenTrain* value lower than 0.5, *ClusterSet* value lower than 0.21) and the remaining 128 points were monomorphic among all samples. After posterior analyses, 74 % (110/128) of those monomorphic SNPs were identified as SNP mostly associated with a transversion mutation type (Transition/Transversion ratio of 0.35), while only 34 % (80/235) was transversion in polymorphic SNPs (transition/transversion ratio of 1.9). For further analyses following such a protocol, we thus recommend to select preferentially transition-type SNPs, as described in Liu et al. (2012).

A 235-SNP set (see Supplemental Table 2) was used to genotype 467 accessions (Supplemental Table 3). Our set of 235 polymorphic markers consists of 188 African-specific variants (polymorphic in cultivated and wild African rice species, monomorphic in Asian rice species) and 47 *Oryza* variants (polymorphic within Asian as well as in African rice species). These *Oryza* variants are composed of 13 SNPs from our own set and of 34 SNPs transferred from the 44 k *O. sativa* chip (Zhao et al. 2011). These *Oryza* variants allowed to validate the methodological choice of our SNP set, as we obtained for *O. sativa* accessions the expected genetic described in Garris et al. (2005) (Fig. 2).

Artificial hybrids exist, not natural

Artificial control hybrids as well as CSSL controls have an intermediate position between the *O. sativa* and *O. glaberrima* groups in the PCA (Principal Component Analysis; Fig. 3), as expected. We also noted that none of the accessions from our collection clusters with the hybrid controls, suggesting that none of the individuals in our sample can be classified as hybrid. Although we cannot exclude old introgression events, we do not observe on the set of the tested accessions any plant with a balanced *sativa*/*glaberrima* genome composition, except for the CSSL and lab-made hybrid plants. Thirteen accessions classified as *O. glaberrima* species were clustered with *O. sativa*; among these, four (ID # 144, 145, 148 and 149) correspond to previously reported hybrids (Semon et al. 2005; Barry et al. 2007), the nine remaining being ID # 1, 224, 420, 421, 422, 493, 494, 499 and 510. Those 13 accessions were analyzed together with the *O. sativa* accessions using distance-based methods, to clarify their nature. PCA and NJ showed the classical structure of *O. sativa* accessions, with 144, 145, 148, 149, 224 and 510 accessions clustered with the group III of *O. sativa*. The second set of these *O. glaberrima* accessions (1, 420, 421, 422, 493, 494 and 499) clustered with the *indica* group (Fig. 4). In the *STRUCTURE* analysis, whatever is the number of populations (K), all those 13 individuals always clustered with *O. sativa* lines (Supplementary Fig. 1). These 13 accessions previously identified as *O. glaberrima* or of hybrid origin rather belong to the *O. sativa* species.

In addition, one weedy *O. barthii* accession (520A) clustered with the *O. sativa/O. longistaminata* group on the PCA. After having carefully checked its phenotype, this accession was reclassified as an *O. longistaminata*. In the following, we thus will refer to 101 *barthii* and no more to 102.

Genetic structure of cultivated rices

The *O. sativa* group is clearly separated from *O. glaberrima* and *O. barthii* using the PCA (Fig. 3) and on the Bayesian analysis using *STRUCTURE* (Supplemental Fig. 1; Supplemental Fig. 2a, b). In the PCA analyses, *O. sativa* accession distribution is rather limited (Fig. 3). In other words, their genetic variances were smaller in the first two axes than the variance from *O. glaberrima.* This is explained by the nature of the SNPs we used, designed to be more variable within *O. glaberrima* and *O. barthii*, than within *O. sativa*. This obviously does not mean that the African species are more variable than *O. sativa.*

The genetic differentiation index F_{ST} between these three species were of 0.212 ($p < 0.05$) between *O. barthii* and *O. glaberrima,* 0.435 ($p < 0.05$) between *O. barthii* and *O. sativa,* and 0.539 ($p < 0.05$) between *O. glaberrima* and

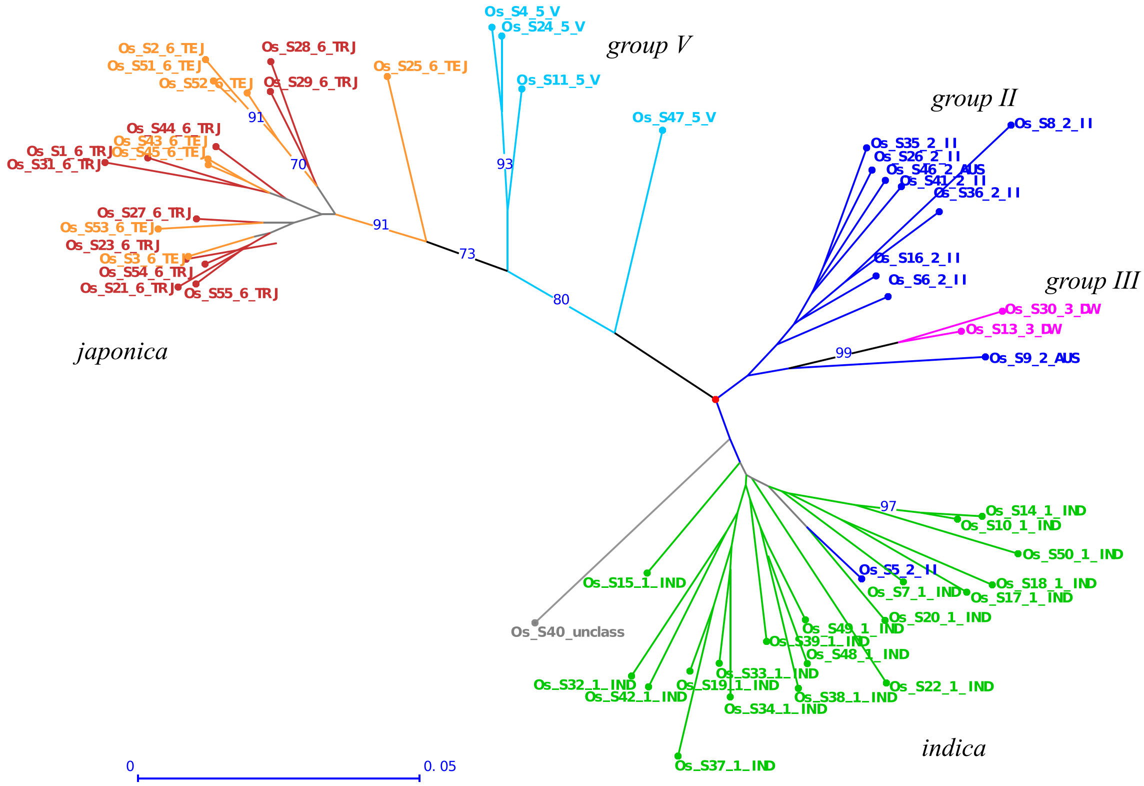


Fig. 2 Genetic structure of O. sativa varieties using the SNP set from the current study. *O. sativa ssp indica* and *japonica* as well as groups II, III and V are shown

O. sativa. As expected, *O. glaberrima* and *O. barthii* are closer together than any of them compared to *O. sativa*. The closer relationship between *O. barthii* and *O. sativa* compared to *O. glaberrima* vs *O. sativa* can be explained as the wild African rice being closer in terms of origin to Asian cultivated rice (Vaughan et al. 2005, 2008).

The observed heterozygosity (*Ho*) was 0.019, 0.014 and 0.080 for *O. barthii*, *O. glaberrima* and *O. sativa*, respectively, and the expected heterozygosity (*He*) 0.399, 0.317 and 0.346, respectively. These diversity indexes were significantly different between *O. barthii* and *O. glaberrima* ($p < 0.001$ and $p < 0.001$, respectively), indicating a high level of inbreeding. Genetic diversity values (*Hs*) were 0.379 (s.d 0.141), 0.288 (s.d 0.172) and 0.073 (s.d 0.161) for *O. barthii*, *O. glaberrima* and *O. sativa*, respectively. The low level of variability in *O. sativa* observed here is linked to the SNP choice (see the previous part). Fixation values between these three species were $F_{IS} = 0.953$, $F_{ST} = 0.383$ and $F_{IT} = 0.971$.

African rice diversity and genetic structure

Distance-based methods and genetic indexes confirmed, as expected for a wild species, that *O. barthii* is more diverse than *O. glaberrima* (Li et al. 2011). Even if *O. glaberrima* and its ancestor *O. barthii* are closely clustered and slightly overlapping, these two species are nevertheless well separated (Fig. 5; Supplemental Fig. 3; $F_{ST} = 0.212$). Overlapping of *O. barthii* in *O. glaberrima* distribution could be explained by the presence of 15 weedy forms of *O. barthii* within the *O. glaberrima* cloud (512A1, 514G1, 515A1, 517A1, 518G1, 521A1, 528A1, 530A1, 531A1, 532A1, 534A1, 535A1, 536A1, 557A1 and 559A1). These weedy *O. barthii* accessions are phenotypically different from true wild forms of *O. barthii* (data not shown). Only one weedy accession of *O. barthii* clustered with the wild forms (513A1) and only one accession of true wild *O. barthii* clustered with *O. glaberrima* (521W1). We were able to confirm the higher variability of *O. barthii* wild forms when comparing with weedy forms included in the *O. glaberrima* variability (Fig. 5; Supplemental Fig. 3). *STRUCTURE* analyses, using the method described in Evanno et al. (2005), confirmed the PCA and NJ results, with a maximal ΔK occurred at $K = 2$, followed by $K = 3$ (Supplemental Fig. 2c, d). When $K = 2$, *O. glaberrima* and *O. barthii* are well separated. When $K = 3$, groups were formed as follows: one group containing the wild *O. barthii* accessions, and two *O. glaberrima* groups (*Ogla_I* and

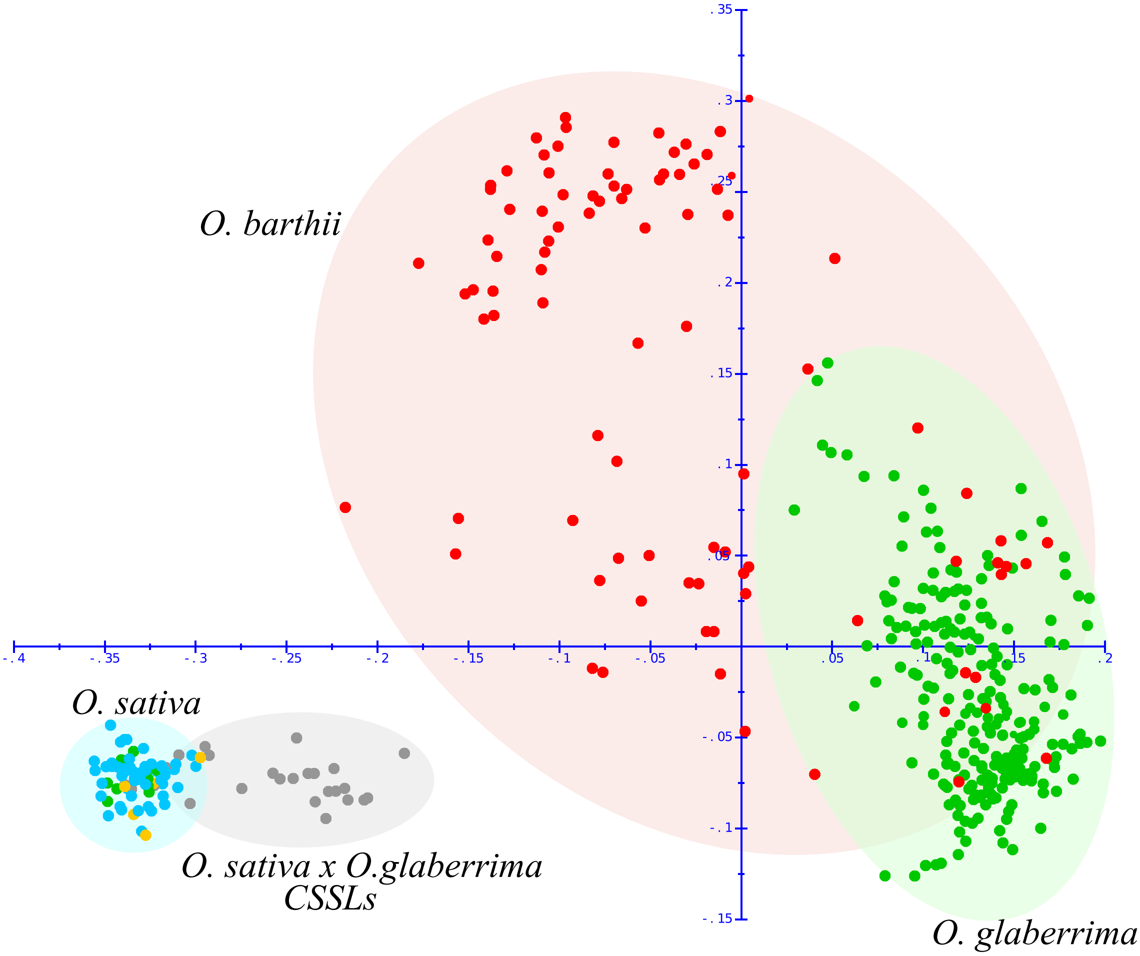


Fig. 3 Principal component analysis (PCA) of 279 *O. glaberrima* (*green*) (including four reported hybrids), 101 *O. barthii* (*red*), 53 *O. sativa* (*blue*), 29 *O. sativa* × *O. glaberrima* controls (*gray*) and 5 *O. longistaminata* (*yellow*; 4 initially known plus the *O. barthii* 520A accession, see text), based on 235 SNPs. The first eigenvector (*x* axis) explained 25.4 % and the second (*y* axis) explained 8.2 % of variation (color figure online)

Ogla_II) containing each one 5 and 11 *O. barthii* weedy accessions, respectively (see below).

Geographical structure of *O. barthii*

In the *STRUCTURE* analysis for the 101 *O. barthii*, we found that the maximal ΔK occurred at $K = 2$, with the next largest peak at K = 3 (Supplemental Fig. 2e, f; Supplemental Fig. 4a). At $K = 2$, *O. barthii* was divided based on its geographic distribution in *Obar_I* containing Niger River Delta region and West Africa accessions, and *Obar_II* group containing Chad Lake region, Central and Austral Africa varieties. At $K = 3$, *O. barthii* was divided into *pop1* (Central Africa), *pop2* (Delta of Niger River and West Africa) and *pop3* (Chad Lake region and Austral Africa varieties; Supplemental Fig. 4a). Using an arbitrary cut-off value of 70 % of ancestry for assignment, 19, 31 and 28 plants were attributed to the *pop* 1, 2 and 3, respectively; 6, 6 and 11 intermediate varieties were detected between 1–2, 1–3 and 2–3 groups, respectively. The relationship based on *STRUCTURE* ancestry analysis is shown in the Supplemental Fig. 4b.

The clusters identified in the *STRUCTURE* analysis were used to compare the genetic diversity of *O. barthii.* When $K = 2$, the population pairwise F_{ST} value between *Obar_I* and *Obar_II* was of 0.315 ($p < 0.05$). The observed heterozygosity *(Ho)* was of 0.013 and of 0.005 for *Obar_I and Obar_II* ($p < 0.001$). These values were lower than the expected heterozygosity (*He*), 0.365 and 0.338, respectively ($p < 0.001$). Similarly, when $K = 3$, the observed heterozygosity (*Ho*) was of 0.001, 0.021 and 0.008 for *pop1*, *pop2* and *pop3*, respectively. The expected heterozygosities *(He)* were 0.371, 0.332 and 0.324, respectively. All the pairwise diversity comparisons were significantly different at $p < 0.05$, except for the *He* between *pop2* and *pop3* ($p = 0.74$). The inbreeding index F_{IS} for these three clusters were 0.941, 0.996 and 0.974, respectively, confirming the

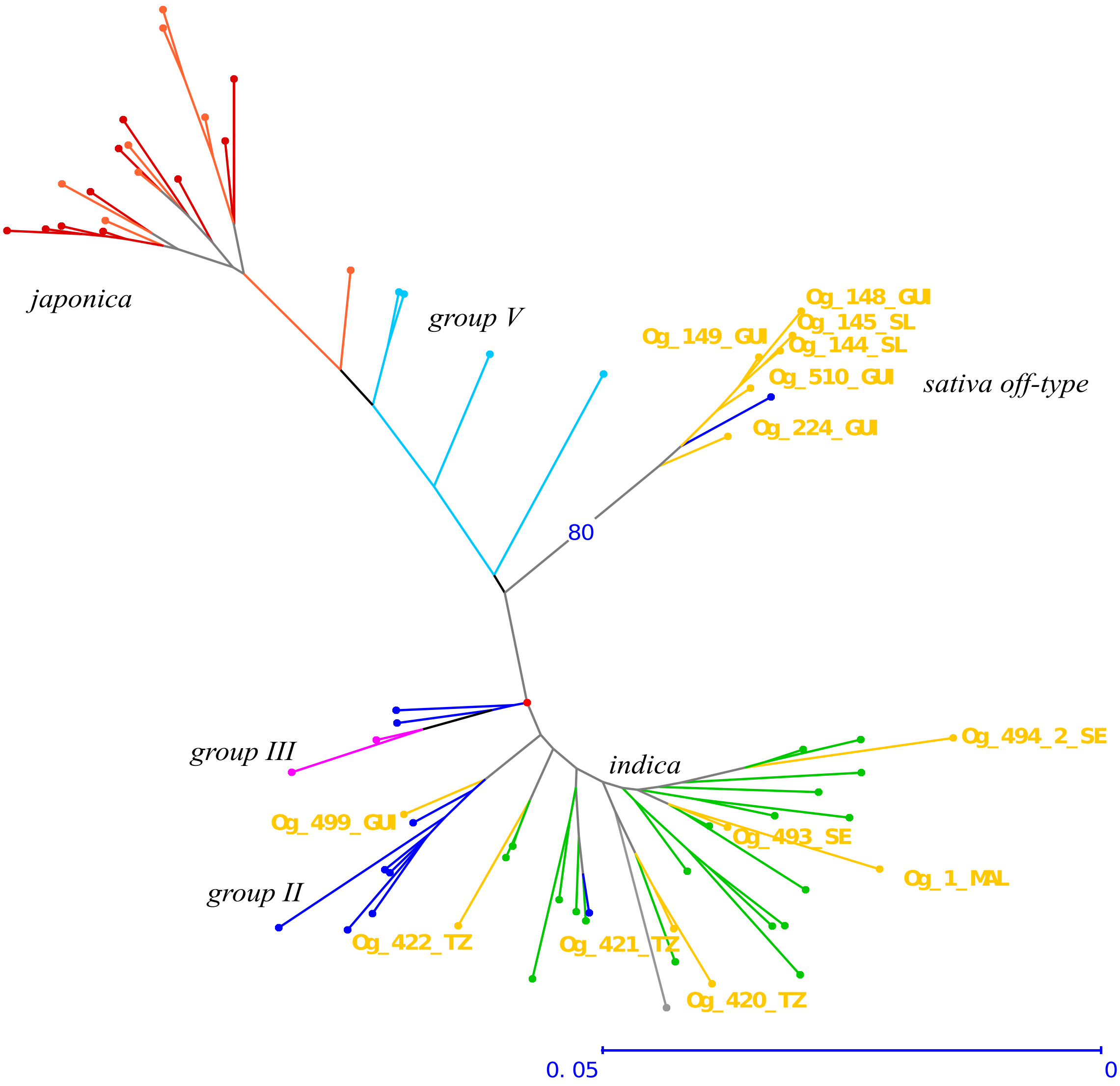


Fig. 4 Neighbor-Joining (NJ) tree including the complete set of *O. sativa* accessions and 13 African varieties (4 varieties reported as natural hybrids and 9 accessions firstly classified as *O. glaberrima*)

high level of selfing (0.87 for the whole species, Nabholz et al., 2014). F_{ST} value between *pop1* and *pop2* was of 0.306 ($p < 0.001$), between *pop1* and *pop3* of 0.244 ($p < 0.001$), and between *pop2* and *pop3* of 0.311 ($p < 0.001$), showing a good separation with limited outcrossing.

No ecological structure observed for *O. glaberrima*

The structure of *O. glaberrima* was then assessed using *STRUCTURE* (Fig. 6b). The maximal ΔK suggested two groups (Supplemental Fig. 2g, h; Supplemental Fig. 5). At $K = 2$, *O. glaberrima* is composed of *Ogla_I* and *Ogla_II* groups, representing 111 and 155 varieties (Fig. 6b), respectively. A total of 52 intermediates between groups *I* and *II* were detected when the arbitrary ancestry percentage of 70 % was applied. Similar to *O. barthii*, the clusters identified in the *STRUCTURE* analysis were used to compare the genetic diversity of *O. glaberrima*. Population pairwise F_{ST} value between *Ogla_I* and *Ogla_II* was of 0.282 ($p < 0.05$), suggesting that these two groups are moderately different and that a large fraction of their genetic diversity is shared. The observed heterozygoty (*Ho*) was of 0.009 and of 0.019 for *Ogla_*I and *Ogla_II*, respectively, with a significant difference ($p < 0.001$). The expected heterozygoty (*He*) are also significantly different ($p < 0.001$): 0.338 and 0.287 for *Ogla_*I and *Ogla_II*. The inbreeding index F_{IS} was 0.970 and 0.932 for *Ogla_I* and *Ogla_II*, respectively (no significant difference). However, unlike *O. barthii*, *O. glaberrima* structure did not translate into a clear geographical pattern, or a clear ecological one.

Discussion

O. glaberrima presents several key advantageous traits for *O. sativa* improvement, such as virus and bacteria resistance, salt and iron tolerance, and so on (Linares 2002). However, to be really efficient in such breeding programs, we first need a comprehensive picture of the diversity in *O. glaberrima*, and of its wild ancestor *O. barthii*. Our study allows a large analysis of these species by testing the diversity of a collection of individuals from various origins and ecologies as well as the largest *O. barthii* sample set to date. Moreover, we included individuals of interest in current breeding programs, e.g. *Tog10434/Og_161, IRGC112577/Og_87 and Tog7291/Og_172*, that show high levels of resistance to RYMV (Thiémélé et al. 2010; Orjuela et al. 2013). To analyze African rice diversity, we built a fast and low-cost 384 SNP *VeraCode* Chip designed to maximize African rice diversity. We did not directly used the *O. sativa* 44 k chip (Zhao et al. 2011), since it was developed for the Asian species and consequently presents specific bias. Using an

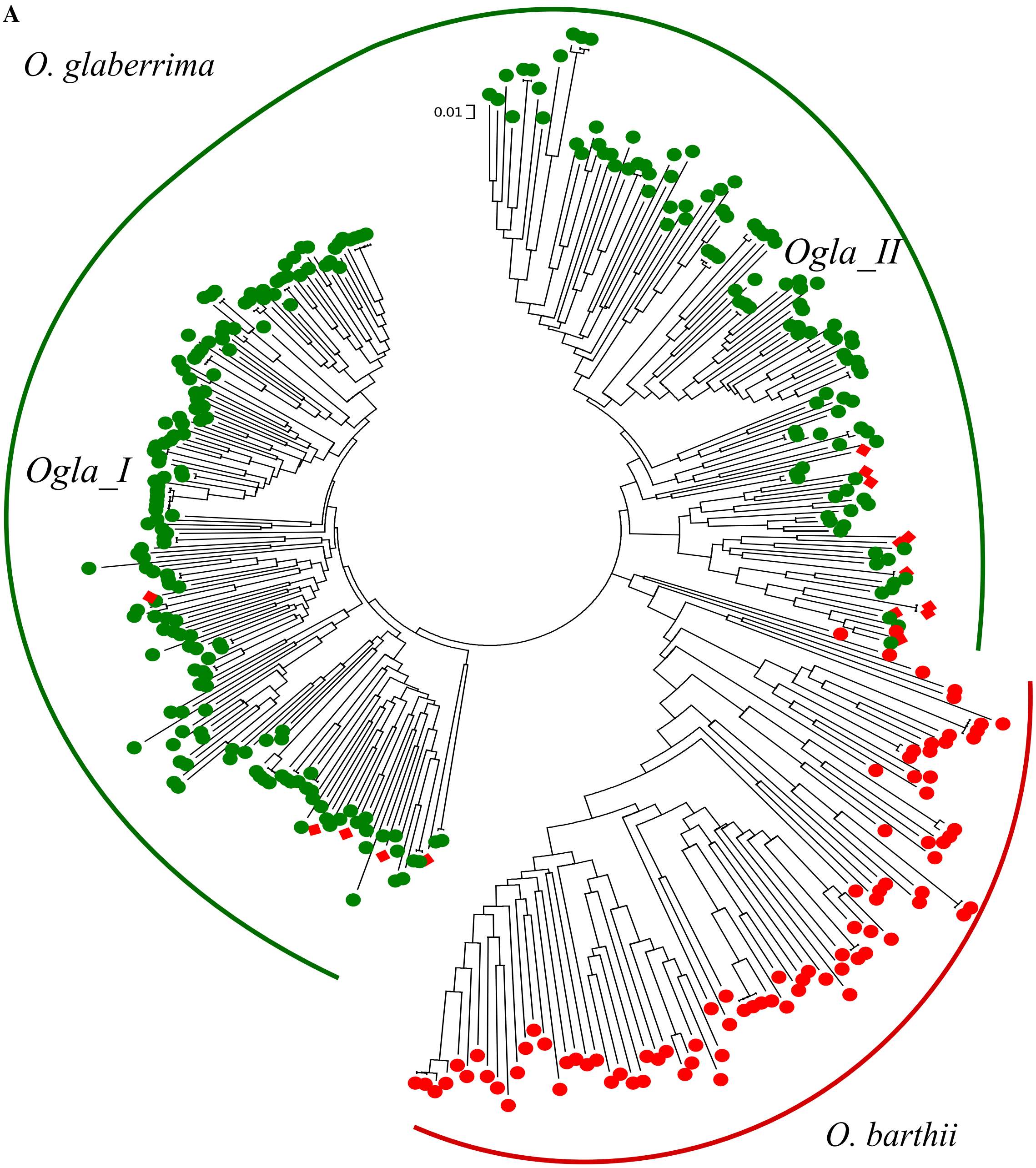


Fig. 5 **a** NJ tree of *O. glaberrima* (*green*) and *O. barthii* (*red*). **b** PCA for the same *O. glaberrima* (*green*) and *O. barthii* (*red*). The two analyses are based on the 235 SNP sets (color figure online)

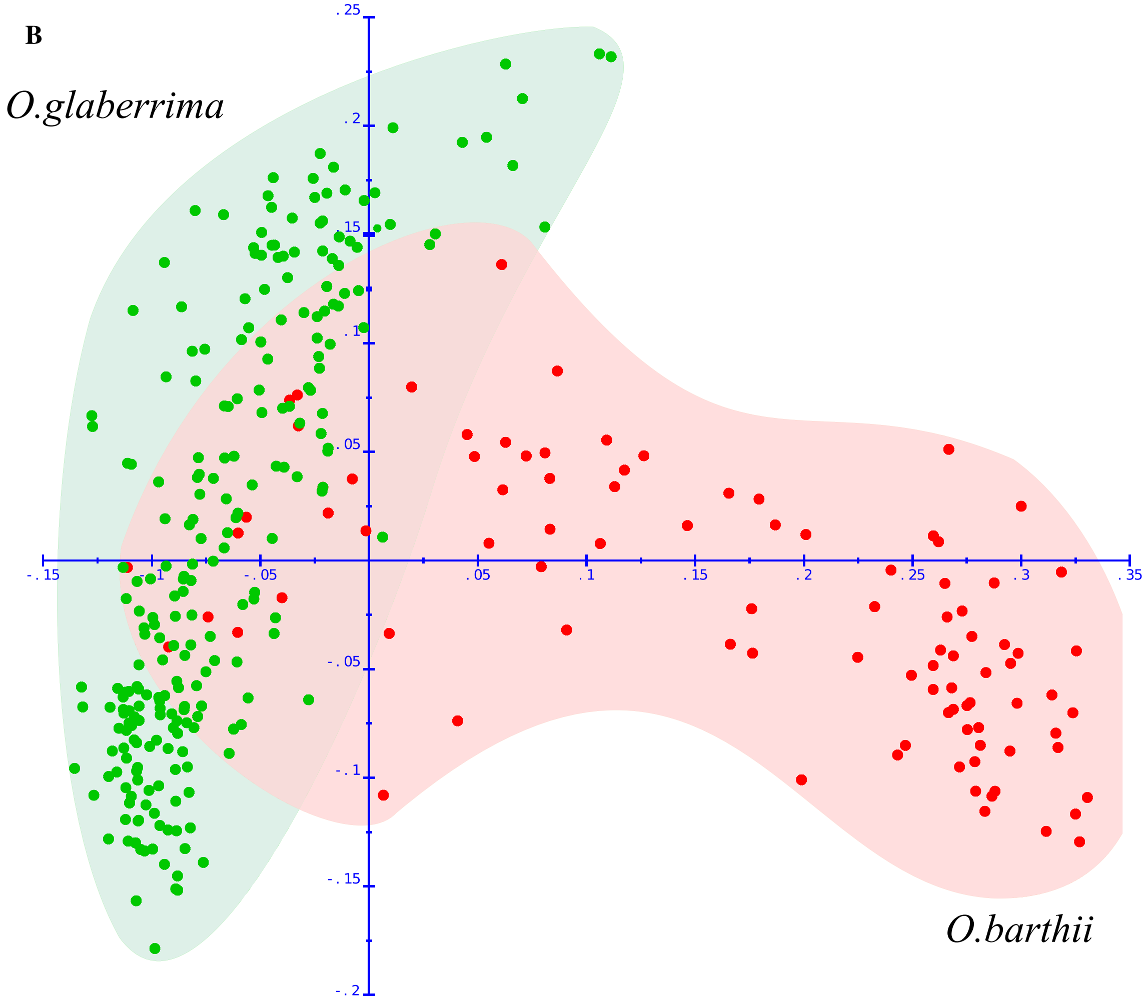


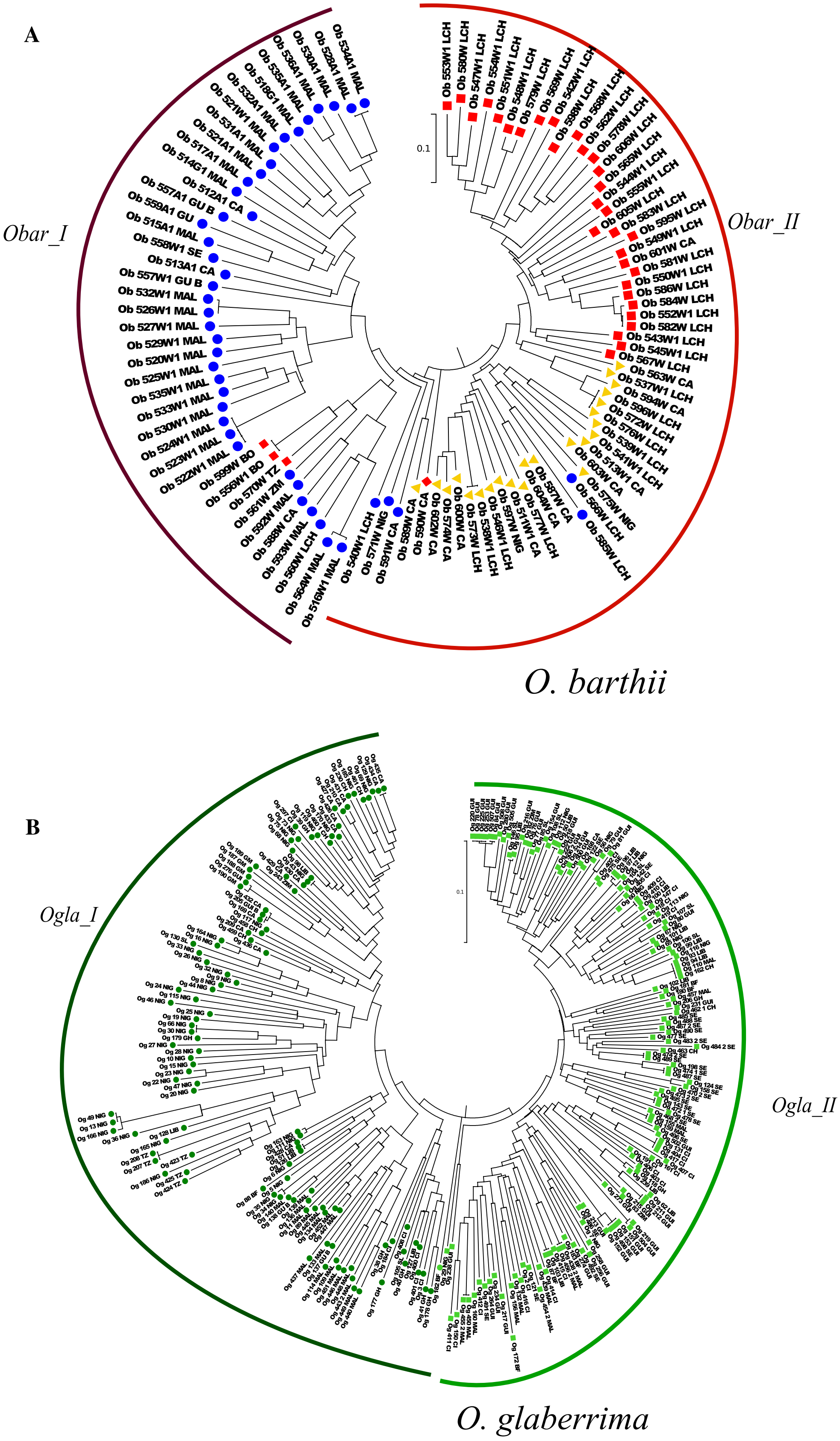
Fig. 5 continued

efficient bioinformatic methodology based on affordable sequencing cost, we were able to identify enough SNP loci for our specific purpose, leading the way for the one-shot chip design at low cost. The way we selected the SNPs lead itself to a biased estimation of the Asian Rice variability, rending it lower than the true level. However, as we focused on African rices, the current biases are not relevant.

As expected, we were able to easily separate the three genetic pools (*O. sativa*, *O. glaberrima* and *O. barthii*) using this set of markers. The 47 polymorphic markers within Asian and within African rice plants (*Oryza* variants) were even enough powerful in the highly variant Asian rice to separate the five already known groups of *sativa* (Glaszmann, 1988; Garris et al. 2005). These markers should represent ancient polymorphism shared between the Asian and African branches of the AA genomes. They are not specific for Asian vs African, however, as they are polymorphic in the two pools. At the opposite, the remaining 188 markers are monomorphic in Asian only, and polymorphic in the two African species. These markers are fixed either in the whole *O. sativa* species or only in the subset of Asian rice varieties we tested. Nevertheless, none of the 47 or 188 SNPs can be used to individually separate Asian and African species, as they are all polymorphic within the African species. However, used in global, they can help us to separate the two groups: Asian and African.

Previous studies (Semon et al. 2005; Barry et al. 2007; Nuijten et al. 2009) reported natural hybrids between *O. sativa* and *O. glaberrima*; although all the artificial hybrids were clearly located between Asian and African areas, these reported natural hybrids did not. In our assay, these hybrids are tightly linked to the O. sativa control varieties, close to group III plants (Garris et al. 2005). After more phenotypic and genetic controls, it appeared that they are not interspecific hybrids, but rather true *O. sativa* varieties. They do not clearly fit in as *O. sativa ssp indica* or *O. sativa ssp japonica* varieties, and we hypothesized that they were probably introduced a long time ago in Africa, through Western Asia.

Fig. 6 **a** NJ tree for the 101 accessions of *O. barthii*. The *dark red line* represents the *Obar_I* group, including 40 varieties; the *light red one* represents the *Obar_II* group, containing 61 accessions. The *blue dot*, *yellow triangle* and *red square* correspond to the three genetic groups found by the Evanno method as the second maximal Δ$K = 3$ (Supplemental Fig. 2). **b** NJ tree for the 266 *O. glaberrima* accessions. The *dark green curve* represents the *Ogla_I* group, including 111 varieties; the *light green curve* represents the *Ogla_II* accessions containing 155 accessions (color figure online)



However, because of the medium resolution (1 locus per 1.6 M on average, higher than any of the previous studies) and the type of polymorphism of our markers (SNP), small introgressions resulting from old hybridization might be missed. Such strong ancient, complex hybridization events result in very small alien introgressions and are by nature difficult to detect with medium-coverage biallelic marker sets.

Moreover, bias created by markers' design strongly impacts the genetic level of detail one can have on intra-specific variation. Previous studies' (Semon et al. 2005; Barry et al. 2007) experiments used, respectively 93 and 11 specific indica/japonica polymorphic SSR markers to study *O. glaberrima*. Consequently, a weak resolution of the diversity of the *O. glaberrima/O. barthii* group was achieved. In our study, we use 235 markers specifically designed for the study of the *O. glaberrima/O. barthii* group, which led to a better intra-specific resolution of the diversity for this specific group. It is noteworthy that analyses using DNA sequence do not have the same bias problem (Li et al. 2011), as well as studies using specifically chosen molecular markers (Dramé et al. 2011). In the same way, our choice of SNP polymorphic within the two African rice species leads to a reduction of the observed Asian Rice diversity, as only 47 SNPs upon 235 are polymorphic within this species.

As for the diversity of the African rices only, we found some intermediate plants between the cultivated and wild compartment. Weedy *O. barthii* could result from interspecific crosses between *O. barthii* and *O. glaberrima*, known these two species are inter-fertile. These weedy accessions were collected in Delta Niger River (mainly Cameroon and Mali) and present a genetic mix between the wild and the cultivated species. The presence of wild *O. barthii* in the vicinity of *O. glaberrima* fields probably favored gene flows, in the same way as described for *O. sativa* and *O. rufipogon* (Zhu et al. 2007; He et al. 2011). We also found that 14 of the 20 *O. barthii* tested in Li et al. (2011) are detected as intermediate (maybe weedy) type in our analysis (individuals bar_ZAM1, bar_TAN, bar_CAM1, bar_MAL1, bar_SIE2, bar_SIE1, bar_MAU, bar_NIG, bar_GUI2, bar_GUI1, bar_BUR, bar_MAL2, bar_SEN, bar_GAM1). The presence of these intermediate individuals might hamper inferences about domestication. However, even with these weedy forms, the F_{ST} value between wild and cultivated species indicates a clear genetic differentiation between the two genetic pools.

The structure of *O. barthii* shows a clear geographical pattern, with two main diversity regions are identified: the Chad Lake and the Delta Niger River (plus Austral and East African samples), with an additional minor one covering Central Africa. Few admixtures are observed in our collection between the three groups. We also found a high level of inbreeding in this species, as in Nabholz et al. (2014), as well as limited gene flows between these three groups, and a higher general diversity in the wild species compared to the cultivated one, as expected. Such radiation could have different origins, to be investigated: ecological or geographical barriers, human spreading, sampling missing in some regions, etc.

In *O. glaberrima*, we could not decipher a clear geographical pattern, but we detect two genetically separated groups: *Oglab_I* and *Oglab_II* (Fig. 6b). No correlation between these two groups and any phenotypic feature assayed so far (grain shape, pericarp color, panicle structure, etc.; data not shown) could be established. Nevertheless, we can observe that the two NERICA parents, TOG5681 and CG14, belong to two different groups: *Oglab_I* and *Oglab_II*, respectively. As they represent two extreme different ecotypes, irrigated and upland, these two groups might harbor some particular ecological adaptation found through the geographic distribution of *O. glaberrima*. To reach more precision, the use of higher density marker set of complete genome sequencing might be necessary, as well as more phenotyping. Such approaches could certainly refine the analysis presented here and would also allow genome-wide association studies.

The knowledge generated by our present results and the following genomic investigations will open the way to an efficient use of *O. glaberrima* and *O. barthii* in Asian and African rice breeding programs. Indeed, in regard of the lower variability of *O. glaberrima* compared to *O. barthii*, the wild relative of the African rice might be worth to be investigated for the Asian rice improvement. It is thus of high importance to more deeply analyze the diversity of *O. barthii*, to fully unravel the potentiality of African rice species.

Data archiving All sequences data used to generate SNP are available through SRA BioProjects PRJNA227059 for *Oryza glaberrima* and PRJNA227060 for *Oryza barthii*. They are also available through the IRD Oryza website, http://oryza.mpl.ird.fr/data.

Author contributions JO and FS performed the analyses; SC, HA and HC managed the original collection; KS multiplied the collection and provided seeds; YV and ML performed statistical tests and validated the analyses; FS and AG managed the whole project.

Acknowledgments Authors want to thank Dr MN Ndjiondjop from AfricaRice Center (Cotonou, Benin) for her help in the African rice accessions choice, Dr Laurence Albar and Valérie Poncet for useful discussions, Mme Christine Tranchant for her informatics support, and Dr Timothy Tranbarger for his help in English corrections. All the analyses were performed on the Bioinformatics platform from IRD (http://bioinfo.ird.fr). This project was supported by a Grant from Agropolis Foundation (ArCad program) and from GRiSP project (MENERGEP NewFrontier program).

Conflict of interest The authors declare no conflict of interest.

References

Albar L, Ndjiondjop M-N, Esshak Z, Berger A, Pinel A, Jones M et al (2003) Fine genetic mapping of a gene required for Rice yellow mottle virus cell-to-cell movement. Theor Appl Genet 107:371–378

Barry MB, Pham JL, Noyer JL, Billot C, Courtois B, Ahmadi N (2007) Genetic diversity of the two cultivated rice species (*O. sativa* & *O. glaberrima*) in Maritime Guinea. Evidence for interspecific recombination. Euphytica 154:127–137

Bezançon G (1993) Le riz cultivé d'origine Africaine *Oryza glaberrima* Steud, et les formes sauvages et adventices apparentées : diversité, relations génétiques et domestication

Billot C, Droc G, Courtois B, Farouk A, Ahmadi N, Clément G et al (2007) *HaplOryza*—SNP analysis of the genetic diversity along the rice genome. http://www.generationcp.org/communications/programme-publications/project-briefs/doc_download/78-2006-executive-summaries

Dramé KN, Sanchez I, Gregorio G, Ndjiondjop M-N (2011) Suitability of a selected set of simple sequence repeats (SSR) markers for multiplexing and rapid molecular characterization of African rice (*Oryza glaberrima* Steud.). Afr J Biotechnol 10:6675–6685

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software *STRUCTURE*: a simulation study. Mol Ecol 14:2611–2620

Excoffier L, Laval G, Schneider S (2005) *Arlequin* (version 3.0): an integrated software package for population genetics data analysis. Evolut Bioinform Online 1:47–50

Garavito A, Guyot R, Lozano J, Gavory F, Samain S, Panaud O et al (2010) A genetic model for the female sterility barrier between Asian and African cultivated rice species. Genetics 185:1425–1440

Garris AJ, Tai TH, Coburn J, Kresovich S, McCouch S (2005) Genetic structure and diversity in *Oryza sativa* L. Genetics 169:1631–1638

Glaszmann JC (1988) Geographic pattern of variation among Asian native rice cultivars (*Oryza sativa* L.) based on fifteen isozyme loci. Genome 30(5):782–792

Gridley HE, Jones MP, Wopereis-Pura M (2002) Development of new rice for Africa (NERICA) and participatory varietal selection. In: Breeding rainfed rice for drought-prone environments: integrating conventional and participatory plant breeding in South and Southeast Asia: proceedings of a DFID Plant Sciences Research Programme/IRRI Conference, pp 12–15

Gutiérrez AG, Carabalí SJ, Giraldo OX, Martínez CP, Correa F, Prado G et al (2010) Identification of a Rice stripe necrosis virus resistance locus and yield component QTLs using *Oryza sativa* × *O. glaberrima* introgression lines. BMC Plant Biol 10:6

He Z, Zhai W, Wen H, Tang T, Wang Y, Lu X et al (2011) Two evolutionary histories in the genome of rice: the roles of domestication genes. PLoS Genet 7:e1002100

Huang P, Molina J, Flowers JM, Rubinstein S, Jackson SA, Purugganan MD et al (2012) Phylogeography of Asian wild rice, *Oryza rufipogon*: a genome-wide view. Mol Ecol 21:4593–4604

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

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N et al (2009) The sequence alignment/Map format and *SAMtools*. Bioinformatics 25:2078–2079

Li Z-M, Zheng X-M, Ge S (2011) Genetic diversity and domestication history of African rice (*Oryza glaberrima*) as inferred from multiple gene sequences. Theor Appl Genet 123:21–31

Linares OF (2002) African rice (*Oryza glaberrima*): history and future potential. Proc Natl Acad Sci 99:16360–16365

Liu Q, Guo Y, Li J, Long J, Zhang B, Shyr Y (2012) Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data. BMC Genom 13:S8

Ma J, Bennetzen JL (2004) Rapid recent growth and divergence of rice nuclear genomes. Proc Natl Acad Sci USA 101:12404–12410

Martin M (2011) *Cutadapt* removes adapter sequences from high-throughput sequencing reads. EMBnet J 17:10–12

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A et al (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303

Milne I, Shaw P, Stephen G, Bayer M, Cardle L, Thomas WTB et al (2010) *Flapjack*–graphical genotype visualization. Bioinformatics 26:3133–3134

Murray SS (2004) Searching for the origins of African rice domestication. Antiquity 78

Nabholz B, Sarah G, Sabot F, Ruiz M, Adam H, Nidelet S, Ghesquière A, Santoni S, David J, Glemin S (2014) Transcriptome population genomics reveals severe bottleneck and domestication cost in the African rice (*O. glaberrima*). Mol Ecol. doi:10.1111/mec.12738

Nuijten E, van Treuren R, Struik PC, Mokuwa A, Okry F, Teeken B et al (2009) Evidence for the emergence of new rice types of interspecific hybrid origin in West African farmers' fields. PLoS ONE 4:e7335

Orjuela J, Thiémélé D, Kolade O, Chéron S, Ghesquière A, Albar L (2013) A recessive resistance to Rice yellow mottle virus is associated with a rice homolog of the CPR5 gene, a regulator of active defence mechanisms. Mol Plant Microbe Interact 26(12):1455–1463

Perrier X, Flori A, Bonnot F (2003) Data analysis methods. In: Genetic diversity of cultivated tropical plants, pp 43–76

Portères R (1962) Primary cradles of agriculture in the African continent. J Afr Hist 3:195–210

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Rousset F, Raymond M (1995) Testing heterozygote excess and deficiency. Genetics 140:1413–1419

Sano Y (1990) The genic nature of gamete eliminator in rice. Genetics 125:183–191

Second G (1982) Origin of the genic diversity of cultivated rice (*Oryza* spp.): study of the polymorphism scored at 40 isozyme loci. Jpn J Genet 57:25–57

Semon M, Nielsen R, Jones MP, McCouch SR (2005) The population structure of African cultivated rice *Oryza glaberrima* (Steud.): evidence for elevated levels of linkage disequilibrium caused by admixture with *O. sativa* and ecological adaptation. Genetics 169:1639–1647

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) *MEGA5*: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731–2739

Thiémélé D, Boisnard A, Ndjiondjop M-N, Chéron S, Séré Y, Aké S et al (2010) Identification of a second major resistance gene to Rice yellow mottle virus, RYMV2, in the African cultivated rice species, *O. glaberrima*. Theor Appl Genet 121:169–179

Thomson MJ, Zhao K, Wright M, McNally KL, Rey J, Tung C-W et al (2011) High-throughput single nucleotide polymorphism genotyping for breeding applications in rice using the *BeadXpress* platform. Mol Breed 29:875–886

Vaughan DA, Kadowaki K, Kaga A, Tomooka N (2005) On the phylogeny and biogeography of the genus oryza. Breed Sci 55:113–122

Vaughan DA, Lu B-R, Tomooka N (2008) The evolving story of rice evolution. Plant Sci 174:394–408

Zhao K, Tung C-W, Eizenga GC, Wright MH, Ali ML, Price AH et al (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. Nat Commun 2:467

Zhu Q, Zheng X, Luo J, Gaut BS, Ge S (2007) Multilocus analysis of nucleotide variation of *Oryza sativa* and its wild relatives: severe bottleneck during domestication of rice. Mol Biol Evol 24:875–888