nature genetics

Domestication history and geographical adaptation inferred from a SNP map of African rice

Rachel S Meyer[1,2], Jae Young Choi[1], Michelle Sanches[1], Anne Plessis[1], Jonathan M Flowers[1,2], Junrey Amas[3], Katherine Dorph[1], Annie Barretto[3], Briana Gross[4], Dorian Q Fuller[5], Isaac Kofi Bimpong[6], Marie-Noelle Ndjiondjop[7], Khaled M Hazzouri[2], Glenn B Gregorio[3] & Michael D Purugganan[1,2]

African rice (*Oryza glaberrima* Steud.) is a cereal crop species closely related to Asian rice (*Oryza sativa* L.) but was independently domesticated in West Africa ~3,000 years ago[1–3]. African rice is rarely grown outside sub-Saharan Africa but is of global interest because of its tolerance to abiotic stresses[4,5]. Here we describe a map of 2.32 million SNPs of African rice from whole-genome resequencing of 93 landraces. Population genomic analysis shows a population bottleneck in this species that began ~13,000–15,000 years ago with effective population size reaching its minimum value ~3,500 years ago, suggesting a protracted period of population size reduction likely commencing with predomestication management and/or cultivation. Genome-wide association studies (GWAS) for six salt tolerance traits identify 11 significant loci, 4 of which are within ~300 kb of genomic regions that possess signatures of positive selection, suggesting adaptive geographical divergence for salt tolerance in this species.

We used paired-end (2 × 100-bp) Illumina sequencing to resequence the genomes of 93 traditional *O. glaberrima* landraces from across the species range in West and Central sub-Saharan Africa (**Fig. 1a** and **Supplementary Table 1**). Most samples originated from a coastal region spanning Senegal to Liberia, as well as from inland areas in Nigeria, Niger, Cameroon, Chad, Mali and Burkina Faso. Four landraces were sequenced deeply (~30–73× mean nuclear genome coverage depth), and the remaining accessions were sequenced to an average depth of ~14.61× (with <4% missing genotype calls; **Supplementary Table 1**). This yielded 381 Gb of mappable sequence when aligned to the *O. glaberrima* CG14 reference genome sequence[3].

After the application of quality control filters, we identified 2,317,937 SNPs, or approximately 7.32 SNPs/kb, in African rice (**Fig. 1b**,**c** and **Supplementary Figs. 1–3**). We estimated nucleotide diversity (π) to be 0.0034 ± 0.0032 (s.d.), comparable to previous estimates using genome-wide data[3]. We validated genotype calls by Sanger sequencing of 51 SNPs and found >93% accuracy (**Supplementary Table 2**). We observed 905,654 SNPs (39%) in genic regions, including 51,296 synonymous, 54,833 nonsynonymous, 110,390 intronic, 19,860 upstream and 667,237 downstream SNPs. There were 1,105 nonsense mutations that truncated the encoded proteins. Linkage disequilibrium (LD) was substantial, with r^2 reaching half its maximum value at ~175 kb and approaching baseline at ~300 kb (**Fig. 1d**).

Principal-component analysis (PCA) of SNP variation using EIGENSTRAT identified ten significant components ($P < 0.0001$)[6], with the top three components each explaining <4% of the variance (**Fig. 2a** and **Supplementary Table 3**). The top two eigenvectors were strongly correlated with geography: PC1 was correlated with an east–west cline ($r = -0.77$, $P < 2.2 \times 10^{-16}$) and PC2 was correlated with a north–south cline ($r = -0.57$, $P < 1.5 \times 10^{-9}$) (**Fig. 2b**)[7]. To generate an alternative view of population stratification, we used the population clustering program STRUCTURE, which inferred the optimal number of genetic clusters comprising the *O. glaberrima* landrace genomes to be $K = 6$ (**Fig. 2c**; for other K values, see **Supplementary Fig. 4**)[8]. We found that all the landraces predominantly belonged to one cluster, with various levels of genomic contribution from five other ancestral populations.

To examine spatial genetic variation, we chose 11° N latitude to divide the arid north from the tropical south and 6° W longitude to separate the western Atlantic coast from eastern inland areas (**Figs. 1a** and **2b**). These divisions are consistent with the clines observed in PCA and define northwest (NW) and southwest (SW) coastal as well as northeast (NE) and southeast (SE) inland populations. The NE inland quadrant encompasses a hypothesized inland center of origin in the middle Niger River of Mali suggested by Portères[1,3,9,10]. The division between the NW and SW coastal regions also separates two proposed centers of secondary diversification[9,10]—a northern region centered on the Casamance in Senegal and a southern area in the Guinea Highlands between Sierra Leone and Liberia. For a review of proposed domestication and diversification centers, see the **Supplementary Note**.

We used TreeMix to examine the topology of relationships and migration history among populations[11]. Using the wild progenitor *Oryza barthii* A. Chev as an outgroup, we observed an older split

[1]Department of Biology, Center for Genomics and Systems Biology, New York University, New York, New York, USA. [2]Center for Genomics and Systems Biology, New York University Abu Dhabi, Abu Dhabi, UAE. [3]Plant Breeding, Genetics and Biotechnology Division, International Rice Research Institute, Los Baños, Philippines. [4]Department of Biology, University of Minnesota, Duluth, Minnesota, USA. [5]Institute of Archaeology, University College London, London, UK. [6]AfricaRice Sahel Station, Saint-Louis, Senegal. [7]AfricaRice Centre, Cotonou, Benin. Correspondence should be addressed to M.D.P. (mp132@nyu.edu).

Received 7 March; accepted 1 July; published online 8 August 2016; doi:10.1038/ng.3633

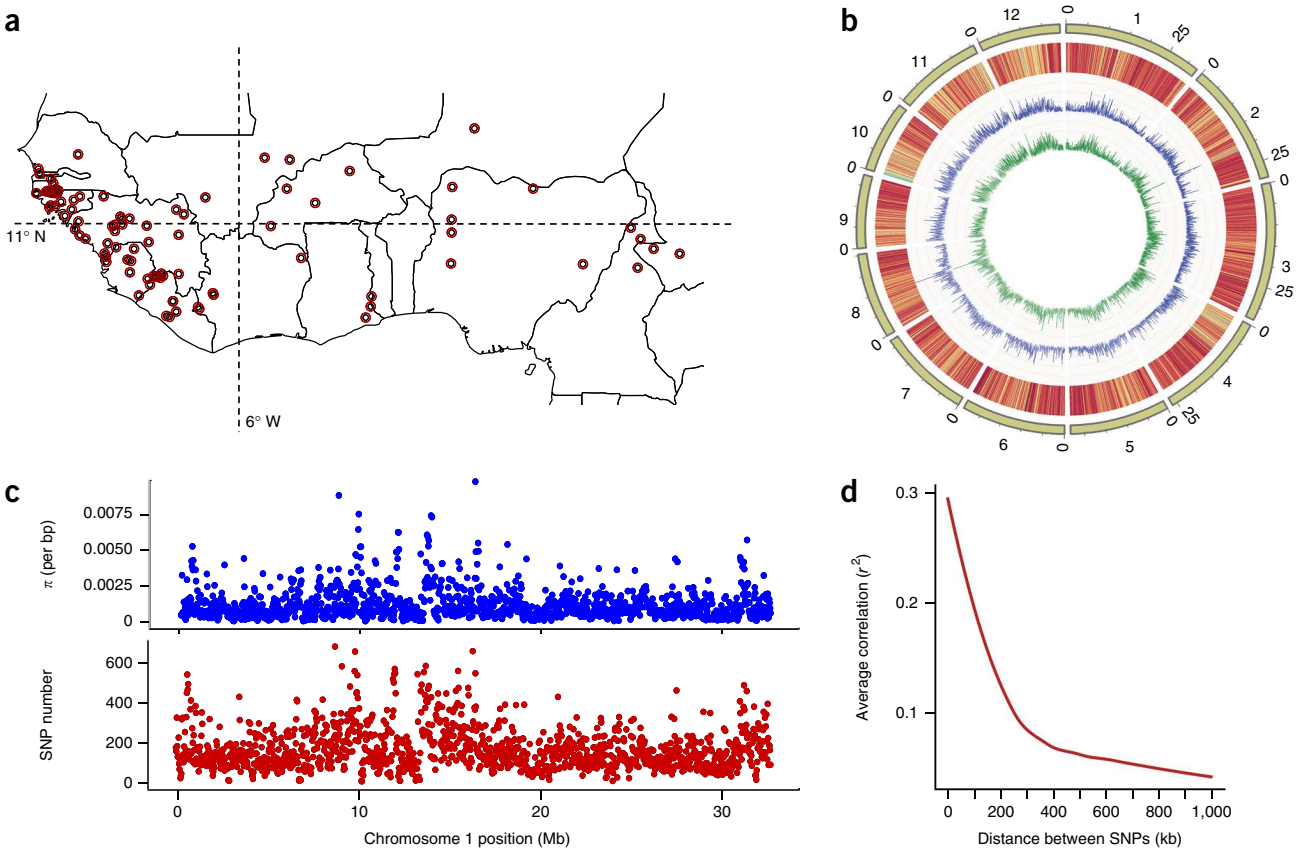

Figure 1 A SNP map for African rice. (**a**) Location of sampled landraces in West Africa, with the latitude (11° N) and longitude (6° W) lines demarcating the four geographical quadrants delimited in this study represented by dashed lines. The map was generated using ggmap[7]. (**b**) Circos plot showing SNP diversity across the 12 chromosomes of *O. glaberrima*. The chromosomes are numbered. The outer circle represents SNP density, the middle (blue) circle shows nucleotide diversity (π) and the inner (green) circle depicts the population mutation parameter (θ_W). (**c**) Close-up view of variation across chromosome 1, with the plots showing nucleotide diversity (top) and SNP number (bottom) in each 25-kb window. (**d**) Relationship of mean LD with genomic distance.

between coastal and inland populations and a more recent separation of northern and southern populations (**Fig. 3a**). Even without migration ($m = 0$), this topology accounted for >99% of the variance in SNP data. Greater model support was provided by inferred gene flow from the SW coastal population to the SE inland population ($m = 1$) (**Fig. 3a**), but this only marginally improved model fit.

Domestication is typically accompanied by population bottlenecks[12,13]. To examine whether this was the case in *O. glaberrima*, we applied a multiple sequentially Markovian coalescent model on two haplotypes (PSMC')[14]. Because *O. glaberrima* is a recently evolved, predominantly selfing species, we adopted a strategy of creating pseudodiploid genomes from data for two individuals from different populations, similar to what has been done in other inbreeding species such as *Caenorhabditis elegans*[15]. As lower sequencing depth is likely to bias this analysis by underestimating polymorphism levels, for each pseudodiploid genome, we used only genomes that were sequenced at >20× coverage and had comparable sequencing depth. Using this approach, we found that the PSMC' profiles indicated a reduction in effective population size (N_e) starting around 13,000–15,000 years ago from ~60,000 to a minimum of ~3,000 approximately 3,500 years ago (**Fig. 3b**). This severe bottleneck during the domestication of African rice[16] is similar to what has been observed in other annual crop species[12,13,17–19]. In contrast, no severe bottleneck was evident in wild *O. barthii* (**Fig. 3b**).

For *O. glaberrima*, the recent maximum N_e observed at ~15,000 years ago coincides with an increase in precipitation in West Africa after deglaciation leading into the start of the early Holocene African Humid Period (AHP)[20,21]. Recognizably domesticated African rice, however, does not appear in the archaeological record until ~2,800–2,400 years ago, from sites in the inland Niger River delta in Mali[22,23]. Interestingly, our PSMC' results indicate that the minimum plateau in N_e for African rice occurred close to the dates corresponding to the earliest archaeological evidence. The analysis suggests an early onset of the bottleneck, and the prolonged decrease in N_e may have resulted from various factors, including regional climate change or some other abiotic factor. One possibility, however, is that the bottleneck may have resulted from a protracted period of low-intensity cultivation and/or management before full domestication ~3,500 years ago, just after a peak in human population growth in western Africa occurring 4,000–5,000 years ago[21]. Archaeobotanical evidence for this protracted use is elusive, as remains in West Africa from before 5,000 years ago are extremely rare[24]. Ceramic finds from the Early Holocene, however, do suggest early consumption of grass grains in the region[25].

The genome-wide SNP map allows us not only to examine aspects of the domestication of *O. glaberrima* but also post-domestication spread and subsequent adaptation to local environments. One key trait likely associated with geographical adaptation in African rice

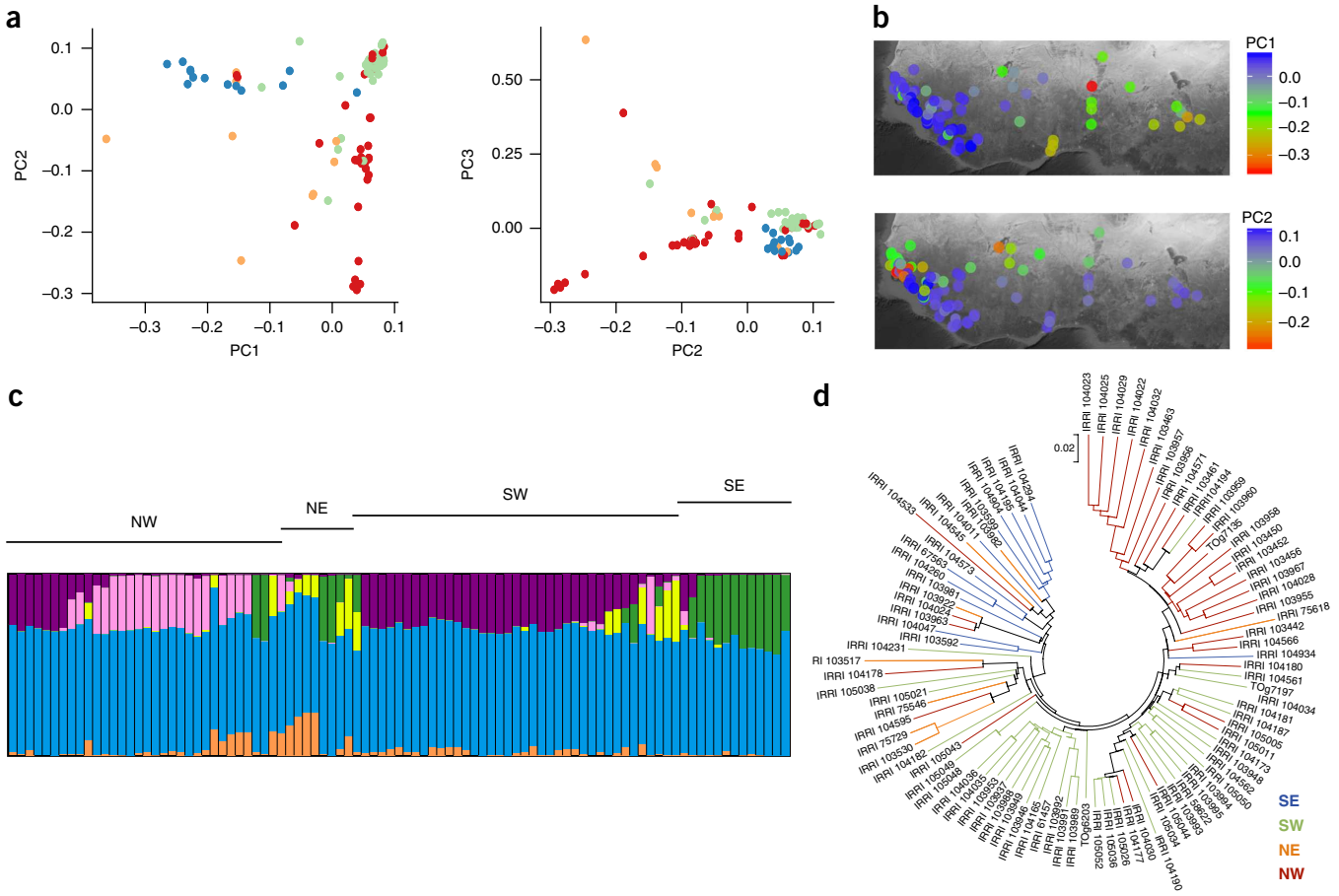

Figure 2 Population structuring in African rice. (**a**) Principal components of SNP variation. Samples from the NW coastal (red), SW coastal (green), NE inland (orange) and SE inland (blue) populations are shown. The plots show the first three principal components. (**b**) PC1 and PC2 scores for each *O. glaberrima* accession are shown, with the geographical location of the samples plotted on the map[7]. Visually, the east–west cline for PC1 and the north–south cline for PC2 are evident. (**c**) STRUCTURE plot for African rice, showing the distribution of the $K = 6$ genetic clusters. The four different West African populations are indicated. (**d**) Neighbor-joining clustering of landraces based on genetic distance. Branch color indicates membership in one of the four geographical populations. The scale bar shows substitutions per site.

is salinity tolerance[26,27]. Arid regions of northern West Africa have higher salinity levels, associated with saltwater intrusion into rivers that can reach up to 250 km inland[1]. We interviewed African farmers in inland Togo and coastal Senegal (high salinity) about the mitigation of salt stress (**Supplementary Table 4**). In Senegal, although efforts are made to control soil salinity, which affects most plant developmental stages, the major strategy was to farm salt-tolerant varieties: ~25% of the *O. glaberrima* varieties used by farmers in this area were reported to be salt tolerant.

To examine phenotypic variation in salt tolerance, we measured several salinity-associated fitness traits in 121 landrace seedlings at early and late stages of salt exposure. These traits included visual symptoms of plant salt stress described using the standard evaluation system (SES) score[28], the percentage of shoots with injury, mortality, and leaf Na^+ and K^+ content (**Supplementary Table 5**). There were significant differences in the phenotypes across the four West African populations, except for the phenotypes corresponding to Na^+ and K^+ content (**Fig. 4a**, **Supplementary Fig. 5** and **Supplementary Table 6**); for example, Kruskal–Wallis tests in experiments of late exposure to salinity for SES ($P < 1.96 \times 10^{-7}$), percentage of shoots with injury ($P < 8 \times 10^{-6}$), mortality ($P < 1.46 \times 10^{-7}$) and fourth-leaf Na^+ concentration ($P < 2.13 \times 10^{-8}$) indicated significant differences.

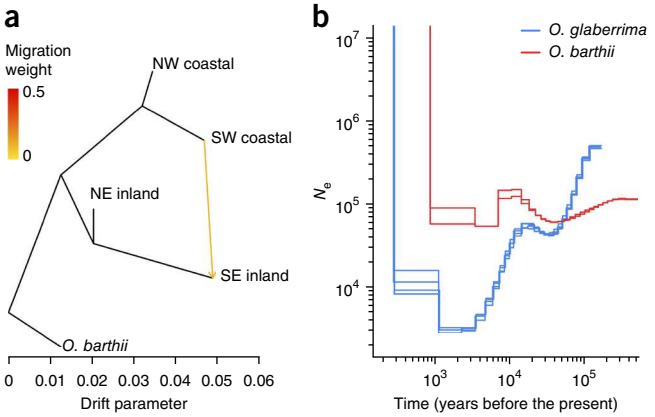

Figure 3 Demography of *O. glaberrima* and *O. barthii*. (**a**) TreeMix analysis of 93 *O. glaberrima* samples divided into four West African geographical quadrants, with *O. barthii* samples serving as the outgroup population. The arrow corresponds to the direction of migration. (**b**) PSMC'-inferred demographic history of *O. glaberrima* and *O. barthii*. Each line represents the past effective population size (N_e) for a pseudodiploid genome generated by combining haploid sequences. The blue line represents a coastal–inland *O. glaberrima* pseudodiploid combination, and the red line represents an intraspecific *O. barthii* combination.

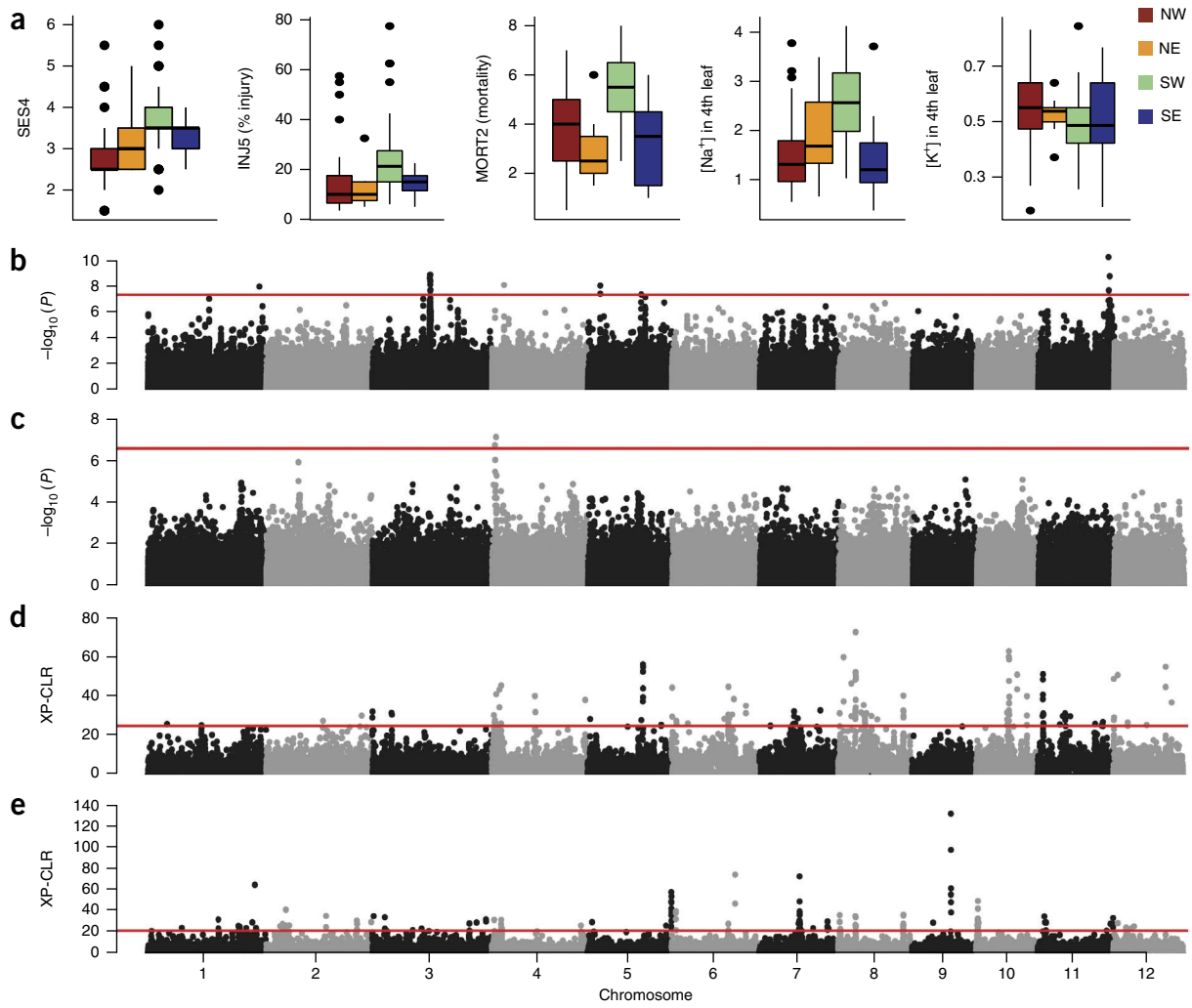

Figure 4 Population phenotypic differentiation, GWAS mapping and selective sweep analysis for salinity tolerance. (**a**) Geographical variation in salt tolerance phenotypes. Bar plots representing the four West African populations (SE inland ($n = 13$), SW coastal ($n = 50$), NE inland ($n = 9$), NW coastal ($n = 49$)) display highest and lowest quartiles around the median (box) and 1.5 times this interquartile range (whiskers). The line in each box is the mean, and the dots beyond the whiskers are outlier values. The specific traits are described in **Supplementary Table 5**. SES4, standard evaluation system score; INJ5, percent shoot injury; MORT2, mortality count in late salinity tests. Na^+ and K^+ concentrations in the fourth leaf are also shown. In all but one trait (K^+ concentration in the fourth leaf), the SW coastal population shows a significant difference in comparison to the other populations. (**b**,**c**) GWAS Manhattan plots for percent leaf injury in late salinity testing (INJ5) using the unpruned SNP set with a linear model (**b**) and for SES score in late salinity testing (SES4) using the LD-pruned SNP set with a linear model (**c**). The red line corresponds to the Bonferroni-adjusted significance threshold. (**d**,**e**) Genome-wide distribution of XP-CLR values using the NW coastal population as the reference and the SW costal population as the object population (**d**) and using the SW coastal population as the reference and the NW coastal population as the object population (**e**). The red line corresponds to the 99.5th-percentile XP-CLR value.

Pairwise population comparisons showed that these differences were driven by reduced salt tolerance in the SW coastal population (**Supplementary Table 7**). Contrasting populations in tests of late salinity, for example, identified Bonferroni-adjusted significant differences ($P < 0.0084$) between the SW coastal population and its sister NW coastal population, where the SW coastal population had higher SES score, percentage of injured shoots and incidence of mortality. There were no significant differences between the NW coastal, NE inland and SE inland populations for these traits (**Supplementary Table 7**). Loss of salinity tolerance in the SW coastal population is possibly associated with the costs of maintaining tolerance in a region of greater rainfall and reduced soil salinity (**Supplementary Fig. 6**)[1,29].

We conducted GWAS mapping with the 93 landraces whose genomes we resequenced[30], using the full set of 1,056,028 SNPs available after removing low-frequency SNPs (minor allele frequency <5%) and a further reduced set of 199,093 SNPs obtained by pruning LD-correlated SNPs ($r^2 > 0.5$). We performed linear[31] and mixed-model[32] association analyses, using the first ten principal components of population structure as covariates. We accepted models across these analyses with a genomic inflation factor (λ) of 1 ± 0.15 in quantile–quantile plots and used a conservative Bonferroni-adjusted threshold ($P < 2.5 \times 10^{-7}$ in the reduced set) to identify significant SNPs (**Fig. 4b**,**c** and **Supplementary Fig. 7**). Although the low sample size limited our power, we found 28 SNPs that exceeded the significance threshold in 11 unique genomic regions (**Supplementary Table 8**).

Seven loci were associated with the percentage of injured shoots, two were associated with SES score and two were associated with both traits. Further work will be necessary to identify specific genes underlying

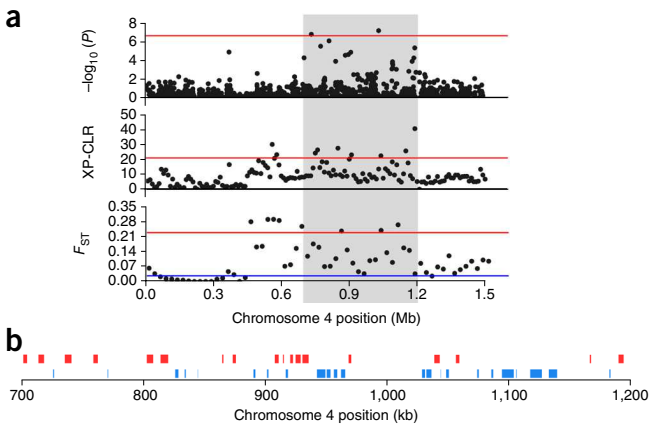

Figure 5 Comparison of GWAS and selected genomic regions at the proximal end of chromosome 4. (**a**) The top Manhattan plot shows GWAS results for the region from 0–1.5 Mb on chromosome 4. The GWAS is for percent leaf injury in late salinity testing using the unpruned SNP set with a linear model. The red line corresponds to the Bonferroni-adjusted significance threshold. The middle plot shows XP-CLR results using the SW coastal population as the reference and the NW coastal population as the object population. The 99.5th-percentile XP-CLR value is represented by the red line. The bottom plot is a sliding window analysis of F_{ST} values across the genomic region. The red line corresponds to the 99.5th percentile for genome-wide F_{ST} values. The blue line represents mean F_{ST} across the genome. The shaded region delimits a common ~500-kb genomic region with elevated GWAS probabilities, XP-CLR likelihoods and F_{ST} values in this genomic locus. (**b**) The positions of 41 gene models in this region are indicated: red, minus-strand genes; blue, positive-strand genes.

these quantitative trait loci (QTLs), although there were plausible candidate genes on chromosome 1. Two orthologs of the *O. sativa* high-affinity potassium transporters *OsHAK5* and *OsHAK6* are found ~75 kb and ~3.6 kb upstream, respectively, of the significant SNP in a shoot injury QTL. Overexpression of *OsHAK5* has been shown to improve salt tolerance[33,34], and qRT–PCR analysis confirmed that the *O. glaberrima* ortholog of *OsHAK5* is significantly upregulated under salt stress ($n = 8$; $P < 0.022$, *t* test). In contrast, the *OsHAK6* ortholog in *O. glaberrima* did not show evidence of upregulation under salt stress ($n = 8$; $P < 0.35$, *t* test).

To identify genomic regions associated with adaptive differentiation in the *O. glaberrima* populations, we used the cross-population composite likelihood ratio (XP-CLR) method to compare the NW coastal and SW coastal populations[35]. Using genomic regions with the top 0.5% of XP-CLR values, we identified 98 selected regions using either the SW coastal or NW coastal population as the test population, with 22 regions overlapping in the two tests (**Fig. 4d**,**e** and **Supplementary Table 9**). These genomic regions ranged in size from ~10 kb to ~760 kb.

We examined any overlap between the QTLs identified by GWAS and the putative selected regions identified by XP-CLR analyses. Two GWAS hits, on chromosomes 5 (14.75 Mb) and 11 (19.23 Mb), were found within 300 kb of a selected region identified by XP-CLR. The most promising region, however, encompassed two salt tolerance GWAS hits on the proximal end of chromosome 4 and overlapped with an inferred selected genomic region in the XP-CLR analysis (**Fig. 5a**). Furthermore, we constructed a genome-wide empirical distribution of the fixation index (F_{ST})[36] between the NW coastal and SW coastal populations and found that the genomic region encompassing these two GWAS loci had mean $F_{ST} = 0.157$. Several SNPs were in the upper 0.5% of the distribution (mean genome-wide $F_{ST} = 0.027$) (**Fig. 5a**, **Supplementary Fig. 8** and **Supplementary Table 10**), providing additional support for adaptive differentiation in this genomic region[37]. We found 41 genes in this area of overlap (**Fig. 5b**); one possible positional candidate gene, *PPI*, encodes a peptidylprolyl *cis*/*trans* isomerase and is a member of a gene family involved in stress response[38] whose members are known to confer seedling salt tolerance in Asian rice[39]. qRT–PCR analysis indicated that the *PPI* gene in this region is significantly upregulated under salt stress ($n = 8$; $P < 0.0017$, *t* test).

In summary, our analysis of African rice provides genetic evidence that may point to an extended period of low-intensity cultivation or management of a wild species before its domestication. There have been two competing hypotheses for the timescale of domesticated crop origins—the rapid and protracted transition models of domestication[40–42]. Our work provides support for the latter hypothesis, and, although other extrinsic (for example, climatic) factors cannot be ruled out, further archaeological and genetic work may help to establish the tempo and mode of the domestication process. Our study also identifies genomic regions associated with geographical differentiation and adaptation to a major abiotic stress factor, salinity, in the West African landscape, documenting crop evolutionary diversification accompanying species range expansion. The genome-wide polymorphism map in *O. glaberrima* presents key information on the evolutionary history of this recently evolved domesticate and offers new tools for mapping agriculturally important genes.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Sequence data have been deposited in the Nucleotide and Sequence Read Archive (SRA) databases. The Illumina raw sequence reads appear in SRA under accession SRP071857 and under BioProject PRJNA315063.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We would like to thank E. Septiningsih for critical discussions. We are grateful to M. Sock and B. Fonton for field assistance, to International Rice Research Institute staff for phenotyping assistance, and to J. Maritz and Z. Joly-Lopez for laboratory assistance. We thank the US Department of Agriculture and International Rice Research Institute for providing germplasm. This work was funded in part by grants from the National Science Foundation Plant Genome Research Program (IOS-1126971), the Zegar Family Foundation and the New York University Abu Dhabi Research Institute to M.D.P., as well as by a National Science Foundation Plant Genome Postdoctoral Fellowship (IOS-1202803) to R.S.M.

AUTHOR CONTRIBUTIONS

R.S.M., G.B.G. and M.D.P. designed the experiments and analyses. I.K.B. and M.-N.N. helped in design and execution of the fieldwork in Senegal and Togo, respectively. R.S.M., M.S., A.P., J.A., A.B., K.D., B.G. and G.B.G. collected the data. R.S.M., J.Y.C., J.M.F., K.M.H. and M.D.P. analyzed the data. R.S.M., J.Y.C., D.Q.F. and M.D.P. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Carney, J.A. *Black Rice: The African Origins of Rice Cultivation in the Americas* (Harvard University Press, 2002).
2. Linares, O.F. African rice (*Oryza glaberrima*): history and future potential. *Proc. Natl. Acad. Sci. USA* **99**, 16360–16365 (2002).

3. Wang, M. *et al.* The genome sequence of African rice (*Oryza glaberrima*) and evidence for independent domestication. *Nat. Genet.* **46**, 982–988 (2014).
4. Sarla, N. & Mallikarjuna, S.B.P. *Oryza glaberrima*: a source for improvement of *Oryza sativa*. *Curr. Sci.* **89**, 955–963 (2005).
5. Agnoun, A. *et al.* The African rice *Oryza glaberrima* Steud: knowledge distribution and prospects. *Int. J. Biol.* **4**, 158–180 (2012).
6. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
7. Kahle, D. & Wickham, H. ggmap: spatial visualization with ggplot2. *R J.* **5**, 144–161 (2013).
8. Falush, D., Stephens, M. & Pritchard, J.K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).
9. Portères, R. in *Papers in African Prehistory* (eds. Fage, J.D. & Oliver, R.A.) 43–58 (Cambridge University Press, 1970).
10. Portères, R. in *Origins of African Plant Domestication* (eds. Harlan, J.R., De Wet, J.M. & Stemler, A.B.) 409–452 (Mouton, 1976).
11. Pickrell, J.K. & Pritchard, J.K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967 (2012).
12. Tenaillon, M.I., U'Ren, J., Tenaillon, O. & Gaut, B.S. Selection versus demography: a multilocus investigation of the domestication process in maize. *Mol. Biol. Evol.* **21**, 1214–1225 (2004).
13. Caicedo, A.L. *et al.* Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS Genet.* **3**, 1745–1756 (2007).
14. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).
15. Thomas, C.G. *et al.* Full-genome evolutionary histories of selfing, splitting, and selection in *Caenorhabditis*. *Genome Res.* **25**, 667–678 (2015).
16. Nabholz, B. *et al.* Transcriptome population genomics reveals severe bottleneck and domestication cost in the African rice (*Oryza glaberrima*). *Mol. Ecol.* **23**, 2210–2227 (2014).
17. Eyre-Walker, A., Gaut, R.L., Hilton, H., Feldman, D.L. & Gaut, B.S. Investigation of the bottleneck leading to the domestication of maize. *Proc. Natl. Acad. Sci. USA* **95**, 4441–4446 (1998).
18. Hyten, D.L. *et al.* Impacts of genetic bottlenecks on soybean genome diversity. *Proc. Natl. Acad. Sci. USA* **103**, 16666–16671 (2006).
19. Zhu, Q., Zheng, X., Luo, J., Gaut, B.S. & Ge, S. Multilocus analysis of nucleotide variation of *Oryza sativa* and its wild relatives: severe bottleneck during domestication of rice. *Mol. Biol. Evol.* **24**, 875–888 (2007).
20. Tjallingii, R. *et al.* Coherent high- and low-latitude control of the northwest African hydrological balance. *Nat. Geosci.* **1**, 670–675 (2008).
21. Manning, K. & Timpson, A. The demographic response to Holocene climate change in the Sahara. *Quat. Sci. Rev.* **101**, 28–35 (2014).
22. Murray, S.S. in *Fields of Change: Progress in African Archaeobotany* (ed. Cappers, R.T.J.) 53–62 (Barkhuis, 2007).
23. Zach, B. & Klee, M. Four thousand years of plant exploitation in the Chad Basin of NE Nigeria. II: Discussion on the morphology of caryopses of domesticated *Pennisetum* and complete catalog of the fruits and seeds of Kursakata. *Veg. Hist. Archaeobot.* **12**, 187–204 (2003).
24. Fuller, D.Q., Nixon, S., Stevens, C.J. & Murray, M.A. in *The Archaeology of African Plant Use* (eds. Stevens, C., Nixon, S., Murray, M.A. & Fuller, D.Q.) 17–24 (Left Coast Press, 2014).
25. Eichhorn, B. & Neumann, K. in *Archaeology of African Plant Use* (eds. Stevens, C., Nixon, S., Murray, M.A. & Fuller, D.Q.) (Left Coast Press, 2014).
26. Temudo, M. Planting knowledge, harvesting agro-biodiversity: a case study of Southern Guinea-Bissau rice farming. *Hum. Ecol.* **39**, 301–321 (2011).
27. Carney, J.A. Landscapes of technology transfer: rice cultivation and African continuities. *Technol. Cult.* **37**, 5–35 (1996).
28. International Rice Research Institute. *Standard Evaluation System for Rice* (International Rice Research Institute, 2014).
29. Matlon, P., Randolph, T. & Guei, R. in *Impact of Rice Research* (eds. Pingali, P.B. & Hossain, M.) 382–404 (International Rice Research Institute, 1998).
30. Huang, X. *et al.* Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* **42**, 961–967 (2010).
31. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
32. Kang, H.M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
33. Yang, T. *et al.* The role of a potassium transporter OsHAK5 in potassium acquisition and transport from roots to shoots in rice at low potassium supply levels. *Plant Physiol.* **166**, 945–959 (2014).
34. Horie, T. *et al.* Rice sodium-insensitive potassium transporter, OsHAK5, confers increased salt tolerance in tobacco BY2 cells. *J. Biosci. Bioeng.* **111**, 346–356 (2011).
35. Chen, H., Patterson, N. & Reich, D. Population differentiation as a test for selective sweeps. *Genome Res.* **20**, 393–402 (2010).
36. Weir, B.S. & Cockerham, C.C. Estimating *F*-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370 (1984).
37. Lewontin, R.C. & Krakauer, J. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* **74**, 175–195 (1973).
38. Sahi, C., Singh, A., Blumwald, E. & Grover, A. Beyond osmolytes and transporters: novel plant salt-stress tolerance-related genes from transcriptional profiling data. *Physiol. Plant.* **127**, 1–9 (2006).
39. Ruan, S.L. *et al.* Proteomic identification of OsCYP2, a rice cyclophilin that confers salt tolerance in rice (*Oryza sativa* L.) seedlings when overexpressed. *BMC Plant Biol.* **11**, 34 (2011).
40. Allaby, R.G., Fuller, D.Q. & Brown, T.A. The genetic expectations of a protracted model for the origins of domesticated crops. *Proc. Natl. Acad. Sci. USA* **105**, 13982–13986 (2008).
41. Fuller, D.Q. Contrasting patterns in crop domestication and domestication rates: recent archaeobotanical insights from the Old World. *Ann. Bot.* **100**, 903–924 (2007).
42. Wilcox, G. in *Biodiversity in Agriculture: Domestication, Evolution, and Sustainability* (eds. Gepts, P. *et al.*) 92–109 (Cambridge University Press, 2012).

ONLINE METHODS

Sample collection and library preparation. *O. glaberrima* seeds were from the International Rice Research Institute (IRRI) and the US Department of Agriculture (USDA) (**Supplementary Table 1**). DNA was extracted from a single seedling leaf using either the DNeasy mini kit (Qiagen) or standard phenol:chloroform:isoamyl alcohol buffer. Libraries were prepared using Illumina TruSeq kits with an insert size of ~380 bp. 2 × 100-bp paired-end sequencing was carried out on an Illumina HiSeq 2500 instrument at the New York University Genome Core with 2–7 libraries per lane. 2 × 100-bp paired-end sequencing was also completed for three *O. barthii* barcoded TruSeq DNA libraries, which were prepared with an insert size of 350 bp and sequenced using half a lane on an Illumina HiSeq 2000 instrument at the University of Minnesota Genomics Center.

Read alignment and SNP calling. Sequencing reads passing Illumina's quality control filter were aligned, using the Burrows–Wheeler aligner (v0.6.1) to map reads to version 1.1 of the Arizona Genome Institute *O. glaberrima* genome[3], which includes the 12 pseudomolecules and 1,939 unassembled scaffolds (Genebuild 2011-05-AGI). Duplicate reads were removed from individual sample alignments using Picard tools (v1.111) MarkDuplicates, and reads were merged using MergeSamFiles. The Genome Analysis Toolkit (GATK; v3.1-1)[43,44] RealignerTargetCreator and IndelRealigner protocol was used for global realignment of reads around indels.

Before processing the final SNP map with 93 *O. glaberrima* samples, we had mapped 99 samples, assumed to be *O. glaberrima* on the basis of their labels, to the *O. sativa* Nipponbare genome (IRGSP 1.0), along with 57 diverse *O. sativa* accessions (I.S. Pires, J.M.F. and M.D.P., unpublished data). SNP calling and filtering (using the methods described below) produced >6 million SNPs that were used in STRUCTURE population clustering analyses to look for mislabeled or admixed samples. Six samples were found not to be true *O. glaberrima* accessions: IRGC lines 103587, 103602, 103961, 104037, 104254 and 106291. These were removed from the study.

SNP calling used the GATK UnifiedGenotyper set for diploids with default filtering settings, similarly to previous studies[45,46]. Base qualities were capped at the mapping quality of the read, and all reads mapping to two or more places were filtered out. Filtering for all SNPs by GATK used settings based on outlier transition/transversion ratios that are enriched in false positives. For the SNP set in 93 *O. glaberrima* samples, the filtering settings were as follows: DP > 10,000, FS > 212, MQ < 11, MQ0 > 5,000, MQRankSum < −46, QD < 0.16 and ReadPosRankSum < −15. This strategy reduced the raw unfiltered set of 2.88 million SNPs to the working set of 2.3 million SNPs (with scaffolds) and 2.14 million SNPs (pseudomolecules only).

The *O. barthii* accessions, including those sequenced as part of this project (IRGC lines 101226, 100941 and 104081) and seven DNA libraries from the public SRA (SRR1206365, SRR1206367, SRR1206368, SRR1206397, SRR1206405, SRR1206412 and SRR1206436), were mapped to the *O. glaberrima* genome. SNPs were called for this set of samples alone and were filtered before merging with the set of SNPs for the 93 *O. glaberrima* accessions. The filtering settings were as follows: DP > 1,500, FS > 74, MQ < 37, MQ0 > 5,000, MQRankSum < −5.50, QD < 0.4 and ReadPosRankSum < −4.

To validate SNPs, we tested 51 SNPs and 131 genotypes by Sanger sequencing of forward and reverse strands. Each region was amplified from at least one of 19 randomly chosen *O. glaberrima* DNA templates and sequenced (**Supplementary Table 2**). Contigs were assembled in Sequencher (Gene Codes) and trimmed of primer sequence. The corresponding coordinates of the trimmed sequences were used to query SNP predictions, and these sites were examined for inconsistency, such as signals of heterozygosity or non-equivalent base calls, in the Sanger sequencing chromatograms. We found that 9 of the 131 genotypes were different between the two methods, giving a 93.1% concordance rate. Eight of the nine were heterozygous in the VCF file but appeared to be homozygous in the Sanger sequencing data. One error consisted of a genotype that matched neither of the variants in the VCF file. Of the concordant assigned genotypes screened, 30 were homozygous for the alternative allele, 7 were heterozygous and 85 were homozygous for the reference allele.

SNP annotation. SnpEff (v3.6c)[47] was used to assign SNP effects on the basis of gene models from AGI v0.1.1 2012 annotation (still current as of February 2016). The SNPs for codons with multiple SNPs were annotated separately, and only canonical transcripts were used. Classification of SNP effects was dependent on the current gene models available; however, given the discrepancies in annotation quality for *O. glaberrima* and *O. sativa*, it is clear that improvement of annotated gene models will change the counts of effects in the SNP set.

Population genetic parameters. The program ANGSD (v0.613)[48] was used to calculate the population genetic statistics θ_W, π and Tajima's D directly from sample BAM files in 25-kb non-overlapping intervals; this was done for the whole sample set and for populations assigned by geographical quadrant.

The same set of SNPs that was used for TreeMix analysis was also used to estimate genome-wide LD. LD was calculated in PLINK (v1.90)[31] using SNP pairs in a 1,000-kb window and at most 99,999 SNPs apart. Genome-wide LD decay was calculated by grouping SNP pairs into 1-kb bins and averaging the squared correlation coefficient (r^2) within bins. Average bin r^2 values were plotted for each chromosome with a line of best fit using LOESS curve fitting.

Population structure analysis and genetic distance relationships. STRUCTURE (version 2.3.4)[8] was run using a reduced SNP set from pseudomolecules only, where a random 4% of the SNPs were retained and subsequently LD pruned in PLINK (v1.90)[31] using the settings -indep 50 5 1.5, leaving 29,983 SNPs. Using settings for admixture and no linkage, STRUCTURE was run with a burn-in of 50,000 replicates and 50,000 Markov chain Monte Carlo (MCMC) iterations following the burn-in step. This was repeated ten times for each K value (K = 1 to 8). Results were analyzed using the EVANNO method with STRUCTURE HARVESTER[49], and CLUMPP (v. 1.1.2)[50] was used to permute run clusters. DISTRUCT[51] was used to plot the results for K = 3 through K = 6 (**Supplementary Fig. 4**). $\Delta(K)$ values indicated that K = 6 was optimal.

PCA was performed using the EIGENSOFT package to run EIGENSTRAT[6] on the LD-pruned pseudomolecule SNP set of 570,728 SNPs. The top ten principal components were used in geographical analysis and in downstream genome-wide association mapping. Results were plotted on geographical maps using ggmap[7].

A neighbor-joining tree was constructed using the filtered SNP set from 93 *O. glaberrima* accessions, with distances calculated using the Gronau method[52], described in Hazzouri *et al.*[46]. Tree construction from the distance matrix used MEGA (v5.2)[53].

Admixture among the *O. glaberrima* accessions from different geographical quadrants was modeled in TreeMix v.1.12 (ref. 11) (**Fig. 1**). Segregating SNPs across the 12 chromosomes were filtered using PLINK (v1.90) to include sites with genotyping rate >90% and exclude sites with minor allele frequency <5%. One hundred SNPs were analyzed as blocks to account for possible LD effects. Admixture trees were built using the ten *O. barthii* accessions as the outgroup while allowing m = 0–10 migration events. The model fit for each migration event was examined by estimating the proportion of variance in relatedness between populations explained by each migration model.

PSMC' analysis. Evolutionary demographic changes in *O. glaberrima* and *O. barthii* were inferred using PSMC' (ref. 14). Samples with >20× genome coverage were used: IRRI_103989, IRRI_103992, IRRI_104011, IRRI_104180 and IRRI_105011 for *O. glaberrima* and ba_100941, ba_101226 and ba_104081 for *O. barthii*. Genotype calls for each position were made using the SAMtools (v1.2)[54] mpileup command, filtering for reads with minimum base and mapping quality scores of 30. The soft-masked *O. glaberrima* genome (version 1.1) was used to identify repetitive regions and mask genotype calls overlapping these repetitive regions. Because of inbreeding for *O. glaberrima* and *O. barthii*, we considered each sample to be a single genomic haplotype following Thomas *et al.*[15]. Occasional heterozygous sites were dealt with by randomly sampling one allele. Single haplotypes were combined with those from other samples to create pseudodiploid genomes for the PSMC' analyses.

Of the five *O. glaberrima* samples, four (IRRI_103989, IRRI_103992, IRRI_104180 and IRRI_105011) were from the western coastal region. Pseudodiploids derived from samples within populations generated spurious PSMC' profiles and were excluded from analysis. Thus, all PSMC' results are from pseudodiploids generated from one coastal and one inland haplotype. For the three *O. barthii*, a pseudodiploid genome generated from samples

ba_100941 and ba_104081 resulted in PSMC' profiles that were similar to those for the *O. glaberrima* combinations. This suggested that ba_100941 and ba_104081 corresponded to more *O. glaberrima*–like *O. barthii* samples, possibly from introgression; thus, results are shown for pseudodiploids generated with the more *O. barthii*–like ba_101226 sample and ba_100941 or ba_104081. Analysis employed default parameters for the PSMC' program. Mutation-scaled time and effective population sizes estimated by MSMC were converted assuming a mutation rate of 6.5×10^{-9} substitutions per site per year[55] and a generation time of one generation per year.

Seedling stage salt tolerance phenotyping. Phenotyping for *O. glaberrima* was performed at IRRI in a phytotron at 25–29 °C. Seeds for 121 *O. glaberrima* landraces were pregerminated for 4 d and then transferred to trays where they were suspended in hydroponic nutrient solution containing 1 g/l Jack's Professional fertilizer 20-20-20 (Jack R. Peters, Inc.). Seedlings were acclimated for 5 d before the onset of test A ('early test') and 12 d before the onset of test B ('late test'); these tests evaluate salt tolerance within the window of the salt-sensitive seedling stage[56,57]. We chose these different start times because there is variation among African rice farmers in practices for transplanting seedlings from non-saline beds to the field where salinity exposure may occur. Tray placements in the phytotron were randomized, and locations were randomly reset every 5 d.

Two people separately evaluated plants using SES score (a score of 1–9 corresponding to the visual appearance of the plant, where 1 corresponds to a completely healthy plant and 9 corresponds to a plant exhibiting the full spectrum of salt-sensitive characteristics)[28] and estimated the percentage of injured plants, at multiple intervals during the tests. Mortality was measured twice during test A. Leaf and shoot Na^+ and K^+ levels were measured once for each test (**Supplementary Table 5**).

In test A, the early salinity test beginning 9 d after seeds were first imbibed, seedlings were exposed to an electric conductivity of 12 dS/m for 12 d and then to 18 dS/m for 6 d. In test B, the late salinity test, exposure to electric conductivity of 12 dS/m occurred 16 d after germination with the electric conductivity increased to 18 dS/m 7 d later. Twenty replicate plants of each landrace were grown as controls, two test replicates of 20 replicate plants each were used in test A and one test replicate of 20 replicate plants was used in test B. Control hydroponic trays were kept at <1 dS/m. For each hydroponic tray, two salt-tolerant (Pokkali IRGC 15368 and FL478), one salt-sensitive (IR29) and one moderately salt-sensitive (IR64) *O. sativa* accessions were grown to confirm the hydroponic solution was acting as expected. Solution pH was balanced to 5.0 every other day, and nutrient solution was refreshed every 5 d.

Cation content is correlated with salt tolerance in rice[58]. In the early test (test A), measurements for the fourth leaf generally correlate best with salt tolerance (G.B.G., unpublished observation). The fourth leaf was collected from three plants expressing the typical phenotype for each test replicate. Test B plants were measured differently; the whole shoots of three typical plants were collected for each landrace. Fresh and dry weights were obtained for the material, dried material was powdered and ions were extracted in acetic acid (0.1 N) overnight at 80 °C. Na^+ and K^+ concentrations in the extracts were determined using a flame spectrometer (Model 420, Sherwood Scientific).

The R (ref. 59) package Pgirmess (v1.64) was used to perform the Kruskal–Wallis test with multiple comparisons; we used this test as not all phenotypes had a normal distribution. *P* values were Bonferroni corrected. Populations from the geographical quadrants tested had the following sample sizes: SE inland, $n = 13$; SW coastal, $n = 50$; NE inland, $n = 9$; NW coastal, $n = 49$.

Selection analyses. An XP-CLR test[35] compared the allele frequency distributions for the NW coastal and SW coastal *O. glaberrima* populations to detect selective sweeps. The parameters for the XP-CLR program (v1.0) were as follows: -w1 0.005 100 100 --p0 CHR# 0.8. XP-CLR scores were estimated across non-overlapping 100-bp windows that were used to estimate a maximum XP-CLR score across 10-kb segments. The 10-kb segments with the top 0.5% of maximum XP-CLR values were considered significant. XP-CLR requires a genetic map for modeling of the allele frequency distribution. Currently, however, there is no genome-wide estimate of the *O. glaberrima* recombination rate; thus, the *O. sativa* average recombination rate of 4.13×10^{-6} cM/bp was used[60]. We note that simulations have shown that the XP-CLR method is robust to misestimates of recombination rate[35].

Genome-wide outlier tests for population differentiation were based on the Lewontin–Krakauer test[37]. F_{ST} between the NW coastal and SW coastal populations was calculated using VCFtools (v0.1.12b)[61] implementing the Weir and Cockerham method[36]. The 25-kb windows with the top 0.5% of F_{ST} values were examined; singleton windows with elevated F_{ST} but without any increase in F_{ST} in neighboring windows were not considered. Peak start and end points were determined for each outlier region on the basis of deviation from the local mean F_{ST}.

Genome-wide association mapping. Linear and full mixed-model association analyses were performed for 18 traits assessed in either early or late tests scored at various time intervals (**Supplementary Table 5**). We used both the full pseudomolecule SNP set (2,138,928 SNPs) and the LD-pruned data set (570,528 SNPs), which were further filtered to retain SNPs with call rate >90% and minor allele frequency >5%, resulting in 1,056,028 and 199,093 SNPs in the full and LD-pruned test sets, respectively. Linear association analysis was conducted in PLINK (v1.07), and the Bonferroni-adjusted *P* values of SNPs were calculated using the --adjust function. Mixed-model association analysis was conducted in EMMAX[31] with a Balding–Nichols kinship matrix. Bonferroni *P*-value correction was performed in R. The top ten EIGENSTRAT principal components were used as covariates in all tests.

Manhattan and quantile–quantile plots were generated in R using the package qqman[62]. Median λ values were obtained from PLINK logs or, for mixed-model association analysis, were calculated in R by dividing the median χ^2 test statistic by the expected distribution given 1 degree of freedom. Where two or more significant SNPs were near each other, the window center was one of the two SNPs or the central significant SNP. Regions syntenic between *O. glaberrima* and *O. sativa* subsp. *japonica* were identified using the Ensembl[63] synteny tool, and gene annotations were scanned for known salt tolerance genes and genes involved in cation transport and stress response. Candidate genes were reciprocally used in a BLAST query of the *O. glaberrima* genome to obtain the coordinates and percent similarity of the orthologous genes.

Gene expression analysis of *PPI* and the *OsHAK5* and *OsHAK6* orthologs. Salt tolerance test B was repeated using eight selected accessions (IRGC lines 103989, 105044, 104030, 105052, 104025, 104022, 104023 and 104047); of these, half were tolerant and half were sensitive to salt. Salinization commenced in the test group at 16 d after germination. At 54 h after salt was applied, leaf tissue from the flag leaf center was collected and frozen in liquid nitrogen for two individuals per accession. RNA was extracted from the tissue for each individual (biological replicates) using the Qiagen RNeasy mini kit. RNA cleanup was performed using the Ambion DNA-free kit (Life Technologies). RNA concentrations were adjusted using the Qubit RNA BR assay. Oligo(dT)-primed cDNA was generated with the SuperScript IV cDNA Synthesis kit (Thermo Fisher).

Real-time PCR primers were designed using Primer3 (v. 0.4.0)[64] with settings according to Thornton and Basu[65]. The target genes were *PPI* and the *O. glaberrima* orthologs of *OsHAK5* and *OsHAK6*. For control genes, we used *UBQ10* and *actin11* (**Supplementary Table 11**). SYBR Green Master Mix (Roche Diagnostics) was used to prepare reactions, with each 25-µl reaction containing 1 µl of each primer at a 10 µM concentration and 61 ng of cDNA. A LightCycler 480 instrument (Roche) was used to run real-time PCR and melting curve analysis. Default settings were used for the SYBR Green dye, which included an initial incubation for 10 min at 95 °C followed by 45 cycles of amplification. The annealing temperature was 57 °C. Melting curves confirmed amplification of single products. Each experiment had two biological replicates. Two-tailed *t* tests were used to determine differences in gene expression. For the *OsHAK5* and *OsHAK6* orthologs, the *t* tests assumed unequal variance, whereas the *t* test for the *PPI* gene assumed equal variance.

Determination of salinity in West Africa. Raster maps of soil conductivity[66] were sourced from the US Geological Survey Earth Resources Observation and Science (EROS) database (July 2015), and a raster map of African cropland data[67] was sourced from the International Institute for Applied Systems Analysis. These maps were intersected using ESRI ArcGIS v. 10.1. Areas with

doi:10.1038/ng.3633

low levels of cultivation (1) had 1–33% land use for cultivation, areas with 34–66% use had moderate levels of cultivation (2), and areas with 67–100% use had high levels of cultivation (3). The current soil electric conductivity layer was used to indicate areas where increased salinity in groundwater would cause oversaturation of salt in the soil. Areas assigned as having 'low', 'moderate', 'severe' and 'very severe' electric conductivity were coded as 1–4, respectively.

Farmer interviews in West Africa. In summer 2015, R.S.M., A.P. and M.S. visited *O. glaberrima* farmers in Togo and Senegal together with AfricaRice staff who assisted in translation between French and the local languages Tem and Éwé in Togo and Serer, Wolof and Pular in Senegal. Work in Senegal spanned both the northern arid regions around Saint-Louis and the Sine-Saloum region extending to the border with the Gambia.

Verbal permission was obtained from village chiefs before interviews began. Interviews involved two to seven people at a time, with each interview lasting ~1 h. Each group was asked a core set of questions together with free questions to increase relevant detail. Answers were recorded for each informant. Notes were taken and cross-checked by two people against an audio recording of the interview. Interview results are presented in **Supplementary Table 4**. In Senegal, some informants donated seeds of varieties to AfricaRice. Institutional review board determination of exempt status was granted through the NYU University Committee on Activities Involving Human Subjects (UCAIHS) before conducting interviews (IRB 12-8968).

43. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
44. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
45. Flowers, J.M. *et al.* Whole-genome resequencing reveals extensive natural variation in the model green alga *Chlamydomonas reinhardtii*. *Plant Cell* **27**, 2353–2369 (2015).
46. Hazzouri, K.M. *et al.* Whole genome re-sequencing of date palms yields insights into diversification of a fruit tree crop. *Nat. Commun.* **6**, 8824 (2015).
47. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}; *iso-2*; *iso-3*. *Fly (Austin)* **6**, 80–92 (2012).
48. Korneliussen, T.S., Albrechtsen, A. & Nielsen, R. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* **15**, 356 (2014).
49. Earl, D.A. & vonHoldt, B.M. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* **4**, 359–361 (2012).
50. Jakobsson, M. & Rosenberg, N.A. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**, 1801–1806 (2007).
51. Rosenberg, N. Distruct: a program for the graphical display of population structure. *Mol. Ecol. Notes* **4**, 137–138 (2004).
52. Gronau, I., Hubisz, M.J., Gulko, B., Danko, C.G. & Siepel, A. Bayesian inference of ancient human demography from individual genome sequences. *Nat. Genet.* **43**, 1031–1034 (2011).
53. Tamura, K. *et al.* MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731–2739 (2011).
54. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
55. Gaut, B.S., Morton, B.R., McCaig, B.C. & Clegg, M.T. Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. *Proc. Natl. Acad. Sci. USA* **93**, 10274–10279 (1996).
56. Pearson, G., Ayers, S. & Eberhard, D. Relative salt tolerance of rice during germination and early seedling development. *Soil Sci.* **102**, 151–156 (1966).
57. Gregorio, G.B., Senadhira, D. & Mendoza, R.D. *Screening Rice for Salinity Tolerance* IRRI Discussion Paper Series 22 (IRRI, 1997).
58. Platten, J.D., Egdane, J.A. & Ismail, A.M. Salinity tolerance, Na^+ exclusion and allele mining of *HKT1*;5 in *Oryza sativa* and *O. glaberrima*: many sources, many genes, one mechanism? *BMC Plant Biol.* **13**, 32 (2013).
59. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2015).
60. Wu, J. *et al.* Physical maps and recombination frequency of six rice chromosomes. *Plant J.* **36**, 720–730 (2003).
61. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
62. Turner, S.D. qqman: an R package for visualizing GWAS results using QQ and Manhattan plots. Preprint at *bioRxiv* http://dx.doi.org/10.1101/005165 (2014).
63. Kersey, P.J. *et al.* Ensembl Genomes 2016: more genomes, more complexity. *Nucleic Acids Res.* **44** D1, D574–D580 (2016).
64. Untergasser, A. *et al.* Primer3—new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115 (2012).
65. Thornton, B. & Basu, C. Real-time PCR (qPCR) primer design using free online software. *Biochem. Mol. Biol. Educ.* **39**, 145–154 (2011).
66. Fischer, G. *et al. Global Agro-ecological Zones Assessment for Agriculture* (IASA and FAO, 2008).
67. Fritz, S. *et al.* Cropland for sub-Saharan Africa: a synergistic approach using five land cover data sets. *Geophys. Res. Lett.* **38**, L04404 (2011).