Domestication history and geographic 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], Annie Barretto[3], Katherine Dorph[1], Briana Gross[4], Dorian Q. Fuller[5], Kofi Bimpong[6], Marie-Noelle Ndjiondjop[7], Glenn B. Gregorio[3] and Michael D. Purugganan[1,2]

[1]Center for Genomics and Systems Biology, Department of Biology, 12 Waverly Place, New York University, New York 10003

[2]Center for Genomics and Systems Biology, New York University Abu Dhabi, Saadiyat Island, Abu Dhabi, United Arab Emirates

[3]Plant Breeding, Genetics and Biotechnology Division, International Rice Research Institute, Los Banos, Laguna, Philippines

[4]Department of Biology, University of Minnesota, 1035 Kirby Drive Duluth, MN 55812

[5]Institute of Archaeology, University College London, 31-34 Gordon Square, London WC1H 0PY UK

[6]AfricaRice Sahel Station, Ndiaye, B.P. 96, Saint Louis, Senegal

[7]AfricaRice Centre, 01 B.P. 2031, Cotonou, Benin

*Corresponding Author: Michael Purugganan Email: mp132@nyu.edu Tel. +1 (212) 998 3801

African rice (*Oryza glaberrima* Steud.) is a cereal crop species that shares a common ancestor with Asian rice (*O. sativa* L.) but was independently domesticated in West Africa ~3,000 years ago.[1-3] African rice is rarely grown outside sub-Saharan Africa, and is of interest because of its tolerance to abiotic stresses.[4,5] Here we describe a map of 2.32 million single nucleotide polymorphisms (SNPs) of African rice from whole genome re-sequencing of 93 landraces. Population genomic analysis reveals a population bottleneck in this species that began ~13-15 thousand years ago (kya), with effective population size reaching its minimum value ~3.5 kya, suggesting a protracted period of population size reduction likely commencing with pre-domestication management and/or cultivation. Genome-wide association studies (GWAS) with 6 salt tolerance traits also identify 11 significant loci, four of which overlap or are within ~300 kb of genomic regions that possess signatures of positive selection, suggesting adaptive geographic divergence for salt tolerance in this species.

We used paired-end (2x100-bp) Illumina sequencing to re-sequence the genomes of 93 traditional *O. glaberrima* landraces from across its 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 inland areas in Nigeria, Niger, Cameroon, Chad, Mali and Burkina Faso. Four landraces were sequenced deeply (~30x-73x mean nuclear genome coverage depth), and the remaining accessions were sequenced to an average depth of ~14.61x (missing genotype calls <4%; Supplementary Table 1). This yielded 381 Gb of mappable sequence when aligned to the *O. glaberrima* CG14 reference genome sequence.[3]

Following the application of quality control filters, we identified 2,317,937 SNPs or approximately 7.32 SNPs/kb, in African rice (Fig 1b,c; Supplementary Figs 1-3). We estimate nucleotide diversity (π) to be 0.0034 ± 0.0032 (± standard deviation) 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 see 905,654 SNPs (39% in genic regions), which include 51,296 synonymous, 54,833 nonsynonymous, 110,390 intron, 19,860 upstream and 667,237 downstream SNPs. There are 1,105 nonsense mutations that truncate encoded proteins. Linkage disequilibrium (LD) is substantial, with r^2 reaching half its maximum value at ~175 kb and approaching the baseline at ~300 kb (Fig 1d).

Principal component analyses (PCA) of SNP variation using EIGENSTRAT revealed 10 significant components ($P < 0.0001$),[7] the top three PCs each explaining <4% of the variance (Fig 2a and Supplementary Table 3). The top two eigenvectors are strongly correlated with geography, PC1 with an east/west (r = -0.77) and PC2 a north/south cline (r = -0.57)[Fig 2b]. An alternative view of populations stratification is offered by the population clustering program STRUCTURE, which infers an optimal number of genetic clusters that comprise *O. glaberrima* landrace genomes as K=6 (Fig 2c; for other K values, Supplementary Fig 4).[8] We find that all landraces predominantly belong to one cluster, with various levels of genomic contributions from 5 other ancestral populations.

To examine spatial genetic variation, we chose 11^o N latitude to divide the arid north from the tropical south, and 6^o W longitude to separate the Western Atlantic coast from eastern inland areas (Fig 1a). These are consistent with clines observed in the PCA

analyses, 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 Porteres.[1,3,9,10] The NW/SW coastal division 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 Guinean highlands between Sierra Leone and Liberia. For a review of proposed domestication and diversification centers, see Supplementary Note 1.

We used TREEMIX to examine the topology of relationships and migration history among populations.[11] Using the wild progenitor *O. barthii* A. Chev as an outgroup, we observe an older split between coastal and inland populations, and a more recent separation of northern and southern populations (Fig 3a). Even without migration (m = 0), this topology accounts for >99% of the variance in SNP data. Greater model support is provided by inferred gene flow from the SW coastal to the SE inland populations (m = 1)[Fig 3a], but this only marginally improves model fit.

Domestication is typically accompanied by population bottlenecks,[12,13] and to examine this we applied a multiple sequentially Markovian coalescent model on two haplotypes (PSMC').[6] PSMC' profiles indicate reduction in effective population size (N_e) that began ~13-15 kya with an N_e ~60,000 and reaching a minimum N_e of ~3,000 at ~3.5 kya (Fig 3b). This severe bottleneck during the domestication of African rice[14] is similar to what is observed in other annual crop species.[12,13,15-17] In contrast, no severe bottleneck is evident in wild *O. barthii* (Fig 3b).

For *O. glaberrima*, the recent maximum N_e observed at ~15 kya coincides with an increase in precipitation in West Africa after deglaciation leading into the start of the

early Holocene African Humid Period (AHP).[18,19] The presence of recognizably domesticated African rice, however, does not appear in the archaeological record until ~2.8-2.4 kya, from sites in the inland Niger delta in Mali.[20,21] Interestingly, our PSMC' results shows that the minimum plateau in N_e for African rice occurs near these earliest archaeological dates. The analysis thus indicates an early onset of the bottleneck, and may point to a protracted period of low-intensity cultivation and/or management before full domestication ~3.5 kya, just after a peak in human population growth in western Africa between 4-5 kya.[19] Archaeobotanical evidence for this protracted utilization is elusive, since remains in West Africa prior to 5 kya are extremely rare.[22] Early ceramic finds in the Early Holocene, however, do suggest early consumption of grass grains in the region.[23]

Post-domestication spread of crop species is associated with adaptation to multiple environments, and one key trait likely associated with geographical adaptation in African rice is salinity tolerance.[24,25] Arid regions of northern West Africa have higher salinity levels, associated with saltwater intrusion into rivers which can reach up to 250 km inland.[1] We interviewed African farmers in inland Togo and coastal Senegal (high salinity) about salt stress mitigation (Supplementary Table 4). In Senegal, despite efforts to control soil salinity, which affects most plant developmental stages, the major strategy was to farm salt tolerant varieties: ~25% of *O. glaberrima* varieties used by farmers in this area were reported as 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 include visual plant salt stress symptoms using the standard evaluation system

(SES) score,[26] percent shoot injury, mortality, leaf Na^+ and K^+ content (Supplementary Table 5). There are significant differences in phenotypes among populations except for those measuring Na^+ and K^+ content (Fig 4a; Supplementary Fig 5; Supplementary Table 6); for example, using Kruskal-Wallis tests in late salinity experiments for SES ($P < 1.96 \times 10^{-7}$), percent injury ($P < 8 \times 10^{-6}$), mortality ($P < 1.46 \times 10^{-7}$) and fourth leaf [Na^+] ($P < 2.13 \times 10^{-8}$).

Pairwise population comparisons show that this difference is driven by reduced salt tolerance in the SW coastal population (Supplementary Table 7). Contrasting populations in late salinity tests, for example, reveals Bonferroni significant ($P<0.0084$) differences between the SW vs. its sister NW population, where the SW population was higher in SES score, percent injury, and mortality phenotypes. There are no significant differences between the NW coast, NE and SE inland populations for these traits (Supplementary Table 7). Loss of salinity tolerance of the SW coast population is possibly associated with costs of maintaining tolerance in a region of greater rainfall and reduced soil salinity (Supplementary Fig 6).[1,27]

We conducted GWAS mapping with the 93 landraces whose genome we re-sequenced,[28] using a full SNP set of 1,056,028 SNPs after removing low-frequency SNPs (minor allele frequency < 5%), and a further reduced set of 199,093 SNPs obtained by pruning LD-correlated ($r^2 > 0.5$) SNPs. We performed linear[29] and mixed-model[30] associations, using the first 10 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 threshold ($P < 2.5 \times 10^{-7}$ in reduced set) to identify significant SNPs (Fig 4b,c, Supplementary Fig 7). Twenty-eight

SNPs that exceed the significance threshold were identified in 11 unique genomic regions (Supplementary Table 8). Seven loci were associated with percent shoot injury, two with SES scores, and two with both traits. Further work will be necessary to identify specific genes underlying these quantitative trait loci (QTLs), although there are plausible candidates genes in a few regions. On chromosome 1, two homologs to 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.[31,32] We also identified a homolog of *Salt Overly Sensitive 4* (*SOS4*)[33] about 73 kb from a significant shoot injury GWAS SNP in chromosome 4.

To identify genomic regions associated with adaptive differentiation between *O. glaberrima* populations, we used the cross-population composite likelihood ratio (XP-CLR) method on the NW vs. SW coast populations.[34] Using the upper 0.5% of CLR as a cut-off, we identified 98 selected genomic regions using either SW or NW coast populations as test populations, with 22 regions overlapping in the two tests (Fig 4d,e and Supplementary Table 9). These genomic regions range in size from ~10 to ~760 kb.

We examined any overlap between QTLs identified by GWAS and putative selected regions identified by XP-CLR analyses. Two GWAS hits, on chromosomes 5 (position 14.75 Mb) and 11 (position 19.23 Mb) are found within 300 kb of a selected region identified by XP-CLR. The most promising region however, encompasses two salt tolerance GWAS hits on the proximal end of chromosome 4 and overlaps 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}) between

NW and SW coastal populations, and find that the genomic region around these two GWAS loci has a mean F_{ST} = 0.157. Several SNPs are in the upper 0.5% of the distribution (mean genome-wide F_{ST} = 0.027) [Fig 5a; Supplementary Fig 8; Supplementary Table 10], providing additional support for adaptive differentiation in this genomic region.[35] We find 41 genes in this area of overlap (Fig. 5b); one possible positional candidate gene encodes a peptidyl-prolyl *cis/trans*-isomerase (PPIase), a member of a gene family involved in stress response,[36] some of which confer seedling salt tolerance in Asian rice.[37]

In summary, our analysis of African rice provides genetic evidence for an extended period of low-intensity cultivation or management of a wild species prior to its domestication. There have been two competing hypotheses for the timescale of domesticated crop origins - the rapid vs. protracted transition models of domestication; our work provides support for the latter.[38-40] Our study also identifies genomic regions associated with geographic 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.

REFERENCES

1. Carney, J.A. *Black Rice: The African Origins of Rice Cultivation in the Americas*. (Harvard University Press, Cambridge, MA, 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. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).

7. Price, A. *et al*. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).

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.) pp. 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.) pp. 409–452 (Mouton, The Hague, 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., U'Ren, J., Tenaillon, O. & Gaut, B.S. Selection versus demography: A multilocus investigation of the domestication process in maize. *Mol. Biol. Evol*. **2**, 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. Nabholz, N. *et al*. Transcriptome population genomics reveals severe bottleneck and domestication cost in the African rice. *Mol. Ecol*. **23**, 2210–2227 (2014).

15. Eyre-Walker, A. *et al.* Investigation of the bottleneck leading to the domestication of maize. *Proc. Natl. Acad. Sci. USA* **95**, 4441–4446 (1998).

16. Hyten, D.L. *et al.* Impacts of genetic bottlenecks on soybean genome diversity. *Proc. Natl. Acad. Sci. USA* **103**, 16666–16671 (2006).

17. Zhu, Q. *et al.* 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).

18. Tjangilli, R. *et al.* Coherent high- and low-latitude control of the northwest African hydrological balance. *Nat. Geosci*. **1**, 670–675 (2008).

19. Manning, K. and Timpson, A. The demographic response to Holocene climate change in the Sahara. *Quat. Sci. Rev*. **101**, 28–35 (2014)

20. Murray, S.S. in *Fields of Change: Progress in African Archaeobotany*. (ed. Cappers, R.T.J.) pp. 53–62 (Barkhuis, Groningen, 2007).

21. 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).

22. 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. and Fuller D.Q) pp. 17–24 (Left Coast Press, Walnut Creek, CA, 2014).

23. Eichhorn, B. & Neumann, K. in *Archaeology of African Plant Use* (eds. Stevens C., Nixon S., Murray M.A. and Fuller D.Q) pp. 83–96 (Left Coast Press, Walnut Creek, CA, 2014).

24. Temudo, M. Planting knowledge, harvesting agro-biodiversity: A case study of Southern Guinea-Bissau rice farming. *Human Ecology* **39**, 301–321 (2011).

25. Carney, J.A. Landscapes of technology transfer: Rice cultivation and African continuities. *Technology and Culture* **37**, 5–35 (1996).

26. International Rice Research Institute. *Standard Evaluation System for Rice*. (International Rice Research Institute, Manila, Philippines, 2014).

27. Matlon, P., Randolph, T & Guei, R. in *Impact of Rice Research* (eds. Pingali, P.B. and Hossain, M.) pp. 382–404 (International Rice Research Institute, Manila, Philippines, 1998).

28. Huang, X. *et al*. Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* **42**, 961–967 (2010).

29. 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).

30. Kang H.M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet*. **42**, 348–354 (2010).

31. Yang, T.Y. *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).

32. 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).

33. Shi, H.Z. *et al*. The *Arabidopsis salt overly sensitive 4* mutants uncover a critical role for vitamin B6 in plant salt tolerance. *Plant Cell* **14**, 575–588 (2002).

34. Chen, H., Patterson, N., & Reich, D. Population differentiation as a test for selective sweeps. *Genome Res*. **20**, 393–402 (2010).

35. Lewontin, R.C. & Krakauer, J. Distribution of gene frequency as test of the theory of the selective neutrality of polymorphism. *Genetics* **74**, 175–195 (1973).

36. 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).

37. 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).

38. 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. U.SA* **105**, 13982–13986 (2008).

39. Fuller D. Q. Contrasting patterns in crop domestication and domestication rates: Recent archaeobotanical insights from the Old World. *Annals Bot*. **100**, 903–924 (2007).

40. Wilcox, G. in *Biodiversity in Agriculture: Domestication, Evolution, and Sustainability* (eds. Gepts, P., Famula, T. et al.) pp. 92–109. (Cambridge University Press, Cambridge, UK, 2012).

FIGURE LEGENDS

Figure 1. A SNP map for African rice. (a) Location of sampled landraces in West Africa, with the latitude (11^{o} N) and longitude (6^{o} W) demarcating the 4 geographic quadrants delimited in this study indicated by dashed lines. (b) CIRCOS plot showing SNP diversity across the 12 chromosomes of *O. glaberrima*. The chromosomes are numbered. The outer circle indicates SNP density, the middle blue circle is nucleotide diversity (π) and the inner green circle depicts the population mutation parameter (θ_W). (c) Close-up of variation across chromosome 1, with blue indicating π (per bp) and red SNP number in each 25-kb window. Position along chromosome is in Mb. (d) Relationship of mean linkage disequilibrium with distance.

Figure 2. Population structuring in African rice. (a) Principal components (PCs) of SNP variation. Samples from the NW coast (red), SW coast (green), NE inland (orange) and SE inland (blue) populations are shown. The plots are for the first 3 principal components. (b) PC1 and PC2 scores for each *O. glaberrima* accession are shown, with geographic location of the samples plotted in the map. Visually, the east/west cline for PC1 and the north/south cline for PC2 are evident. (c) STRUCTURE plot for African rice, showing distribution of the K=6 genetic clusters. The 4 different West African populations are indicated. (d) Neighbor-joining clustering of landraces based on genetic

distance. Colors of the branches indicate membership in one of the 4 geographic populations.

Figure 3. Demography of *O. glaberrima* and *O. barthii*. (a) Treemix analysis of 93 *O. glaberrima* samples divided into 4 West African geographic quadrants, with *O. barthii* samples serving as the outgroup population. Arrow represents the direction of migration. (b) PSMC'-inferred demographic history of *O. glaberrima* and *O. barthii*. Each line represents past effective population sizes for a pseudodiploid genome generated by combining haploid sequences of *O. glaberrima* or *O. barthii*. Blue line represents coastal/inland *O. glaberrima* pseudodiploid combination, and red line represent intraspecific *O. barthii* combinations.

Figure 4. Population phenotypic differentiation, GWAS mapping and selective sweep analysis for salinity tolerance. (a) Geographic variation in salt tolerance phenotypes. Bar plots represent the 4 West African populations display highest and lowest quartiles around the median (box) and 1.5 times this interquartile range (whisker). The line in the box is the mean, while the dots outside the whiskers are outlier values. Specific traits are described in Supplementary Table 5. SES4 is standard evaluation system score, INJ5 is percent shoot injury, and MORT2 is mortality count in late salinity tests. Na^+ and K^+ concentrations in the 4^{th} leaf are also shown. In all but one trait (K^+ in 4^{th} leaf), the SW coast population shows a significant difference compared to other populations. (b) GWAS Manhattan plots for percent leaf injury in late salinity test using unpruned SNP set with linear model, and (d) SES score in late salinity test using LD-

pruned SNP set with linear model. Red line indicates Bonferroni significance threshold. Chromosome numbers are indicated below. (d) Genome-wide distribution of XP-CLR values using NW coastal as reference and SW costal as object population, and (e) SW coastal as reference and NW coastal as object population. Red line indicates the 99.5% percentile XP-CLR value.

Figure 5. Comparison of GWAS and selected genomic regions at proximal end of chromosome 4. (a) The top Manhattan plot shows the GWAS results between position 1 to 1.5 Mb on chromosome 4. The GWAS is for percent leaf injury in late salinity test using unpruned SNP set with linear model. The red line indicates the Bonferroni significance threshold. The middle plot shows the XP-CLR results using SW coastal population as the reference and NW coastal population as the object population. 99.5% percentile XP-CLR value is shown in red line. The bottom plot is a sliding window analysis of F_{ST} values across the genomic region. The red line is the 99.5% percentile for genome-wide F_{ST} values. The blue line indicates the 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 region. (b) Position of 41 gene models in this region are indicated below, where red represents minus strand and blue represents positive strand genes.

ONLINE METHODS

Sample collection and library preparation. Seeds were obtained from the International Rice Research Institute (IRRI) and from the United States Department of Agriculture

(USDA) [see Supplementary Table 1 for accessions]. DNA was extracted from a single seedling leaf using either the DNeasy mini kit (Qiagen, Venlo, Netherlands) or using standard phenol-chloroform-isoamyl alcohol buffer. Libraries were prepared using Illumina TruSeq (San Diego, CA) kits with an average 380 bp insert size. 2 x 100 paired-end sequencing was done for the *O. glaberrima* samples on an Illumina HiSeq 2500 sequencer at the New York University Center for Genomics and Systems Biology Genome Core with 2-7 libraries per lane. 2 x 100 paired-end sequencing was also completed for three *O.barthii* samples on an Illumina HiSeq 2000 sequencer at the University of Minnesota Genomics Center (UMGC). For these, barcoded TruSeq DNA Nano libraries were prepared with a 350-bp insert, and the pooled libraries sequenced using 1/2 lane.

Read alignment and SNP calling. Sequencing reads passing Illumina's quality control filter were aligned using Burroughs-Wheeler Aligner (v0.6.1) to map reads to the Arizona Genome Institute *Oryza glaberrima* genome version 1.1[3] that included both the 12 pseudomolecules and 1,939 unassembled scaffolds (Genebuild version 2011-05-AGI). Duplicate reads were marked and removed from individual sample alignments using Picard-tools (v1.111) MarkDuplicates and then merged using MergeSamFiles. Global realignments of reads around indels was done using the Genome Analysis Toolkit (GATK v3.1-1)[41,42] RealignerTargetCreator/IndelRealigner protocol.

Before processing the final SNP map of 93 *O. glaberrima* samples, we had mapped 99 samples, assumed to be *O. glaberrima* based on the accession labels from IRRI or USDA, to the *O. sativa* Nipponbare reference genome (IRGSP 1.0), along with

57 diverse *O. sativa* accessions (I.S. Pires, J.M. Flowers and M.D. Purugganan, unpublished data). SNP calling and filtering (using the methods described below) produced a set of over 6 million SNPs that were used in STRUCTURE population clustering analyses (see below) to look for mislabeled or admixed samples. Six samples were found to not be true *O. glaberrima*: IRGC 103587, 103602, 103961, 104037, 104254, and 106291. These were removed from the study.

SNP calling was performed using the GATK Unified Genotyper set for diploids using default filtering settings, similar to previous studies.[43,44]. Base qualities were capped at the mapping quality of the read, and all reads mapping to two or more places were automatically filtered out. Filtering for all SNPs was done in GATK using settings based on outlier transition/transversion ratio that are enriched false positives. For the SNP set of 93 *O. glaberrima* samples, filtering was applied with the following settings: DP>10000, FS > 212, MQ < 11, MQ0 > 5000, MQRankSum < -46, QD < 0.16, and ReadPosRankSum < -15. This filtering strategy reduced the raw unfiltered set of 2.88 million SNPs to the working set of 2.3 million (with scaffolds) and 2.14 million SNPs (pseudomolecules only) analyzed in this study.

The three *O. barthii* accessions sequenced as part of this project (IRGC 101226, 100941, and 104081) and 7 publicly available DNA libraries from the SRA archives (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 93 *O. glaberrima* set. Settings used to filter the *O. barthii* set were DP > 1500, FS > 74, MQ < 37, MQ0 > 5000, MQRankSum < -5.50, QD < 0.4, and ReadPosRankSum < -4. This set of *O. barthii*

and *O. glaberrima* SNPs, as well as individual sample BAMs, were used in downstream population analyses that included both species.

To validate SNPs, we tested 51 SNPs and 131 genotypes by Sanger sequencing. Each of regions were amplified in at least one out of 19 haphazardly chosen *O. glaberrima* DNA templates and both forward and reverse strands were Sanger sequenced in multiple DNA templates (Supplementary Table 2). Contigs were assembled in Sequencher (Gene Codes, Ann Arbor, MI) 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 nonequivalent base calls in Sanger sequence chromatograms. We find 9/131 genotypes were different between the two methods, giving a concordance rate of 93.1%. Eight out of 9 errors were false heterozygous calls in the VCF that appeared to be homozygous reference or alternative alleles in the Sanger data. One error was found to be a genotype that was neither of the biallelic variants in the VCF. Of the concordant assigned genotypes screened, 30 were homozygous alternative SNP, 7 were heterozygous, and 85 were homozygous reference allele.

SNP annotation. SNPeff (v3.6c)[45] was used to assign SNP effects based on gene models from the AGI v0.1.1 2012 annotation (still current as of February 2016). Codons with multiple SNPs in the same codon were annotated separately, and only canonical transcripts were used. SNP effect classifications were dependent on the contemporary gene models available; however, given the discrepancies between *O. glaberrima* and *O.*

sativa annotation quality, it is clear that improvement in annotated gene models will change the counts of effects in the SNP set.

Population genetic parameters. The program ANGSD (v0.613)[46] was used to calculate population genetic statistics θ_W, π, and Tajima's D directly from sample BAMs in 25-kb non-overlapping intervals; this was done for the whole sample set and for populations assigned by geographic quadrant.

The same set of SNPs from the TREEMIX analysis (see below) was used to estimate genome-wide linkage disequilibrium (LD). LD was calculated in PLINK (v1.90)[29] using SNP pairs that were within a 1,000-kb window and at a maximum distance of 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 each bin. The average r^2 per bin were plotted for each chromosome and a line of best fit was plotted using LOESS curve fitting.

Population structure analysis and genetic distance relationships. STRUCTURE (version 2.3.4)[8] was run using a reduced SNP set in which SNPs called in scaffolds and not pseudomolecules were removed, 4% of the total SNPs were chosen at random and retained, and then these were LD pruned in PLINK (v1.90)[29] using settings '-indep 50 5 1.5', which left of 29,983 SNPs. Using settings for admixture and no linkage, STRUCTURE was run with a burnin of 50,000 replicates and 50,000 MCMC iterations following the burnin step. This was repeated 10 times for each K value (from 1 to 8). Results were analyzed using the EVANNO method with STRUCTURE HARVESTER[47]

and CLUMPP (v. 1.1.2)[48] was used to permute run clusters. DISTRUCT[49] was used to plot the results of K=3 through K=6 (Supplementary Fig 4). The delta(K) indicates K=6 to be optimal.

Principal component analysis (PCA) was done using the EIGENSOFT package to run EIGENSTRAT[7] on the complete pseudomolecule SNP set that had been LD pruned to 570,728 SNPs. The top ten principal components were used in geographic analysis as well as in downstream genome-wide association mapping.

A neighbor-joining tree was constructed using the filtered SNP set of 93 *O. glaberrima* accessions; with distances calculated using the Gronau method[50], described in Hazzouri et al., (2015)[44]. The tree was constructed from the distance matrix using Mega (v5.2)[51].

TreeMix v.1.12[11] was used to model the admixture among the 93 *O. glaberrima* samples divided by geographic quadrants (see Fig 1). Initially, SNPs segregating across the 12 chromosomes were filtered using PLINK (v1.90) to include sites with greater then 90% genotyping rate and exclude sites with minor allele frequency less then 5%. One hundred SNPs were analyzed as blocks to account for possible effects of LD. Admixture trees were built using the 10 *O. barthii* as the outgroup sample while allowing 0 – 10 migration events. Model fit of 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 the multiple sequentially Markovian coalescent model on two

haplotypes (PSMC').[6] Samples with at least 20X genome coverage was used for the analysis: IRRI_103989, IRRI_103992, IRRI_104011, IRRI_104180, and IRRI_105011 for *O. glaberrima*; ba_100941, ba_101226, and ba_104081 for *O. barthii*. Genotype calls for each genomic position were made using Samtools (v1.2)[52] mpileup command, filtering for reads with a minimum base quality score of 30 and mapping quality score 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. Due to inbreeding for *O. glaberrima* and *O. barthii* we considered each sample as a single genomic haplotype following Thomas et al. 2015.[53] Occasional heterozygous sites were dealt with by randomly sampling one allele. Each single haplotype were then combined with other samples to create pseudodiploid genomes for the PSMC' analyses.

Out of the 5 *O. glaberrima* samples, 4 (IRRI_103989, IRRI_103992, IRRI_104180, and IRRI_105011) were from the west coastal region. Pseudodiploids generated from 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 3 *O. barthii*, a pseudodiploid genome generated from sample ba_100941 and ba_104081 resulted in PSMC' profiles that were similar to the *O. glaberrima* combinations. This suggested ba_100941 and ba_104081 were a more *O. glaberrima*-like *O. barthii* sample, possibly from introgression; thus results are shown for pseudodiploids generated between the more *O. barthii*-like ba_101226 sample and ba_100941 or ba_104081. Default parameters of the PSMC' program were used for the analysis. Mutation scaled time and effective population size estimated by MSMC were

converted by assuming a mutation rate of 6.5×10^{-9} substitutions per site per year[54] and a generation time of one generation per year.

Salt tolerance phenotyping of the seedling stage. Phenotyping for *O. glaberrima* was done at the International Rice Research Institute (IRRI) in a phytotron at 25^{o}-29^{o}C controlled temperature range. Seeds for 121 *O. glaberrima* landraces were pre-germinated for 4 days and then transferred to trays suspended in hydroponic nutrient solution containing 1 g/L of Jack's Professional fertilizer 20-20-20 (Jack R. Peters, Inc), where they acclimated for five days before the onset of Test A (also referred to as the early test), and twelve days before the onset of Test B (late test); these tests evaluate salt tolerance within the window of the salt-sensitive seedling stage[55,56]. We chose these different start times because there is variation in when people transplant African rice from elevated beds, that are watered by rain or well water, to the field where exposure to salinity may occur. Tray placements in the Phytotron were randomized, and locations were haphazardly reset every 5 days.

Two people separately evaluated plants using the standard evaluation score (SES: a visual score of 1-9, where 1 is completely healthy and 9 is exhibiting full salt sensitive characteristics)[26] and estimated percent plant injury, at multiple intervals during the tests. Mortality was measured twice during Test A. Leaf and shoot sodium and potassium levels were measured once for each test. The intervals of when the phenotype measurements were made are in Supplementary Table 5.

In Test A, the early salinity test beginning 9 days after seeds were first imbibed, seedlings are exposed to an electric conductivity (EC) of 12 dSm^{-1} for 12 days, then to 18

dSm^{-1} for 6 days. In Test B, the late salinity test, exposure to EC12 dSm^{-1} salt levels was at 16 days after germination, and the increased to EC18 dSm^{-1} seven days 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 dSm^{-1}. For each hydroponic tray, two salt tolerant (Pokkali IRGC 15368 and FL478), one salt sensitive (IR29) and one moderately salt sensitive (IR64) *O. sativa* checks 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 days.

Cation content is correlated with salt tolerance in rice[57]. In the standard SES test (Test A), the 4th leaf generally best correlates to salt tolerance (Gregorio pers. comm.). The 4th leaf was collected from three plants expressing the typical phenotype per test replicate. Test B plants were measured differently; the whole shoots of three typical plants per landraces were collected. Material fresh and dry weight was obtained, dried material was powdered and ions were extracted in acetic acid (0.1 N) overnight at 80 °C. Na^+ and K^+ in the extracts were determined using a flame spectrometer (Model 420; Sherwood Scientific, Cambridge, UK).

The R[58] package Pgirmess[59] (v1.64) was used to perform the Kruskal-Wallis test with multiple comparisons; we used this test since not all phenotypes had a normal distribution. *P* values were Bonferroni-corrected. Geographic quadrants tested had the following sample sizes: SE inland (n=13), SW coast (n=50), NE inland (n=9), and NW coast (n=49).

Selection analyses. Cross population composite likelihood ratio (XP-CLR) test[26] was used to compare the NW and SW costal *O. glaberrima* population allele frequency distribution to detect selective sweeps. XP-CLR program (v1.0) was used with the following parameter: "-w1 0.005 100 100 –p0 CHR# 0.8". XP-CLR scores were estimated across nonoverlapping 100-bp windows that was then used to estimate a maximum XP-CLR score across 10-kb segments. 10-kb segments within the top 0.5% maximum XP-CLR values were considered significant. XP-CLR requires a genetic map during its modeling of the allele frequency distribution. Currently, however, there is no genome-wide estimate of a *O. glaberrima* recombination map thus the *O. sativa* average recombination rate of 4.13×10^{-6} cM/bp was used.[60] We note that simulations have shown the XP-CLR method is robust to misestimates of recombination rate.[26]

We also did a genome-wide outlier test for population differentiation based on the Lewontin-Krakauer test.[35] F_{ST} between NW vs. SW coast populations was calculated using VCFtools (v0.1.12b)[61] implementing the Weir and Cockerham method (1984).[62] The top 0.5% of 25-kb windows across the genome were examined and singleton elevated windows, without any increase in F_{ST} in neighboring windows, were not considered. Peak start and end points were determined for each outlier region based on deviation from the mean F_{ST} at that position.

Genome-wide association mapping. Linear and full mixed-model associations were performed for 18 traits in either early and late tests, and scored at various time intervals (see Supplementary Table 5). We used both full pseudomolecule SNP (2,138,928 SNPs) and LD pruned datasets (570,528 SNPs) that were further filtered to retain SNPs with

>90% called genotypes and a minor allele frequency >5%, resulting in 1,056,028 and 199,093 SNPs in the full and LD pruned test sets, respectively. Linear associations were conducted in PLINK (v1.07) and Bonferroni significance values of SNPs were calculated using the —adjust function. Mixed-model associations were conducted in EMMAX[30] with a Balding-Nichols kinship matrix and Bonferroni *P* value correction was performed in R. The top ten EIGENSTRAT PCs were used as covariates in all four kinds of tests.

Manhattan and quantile-quantile (Q-Q) plots were made in R using the package qqman.[63] Median lambda values were obtained from PLINK logs or for mixed-model associations, were calculated in R by dividing the median chi-square test statistic by the expected distribution given one degree of freedom. Where two or more significant SNPs were near each other, one of the two or the central significant SNP, was used as the window center. Syntenic regions between *O. glaberrima* and *O. sativa* subsp. japonica were determined using the Ensembl[64] synteny tool based on orthology calculated from collinear gene blocks, 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 gene.

Determination of salinity in West Africa. Raster maps of soil conductivity[65] were sourced from the USGS Earth Resources Observation and Science (EROS) database (July 2015) and a raster map of African cropland data[66] was sourced from the International Institute for Applied Systems Analysis. These were intersected using ESRI ArcGIS v. 10.1. Areas between 1-33% land use for cultivation were coded as 1 for low cultivation.

Areas between 34-66% were coded as 2 for moderate. Areas between 67-100% were coded as 3 for high cultivation. The current soil electric conductivity (EC) layer was used to indicate areas where increased salinity in groundwater as a result of overdrawn irrigation systems would cause oversaturation of salt in the soil. Areas assigned "low", "moderate", "severe", and "very severe" EC were coded 1-4.

Farmer interviews in West Africa. In Summer 2015, Meyer, Plessis, and Sanches visited *O. glaberrima* farmers in Togo and Senegal together with AfricaRice staff that assisted in translation between French and the local languages Tem and Éwé in Togo and Serer, Wolof, and Pular in Senegal. Work in Togo was overseen by Ndjiondjop, who selected the field sites in the Plateaux region of Togo, determined to be suitably representative of the rice agricultural zone through interviews with faculty at University Abomey Calavi Benin and other staff at AfricaRice Cotonou. Work in Senegal was overseen by Bimpong and spanned both the northern arid regions around St. Louis and the Sine Saloum region until the Gambia border.

Village chief verbal permission was obtained before interviews began. Seven interviews were done involving groups of 2 to 7 people at a time, each lasting ~1 hr. Each group was asked a core set of questions and free questions to obtain more relevant detail. Answers were recorded per informant. Notes were taken and cross-checked by two people with an audio recording of the interview. Interviews were analyzed and answers were presented in tables (Supplementary Data 1). In Senegal, some informants donated seeds of varieties to AfricaRice. Institutional Review Board determination of Exempt

status was granted through the New York University University Committee on Activities Involving Human Subjects (UCAIHS) prior to conducting interviews (IRB# 12-8968).

ONLINE METHODS REFERENCES

41. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet*. **43**, 491–498 (2011).

42. McKenna, A. *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. **20**, 1297–1303 (2010).

43. 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).

44. Hazzouri, K.M., *et al.* Whole genome re-sequencing of date palms yields insights into diversification of a fruit tree crop. *Nat. Comm*. **6**, doi:10.1038/ncomms9824 (2015).

45. 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 w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).

46. Korneliussen, T.S., Albrechtsen A. & Nielsen, R. ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics* **15**, doi: 10.1186/s12859-014-0356-4 (2014).

47. Earl, D.A., & vonHoldt, B.M. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Cons. Genet. Resources* **4**, 359-361 (2012).

48. Jakobsson, M. & Rosenberg, N. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**, 1801–1806 (2007).

49. Rosenberg, N. Distruct: a program for the graphical display of population structure. *Mol. Ecol. Notes* **4**, 137-138 (2004).

50. Gronau, I. *et al.* Bayesian inference of ancient human demography from individual genome sequences. *Nat. Genet.* **43**, 1031–1034 (2011).

51. 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).

52. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

53. Thomas, C. *et al.* Full-genome evolutionary histories of selfing, splitting, and selection in *Caenorhabditis*. *Genome Res*. **25**, 667–678 (2015).

54. 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).

55. Pearson, G., Ayers, S. & Eberhard, D. Relative salt tolerance of rice during germination and early seedling development. *Soil Sci.* **102**, 151–156 (1966).

56. Gregorio, G.B., Senadhira, D. & Mendoza, R.D. *Screening Rice for Salinity Tolerance*. IRRI Discussion Paper Series 22 (IRRI, Manila, Philippines, 1997).

57. Platten, J.D., Egdane, J. & Ismail, A. 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).

58. R Core Team. R: *A language and environment for statistical computing.* (R Foundation for Statistical Computing, Vienna, Austria, 2015).

59. Giradoux, P. pgirmess: Data analysis in ecology. R package version 1.6.4 (2016).

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. Weir, B.S. & Cockerham, C.C. Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370 (1984).

63. Turner, S.D. qqman: an R package for visualizing GWAS results using QQ and Manhattan plots. bioRxiv, 005165 (2014).

64. Kersey, P.J., *et al*. Ensembl Genomes 2016: more genomes, more complexity. *Nucleic Acids Res*. **44**, D574-D580 (2016).

65. Fischer, G. *et al*. Global Agro-ecological Zones Assessment for Agriculture (IASA, Laxenburg, Austria and FAO, Rome, Italy, 2008).

66. Fritz, S. *et al*. Cropland for sub-Saharan Africa: A synergistic approach using five land cover data sets. *Geophys. Res. Lett*. **38**, L04404 (2011).

ACCESSION CODES. Sequence data have been deposited in the NCBI Nucleotide and Sequence Read Archive (SRA) databases. The Illumina raw sequence reads appear in SRA as XXXXXXX.

ACKNOWLEDGEMENTS

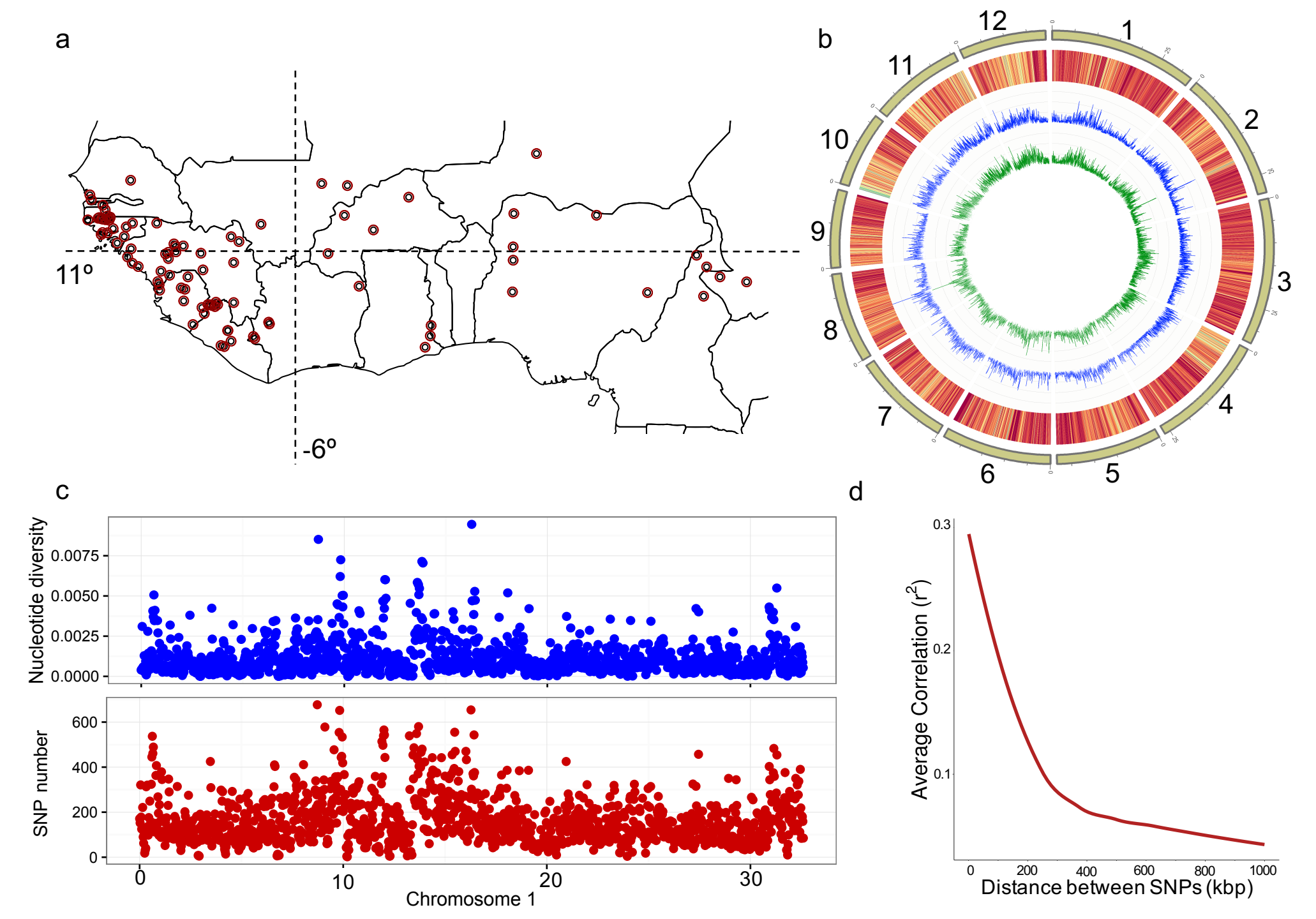
We would like to thank Endang Septiningsih for critical discussions, and Khaled Hazzouri for help in some of the analyses. We are grateful to Mamadou Sock and Bertin Fonton for field assistance, to IRRI staff for phenotyping assistance, and to Julia Maritz for lab assistance. We thank the USDA and IRRI 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 NYU Abu Dhabi Research Institute to MDP, as well as an NSF Plant Genome Postdoctoral Fellowship (IOS-1202803) to RSM.

AUTHOR CONTRIBUTIONS

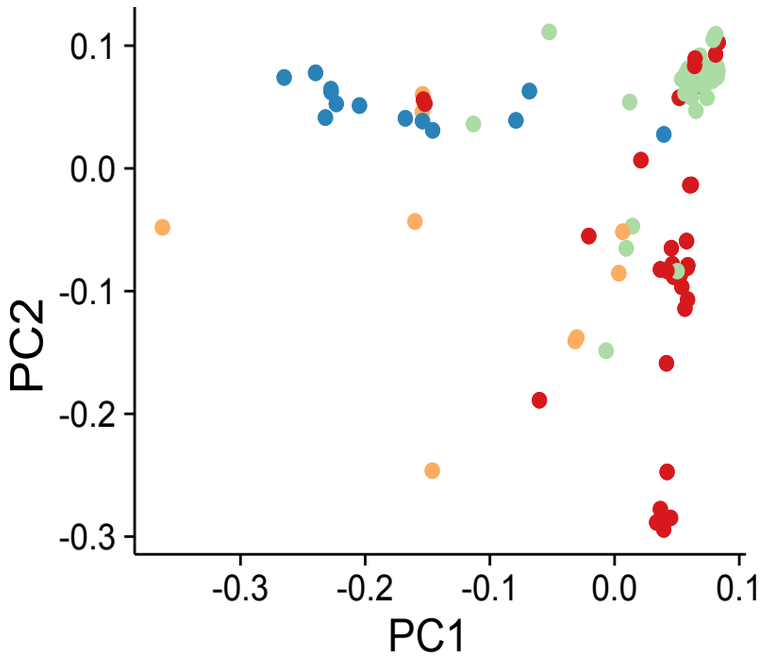
RSM, GG and MDP designed the experiments and analyses. KB and MNN helped in design and execution of the fieldwork. RSM, MS, AP, JA, AB, KD, BG, and GG collected the data. RSM, JYC, JMF and MDP analyzed the data. RSM, JYC, DQL and MDP wrote the paper.

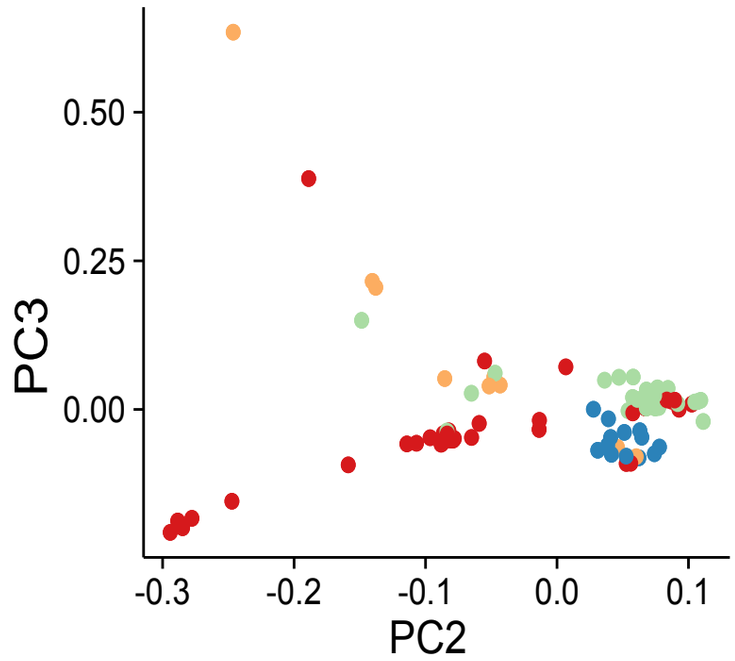
COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

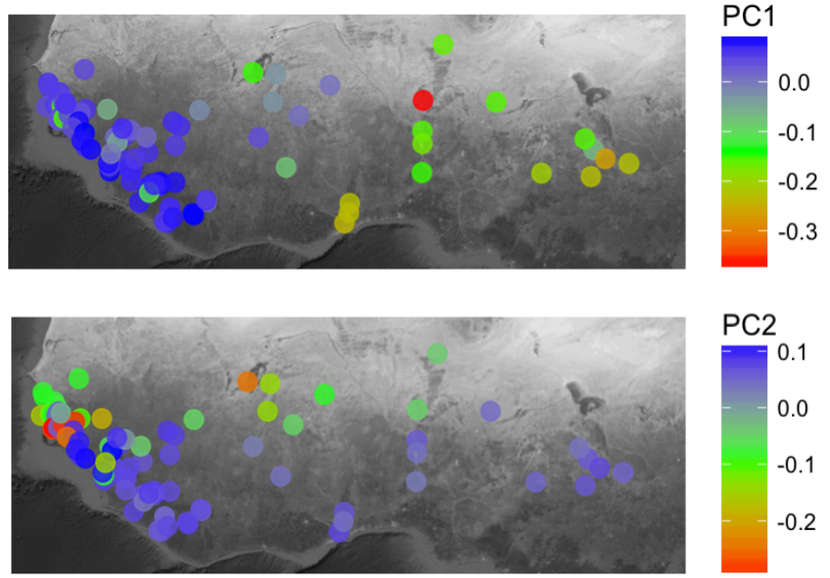

a
11°
-6°
b
12
1
2
3
4
5
6
7
8
9
10
11
c
Nucleotide diversity
0.0075
0.0050
0.0025
0.0000
SNP number
600
400
200
0
0
10
20
30
Chromosome 1
d
Average Correlation (r^2)
0.3
0.2
0.1
0
200
400
600
800
1000
Distance between SNPs (kbp)

a

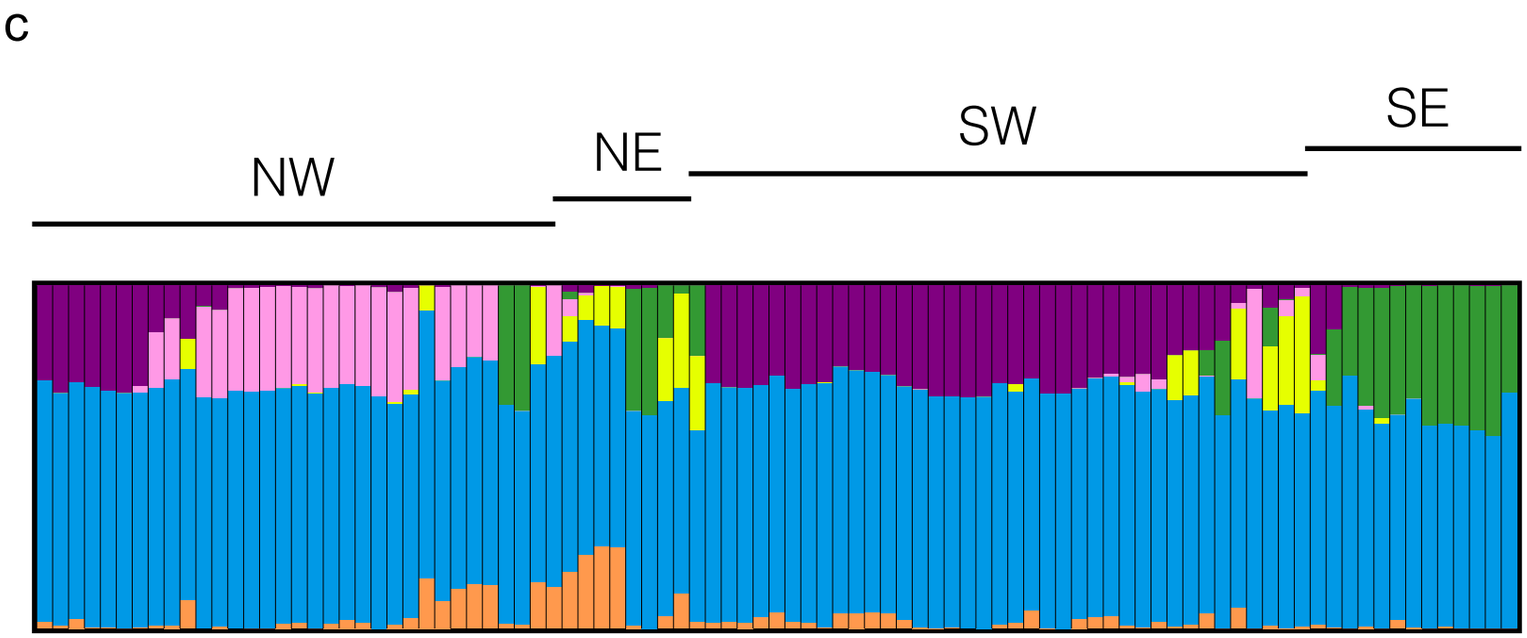

0.1
0.0
-0.1
-0.2
-0.3
PC2
-0.3
-0.2
-0.1
0.0
0.1
PC1


0.50
0.25
0.00
PC3
-0.3
-0.2
-0.1
0.0
0.1
PC2

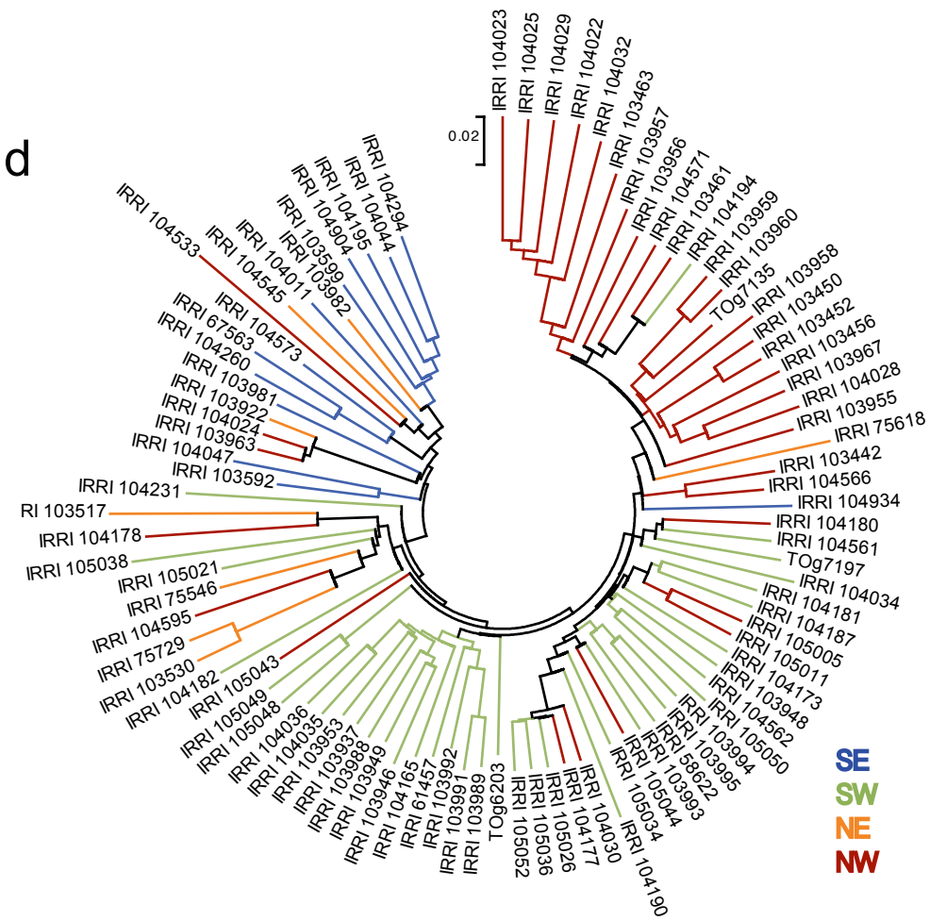
b

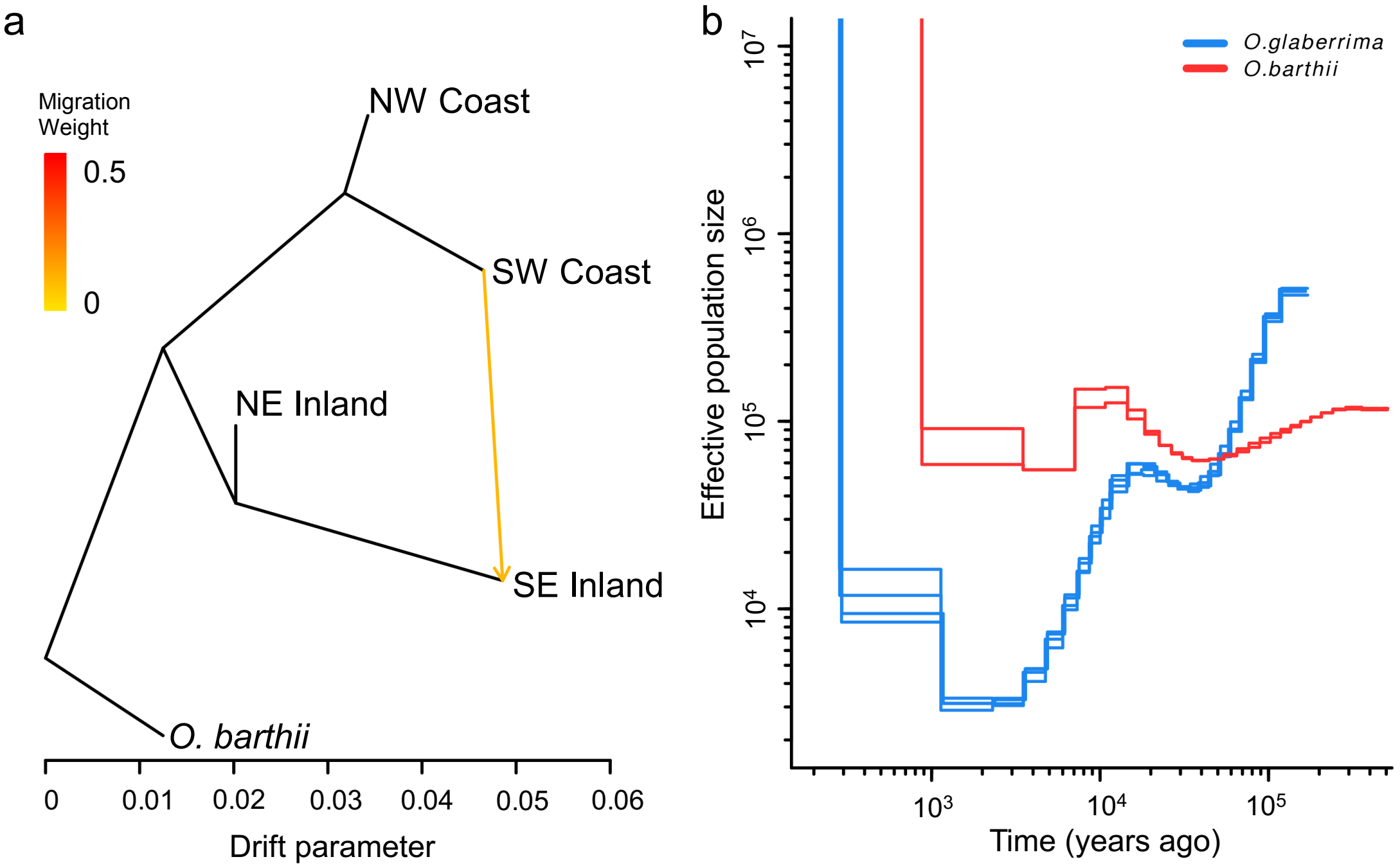

PC1
0.0
-0.1
-0.2
-0.3
PC2
0.1
0.0
-0.1
-0.2

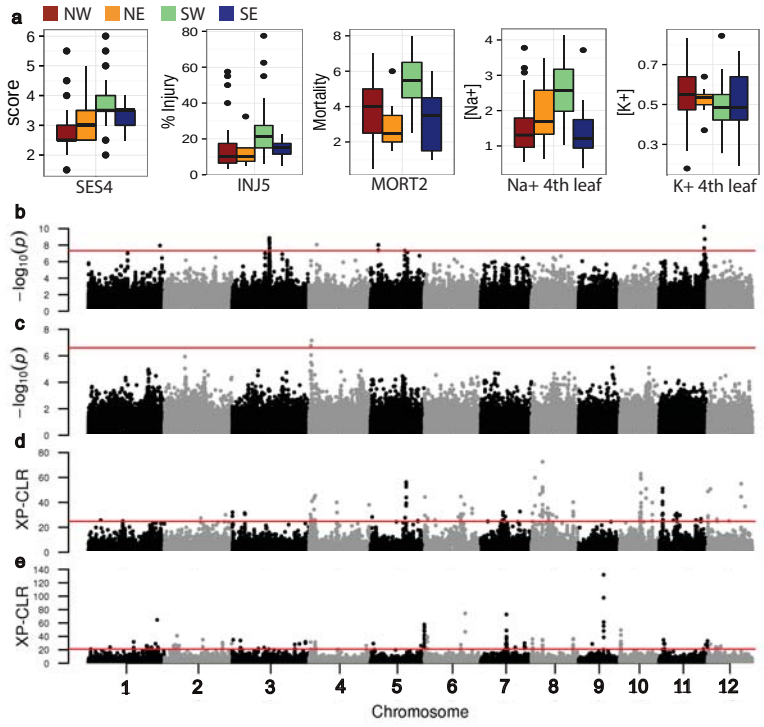
c

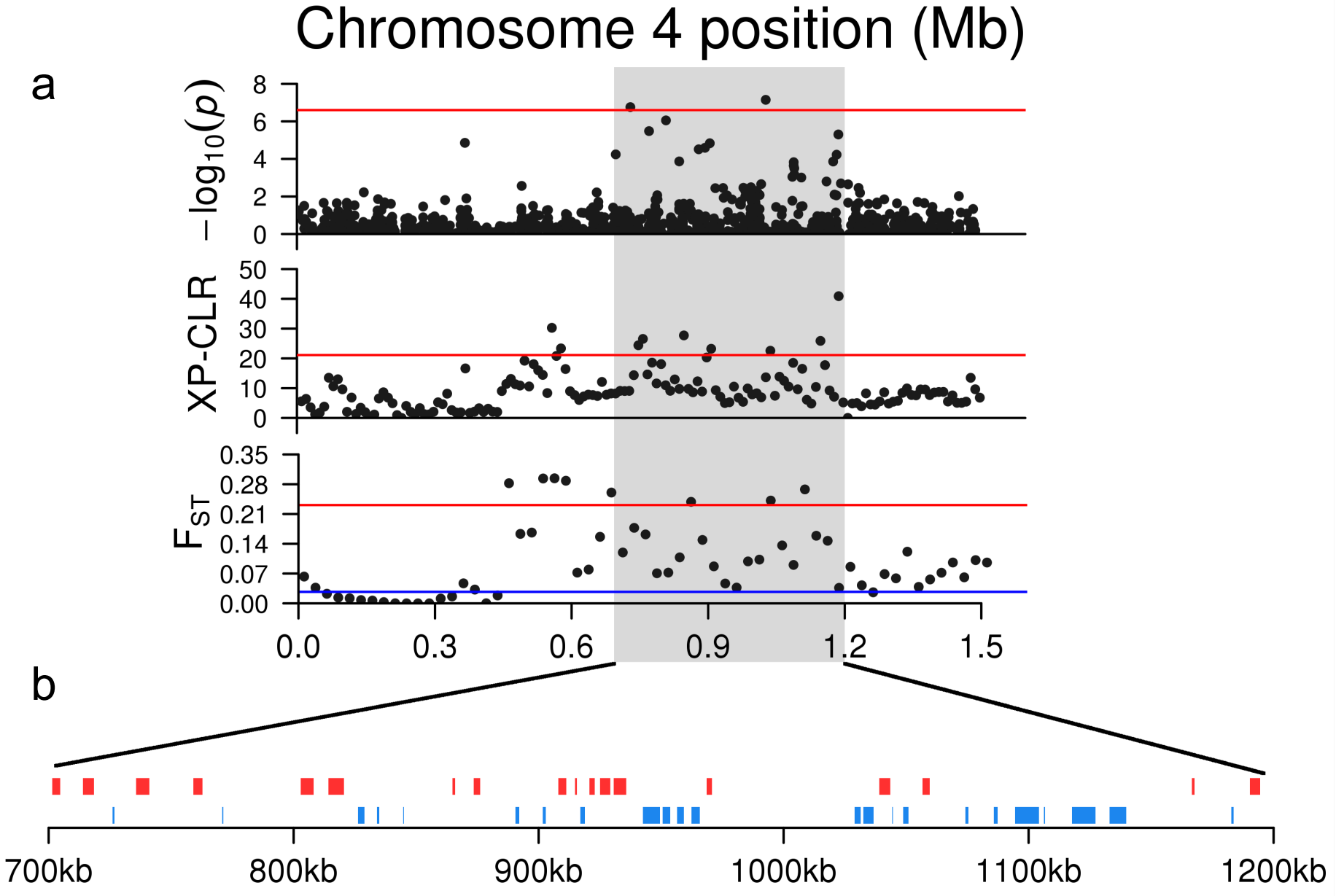

NW
NE
SW
SE

d


0.02
IRRI 104023
IRRI 104025
IRRI 104029
IRRI 104022
IRRI 104032
IRRI 103463
IRRI 103957
IRRI 103956
IRRI 104571
IRRI 103461
IRRI 104194
IRRI 103959
IRRI 103960
IRRI 103958
TOg7135
IRRI 103450
IRRI 103452
IRRI 103456
IRRI 103967
IRRI 104028
IRRI 103955
IRRI 75618
IRRI 103442
IRRI 104566
IRRI 104934
IRRI 104180
IRRI 104561
TOg7197
IRRI 104034
IRRI 104181
IRRI 104187
IRRI 105005
IRRI 105011
IRRI 104173
IRRI 103948
IRRI 104562
IRRI 105050
IRRI 103994
IRRI 103995
IRRI 58622
IRRI 103993
IRRI 105044
IRRI 105034
IRRI 104190
IRRI 104030
IRRI 104177
IRRI 105026
IRRI 105036
IRRI 105052
TOg6203
IRRI 103989
IRRI 103991
IRRI 103992
IRRI 61457
IRRI 104165
IRRI 103946
IRRI 103949
IRRI 103988
IRRI 103937
IRRI 103953
IRRI 104035
IRRI 104036
IRRI 105048
IRRI 105049
IRRI 105043
IRRI 104182
IRRI 103530
IRRI 75729
IRRI 104595
IRRI 75546
IRRI 105021
IRRI 105038
RI 103517
IRRI 104178
IRRI 104231
IRRI 103592
IRRI 103963
IRRI 104047
IRRI 104024
IRRI 103922
IRRI 103981
IRRI 104260
IRRI 67563
IRRI 104533
IRRI 104545
IRRI 104573
IRRI 104011
IRRI 103982
IRRI 103599
IRRI 104904
IRRI 104195
IRRI 104044
IRRI 104294
SE
SW
NE
NW


a
Migration Weight
0.5
0
NW Coast
SW Coast
NE Inland
SE Inland
O. barthii
0
0.01
0.02
0.03
0.04
0.05
0.06
Drift parameter
b
O.glaberrima
O.barthii
Effective population size
10^7
10^6
10^5
10^4
10^3
10^4
10^5
Time (years ago)


a
NW
NE
SW
SE
score
SES4
% Injury
INJ5
Mortality
MORT2
[Na+]
Na+ 4th leaf
[K+]
K+ 4th leaf
b
−log10(p)
c
−log10(p)
d
XP-CLR
e
XP-CLR
1
2
3
4
5
6
7
8
9
10
11
12
Chromosome


Chromosome 4 position (Mb)
a
$-\log_{10}(p)$
XP-CLR
F_{ST}
0.0
0.3
0.6
0.9
1.2
1.5
b
700kb
800kb
900kb
1000kb
1100kb
1200kb

SUPPLEMENTARY NOTE. Domestication theories for African rice.

The most commonly described theory of African rice domestication is the one proposed by French botanist Roland Portères[1,2], a specialist on grain crops. Portères contributed to the meetings on African crop domestication organized by Jack Harlan, which became collected in the book 'Origins of African Plant Domestication'[3], and illustrate the difficulty in confirming origins and timing of African rice domestication processes. This note summarizes the spatial and temporal ideas of African rice domestication based on multiple lines of scholars (botanical, archaeological, glottochronological, and cultural), that provide alternative views we have considered, that are especially important given the paucity of archaeological data in tropical and sub-tropical West Africa.

Pertinent to the timing of domestication is the severity of protraction: protraction is broad-range incipient domestication events rather than events in single specific origins at a specific time point. The archaeological findings at Jenne-Jeno that could only trace domesticated African rice to 1000 BCE[4,5] contributed to Portères rejecting the protracted model and searching for defined Centers of Origin in Africa. He proposed that African rice diffused out from a single origin in the inner Niger River Delta (IND), and that the Mande people (Mandingo) brought it West to Senegambia. He further proposed that rival steppe nomads pushed them, the grain producers, into the forest, which led to Senegambia and the Guinean forest region being called more recent 'secondary centers'.

The assertion by Portères[1,2] that food production in the forest was a recent innovation by grain producers was subsequently overturned[6], has been rejected by some scholars, stating that as a grain specialist, Portères overlooked forest product importance[7]. Some scholarship asserted that along the widespread Savannah, Sudanic, and Guinean zones, ancient people were able to live through exploitation of the environment without needing to deliberately produce food themselves[8]. Sub-Saharan Neolithic pastoralism as early as 7 kya was possible; there is some evidence of hunter-gatherer settlements in Senegal 7 kya[9]. Regardless, some scholars continued to search for a nuclear Mande center in the Sudano-Guinean forest that would suggest an acute center of origin for rice and other crops there, but not enough evidence materialized[10]. The nature of hunter-gatherer and pastoralism in West Africa led Thurnstan Shaw to report that the whole principle of sub-Saharan Centers of Origin is called into question[7], turning attention to a theory of a non-centric protracted origin of African rice.

Climate change causing desiccation also was used in development of another theory that people would have domesticated African rice on the widespread fringe of the tropical West African forest belt where habitat was transforming into savannah 4000 years ago as they could no longer rely on traditional foodsources[11]. Clark proposed that desiccation and increased population would force change. Coursey proposed separate practices of rice and tuber harvesting divided by ethnic boundary, with rice grown in the westerly Ivory Coast regions, and that proposed that people would increase reliance on rice and relax their reliance on forest tubers under climate change[12]. Whether this reliance was intensified wild harvesting or production is unknown. Grain grinding stones widespread early on but could only be associated with cultivated cereals at end of second millennium BC[13]. *Oryza barthii*, the progenitor of African rice, occurred throughout and was proposed as one of the food sources of the Sahel communities in now Senegal that did not turn to cultivation until 3 kya[13].

Despite the lack of consensus on how people in the Savannah adapted or how rice spread to new environments[14], Harlan remained firm that experimentation in uplands and floodplains was also a domestication process, and that multiple non-centric domestication centers should be considered. Some support is offered from other lines of evidence that do suggest longstanding African rice cultivation outside of the Inner Niger delta. Olga Linares used social practices and history to argue the Casamance region of Senegal had African rice by the beginning of the first millennium AD[15]. Using glottochronology, Edda Fields-Black demonstrated that in western coastal regions, both lowland coastal rice (rainfed and stream-irrigated) and rainfed upland rice was developed for over 1000 years[9]. This long timespans for rice agriculture along the coast support Harlan's multiple domestication ideas. Regardless, there is evidence that inland and coastal rice technology was not unidirectional, as proposed by Portères[1], but exchanged as ethnic groups interacted. Dry habitat grain cultivation words from Mande became 'loanwords' in Atlantic languages, and coastal habitat grain cultivation words from Atlantic languages were borrowed by the Mande[9].

While this study does not specifically address the geographic origin of domestication, the distinctness of forest, north Atlantic, and inland northern rice populations (main text Fig. 2), the early divide we observed of coastal and inland populations with possible migration happening laterally within the forest belt (main text Fig 3), and the protracted reduction in effective population size peaking at a time of high human population (main text Fig. 3) is compatible with both centric and non-centric protracted origin(s) of African rice.

References

1. Portères, R. in *Origins of African Plant Domestication* (eds. Harlan, J.R., De Wet, J.M. & Stemler, A.B.) pp. 409–452 (Mouton, The Hague, 1976).
2. Portères, R. Berceaux agricoles primaires sur le continent Africain. *J. African Hist.* 3, 197–199 (1962).
3. Harlan, J.R. and J.M.J. De Wet (eds.), Origins of African Plant Domestication (Mouton, The Hague, 1976).
4. McIntosh, R.J. and S.K. McIntosh.The Inland Niger Delta before the empire of Mali: evidence from Jenne-Jeno. *J. African Hist.* 22, 15–16 (1981).
5. Murray, S.S. in *Fields of Change: Progress in African archaeobotany* (ed. Cappers, R.T.J.) pp. 53–62 (Barkhuis & Groningen University Library, Groningen, 2007).
6. Hladik, C.M., Linares, O.F., Hladik, A., Pagezy, H. et Semple, A. in *Tropical Forests, People*

and Food: Biocultural interactions and applications to development (eds. Hladik, C.M., Hladik, A., Linares, O.F., Pagezy, H, Semple, A. & et Hadley, M.). pp. 3–14 (UNESCO-Parthenon, Paris, 1993).
7. Shaw, T. in *Origins of African Plant Domestication* (eds. Harlan, J.R., De Wet, J.M. & Stemler, A.B.) pp. 107–153 (Mouton, The Hague, 1976).
8. Mokhtar, G. (ed.) Ancient civilization of Africa: Volume 2. pp 331–333. (James Currey, London, 1990).
9. Fields-Black, E.L. Untangling the many roots of west african mangrove rice farming: rice technology in the Rio Nunez region, earliest times to *c*. 1800. *J. African Hist.* 49, 1–21 (2008).
10. Murdock, G.P. *Africa: Its Peoples and their Culture History*. pp 64–76. (New York, McGraw-Hill, 1959).
11. Clark, J.D. in *Background to Evolution in Africa*. (eds. Bishop, W.W. & J.D. Clark.) pp. 601–627 (Chicago: University of Chicago Press, 1967).
12. Coursey, D.G. *Yams*. (London: Longmans, 1967).
13. Clark, J.D. in *Origins of African Plant Domestication* (eds. Harlan, J.R., De Wet, J.M. & Stemler, A.B.) pp. 67–105 (Mouton, The Hague, 1976).
14. Harlan, J.R. Agricultural origins: centers and noncenters. *Science* 29, 468–74 (1971).
15. Olga Linares de Sapir, Shell middens of Lower Casamance and problems of Diola proto-history. *West Afr. J. Arc.* 1, 23–54 (1971).

SUPPLEMENTARY TABLE 4a. Number of farms visited where salt tolerance was discussed during interviews. Some farms had multiple people involved in the interview. In Senegal 14/33 people expressed they practiced transplanting (which could protect young seedlings from some stresses).

Country /Regions	**Total number of farms/total farmers**	**Farms affected by salt stress**	**Farms unaffected by salt stress**	**Farms with ambiguity**	**Farmer response to salt stress**	**Farmers practicing at least one of the responses to salt stress**
Togo/ Plateau	7	0	5	2	n/a	n/a
Senegal/ Saint Louis and Sine Saloum	33/33	32	1	0	Improving drainage, growing salt tolerant varieties, building small dykes, large dykes, planting trees, moving to places with less salt, transplanting	19

SUPPLEMENTARY TABLE 4b. Names of varieties farmers could name when asked what kinds of African rice they were growing within the last 3 years. Farmers were then asked which of these are tolerant to salt. In Senegal, 27% of the varieties were said to be salt tolerant.

Country	**Local variety names**	**Local salt tolerant varieties**
Togo	Café, Há, Moligba, Oletretowê, Pakbalipe, Tretomoli, Trevemoli, Winto, Winto hi, Winto ibo, Wintoshi, Wintoyibo,Yibo	n/a

Senegal	Abibaba, Ahobale, Ahouyet, Bakonda, Banyuno, De semsen, Diep buseyo, Dimba gnyima, Djiep gao, Donathe, Dootir, Lagrat, Luku, Maria masa, Mbora Tell black, Mbora Tell white, Metoro, Momo, Momocoy, Mon, Ndoukouti, Okunda, Ombendah, Omomobalé, Omomokane, Ortora, Padit, Pepermeñe, Pudar, Sam saham, Thiloo, Toubab, Tout yif, Ubale, Ubalule, Undap, Ya binta, Ya cisse, Yaka, Yehoumahk,Simtam/Sintam	Momo, Ndoukouti, Ombendah, Ortora, Pepermeñe, Sintam, Ubalule, Yahoumahk, Yaka, Mbora Tell black, Mbora Tell white,

SUPPLEMENTARY TABLE 4c. Informants in Senegal (n=33) report how many other abiotic factors (drought, flood, heat, high nutrients) affect how the plant responds to salt stress, based on their life experience. They report if the factors exacerbate or relieve the salt stress response.

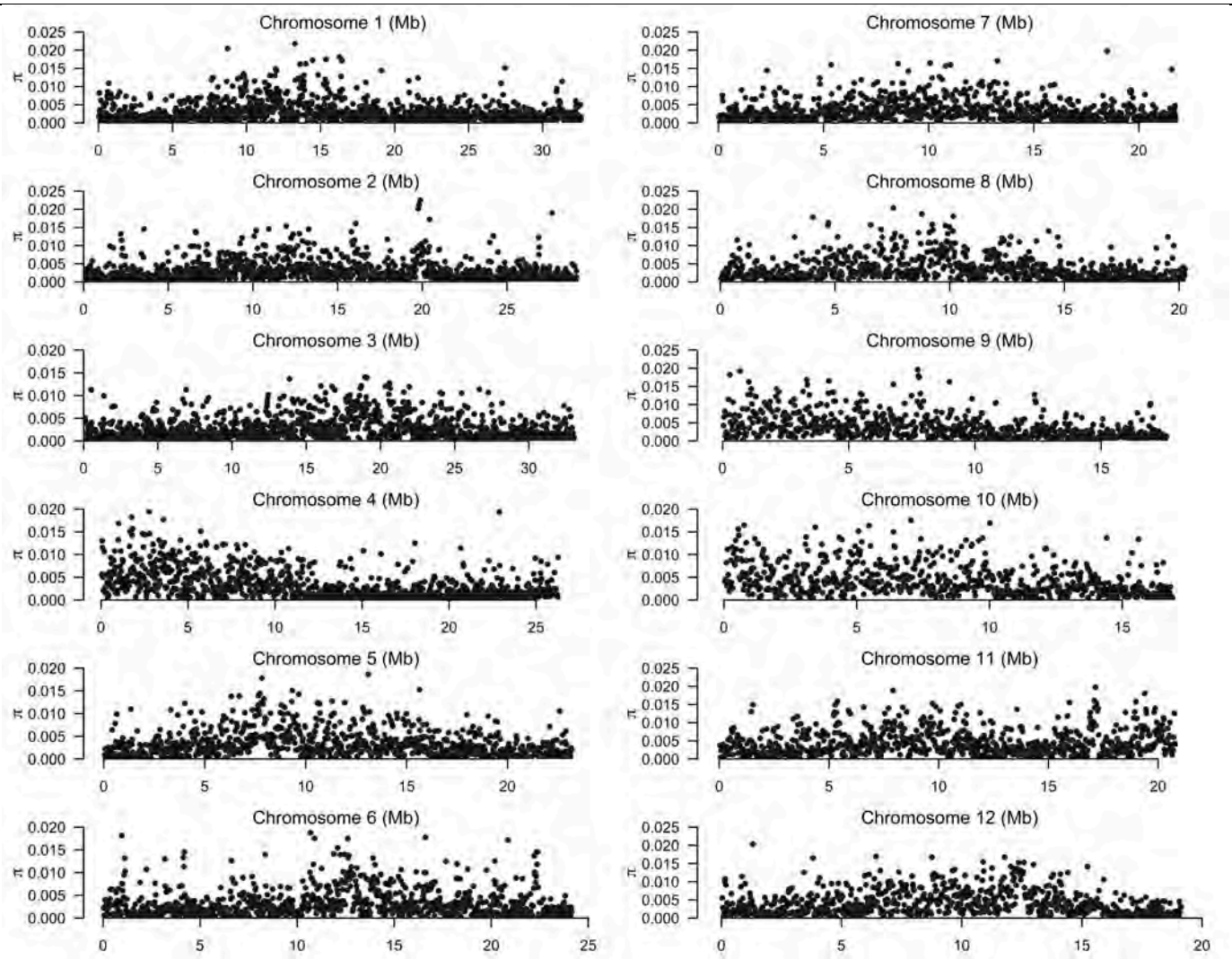
Drought worsens salt stress	**Iron worsens salt stress**	H**eat worsens salt stress**	**Fertilizer/nutrients relieves salt stress**
11	1	2	2

SUPPLEMENTARY TABLE 4d. Informants were asked at what growth stages African rice is most susceptible to salinity.

seedling stage	**heading**	**flowering**
7	4	16

SUPPLEMENTARY TABLE 11. Primers used in qRT-PCR

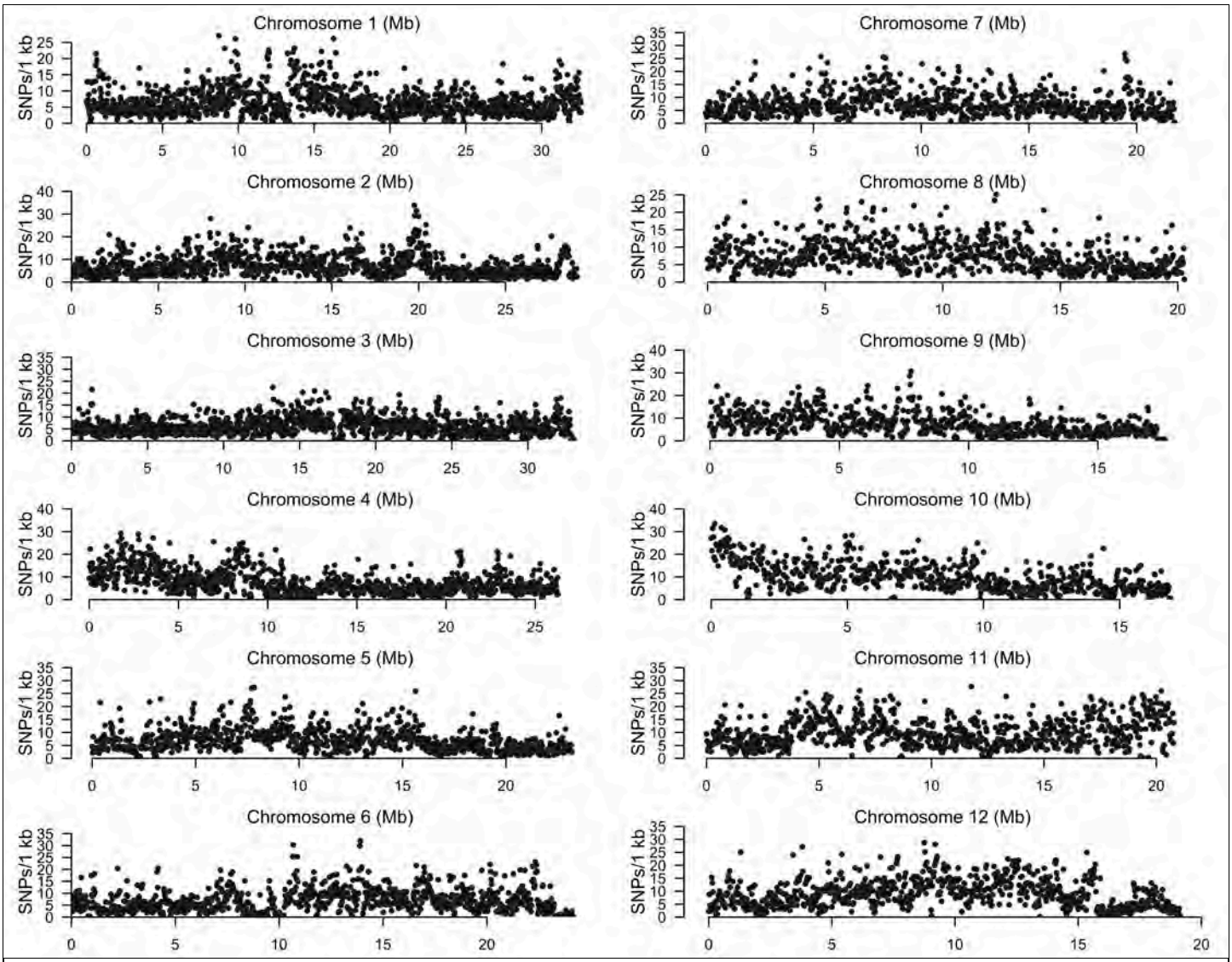
qRT_Actin11_F	CAGCCACACTGTCCCCATCTA
qRT_Actin11_R	AGCAAGGTCGAGACGAAGGA
qRT_OgUBQ10_F	TGGTCAGTAACCAGCCAGTTTGG
qRT_OgUBQ10_R	GCACCACAAATACTTGACGAACAG
qRT_PPI_F1	GCGACCAGATCCTCTCC
qRT_PPI_R1	CGAAGGGTTTCTGCATCT
qRT_HAK6_F	ATCTCTGCAAGCTACTCCA
qRT_HAK6_R	AACACACACGACCATCAA
qRT_HAK5_F	TTTCTGTAGTTCTCCTTCCTT
qRT_HAK5_R	TGTATCTGCAACGTGTTCT



Supplementary Figure 1

Nucleotide diversity (π) across the genome

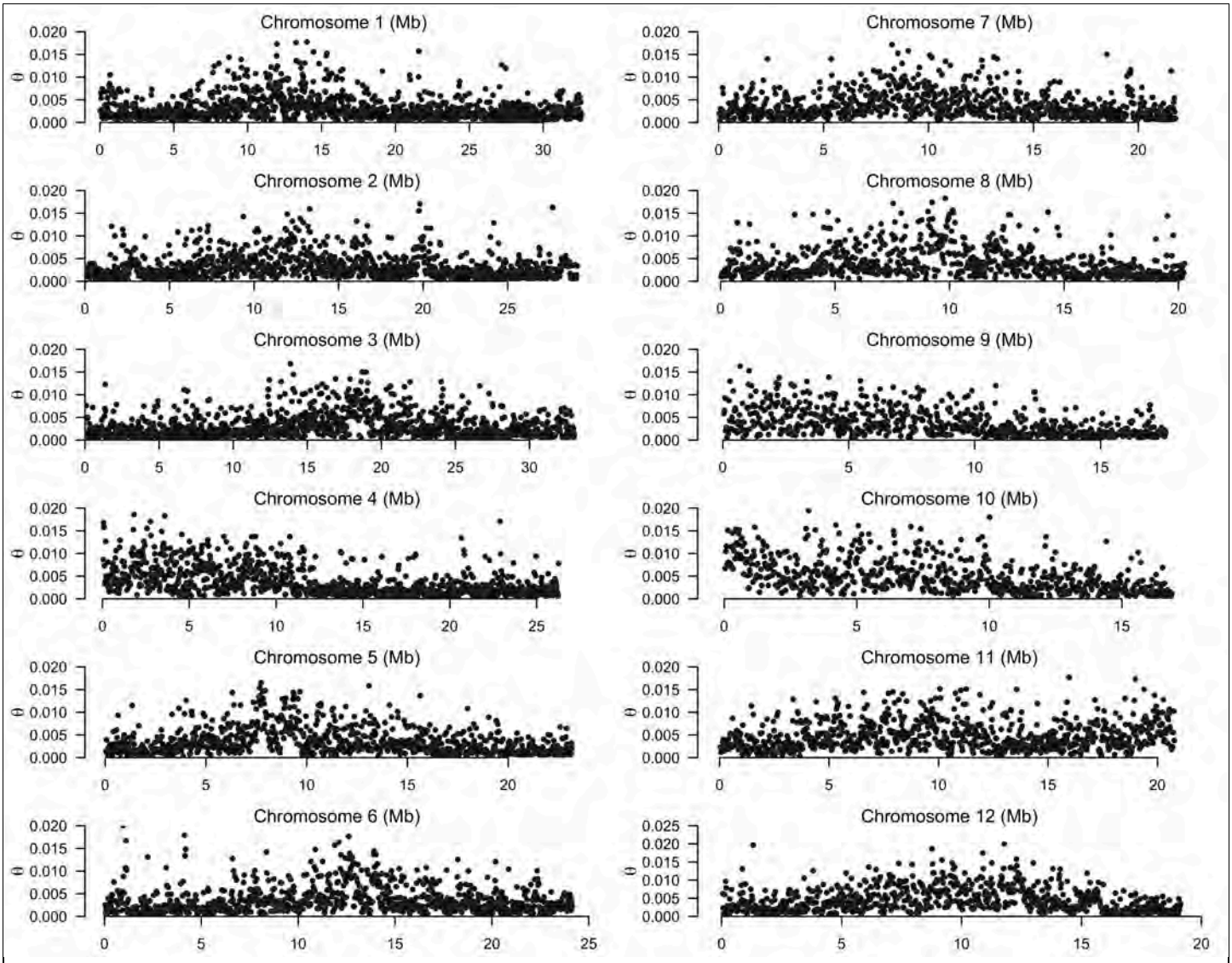
Plots of nucleotide diversity was calculated in 25-kb windows across the 12 African rice chromosomes.



Supplementary Figure 2

SNP density across the genome

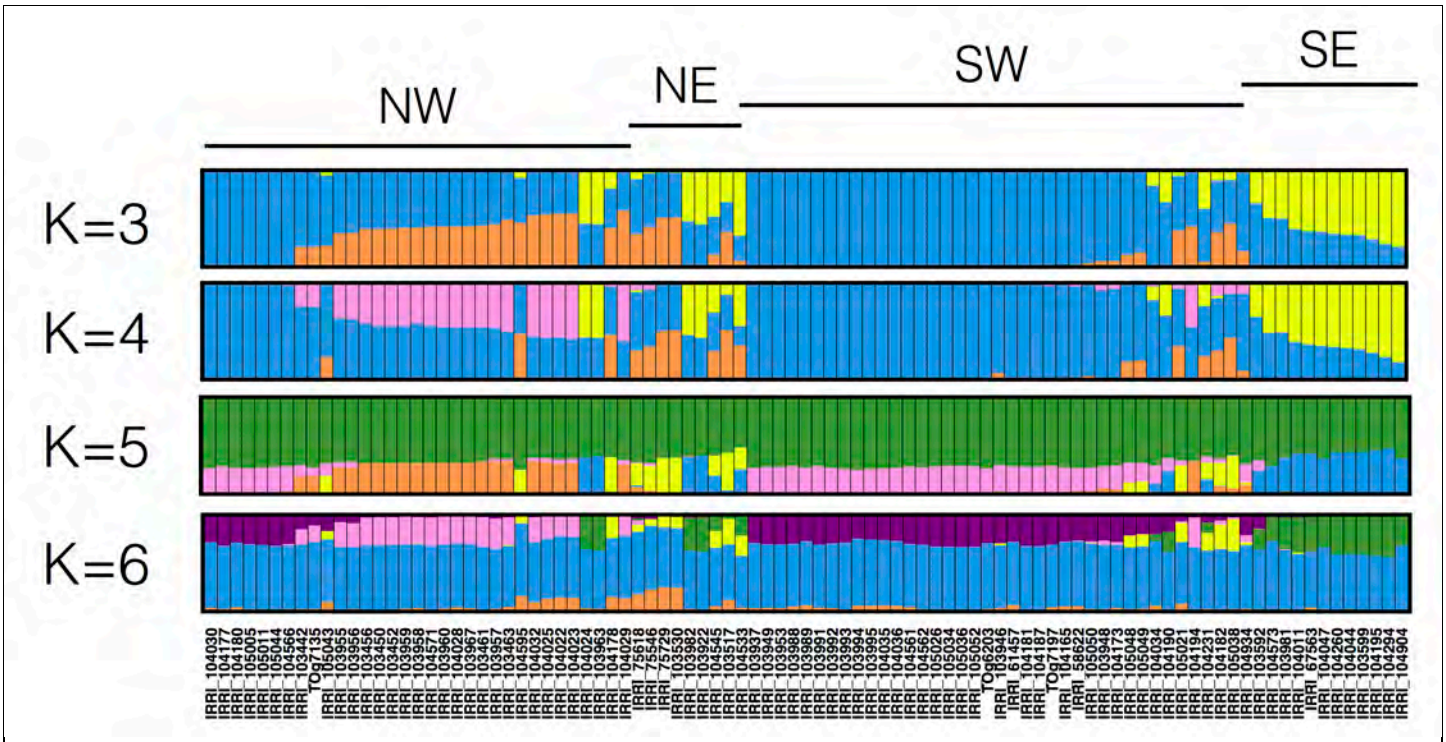
Plots of SNP density was calculated in 25-kb windows across the 12 African rice chromosomes.



Supplementary Figure 3

Population mutation parameter (θ_w) across the genome

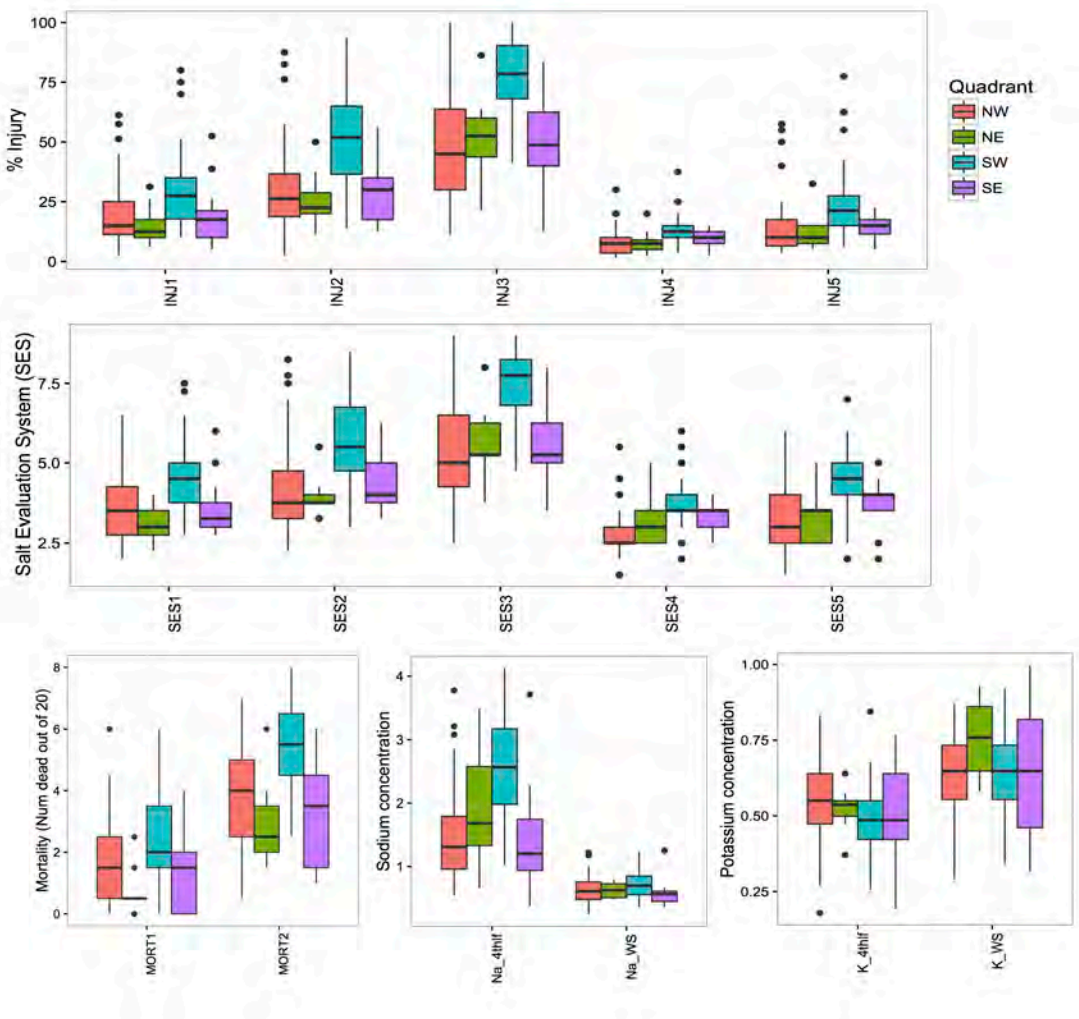
Plots of Watterson's theta calculated in 25-kb windows across the 12 African rice chromosomes.



Supplementary Figure 4

STRUCTURE results for African rice

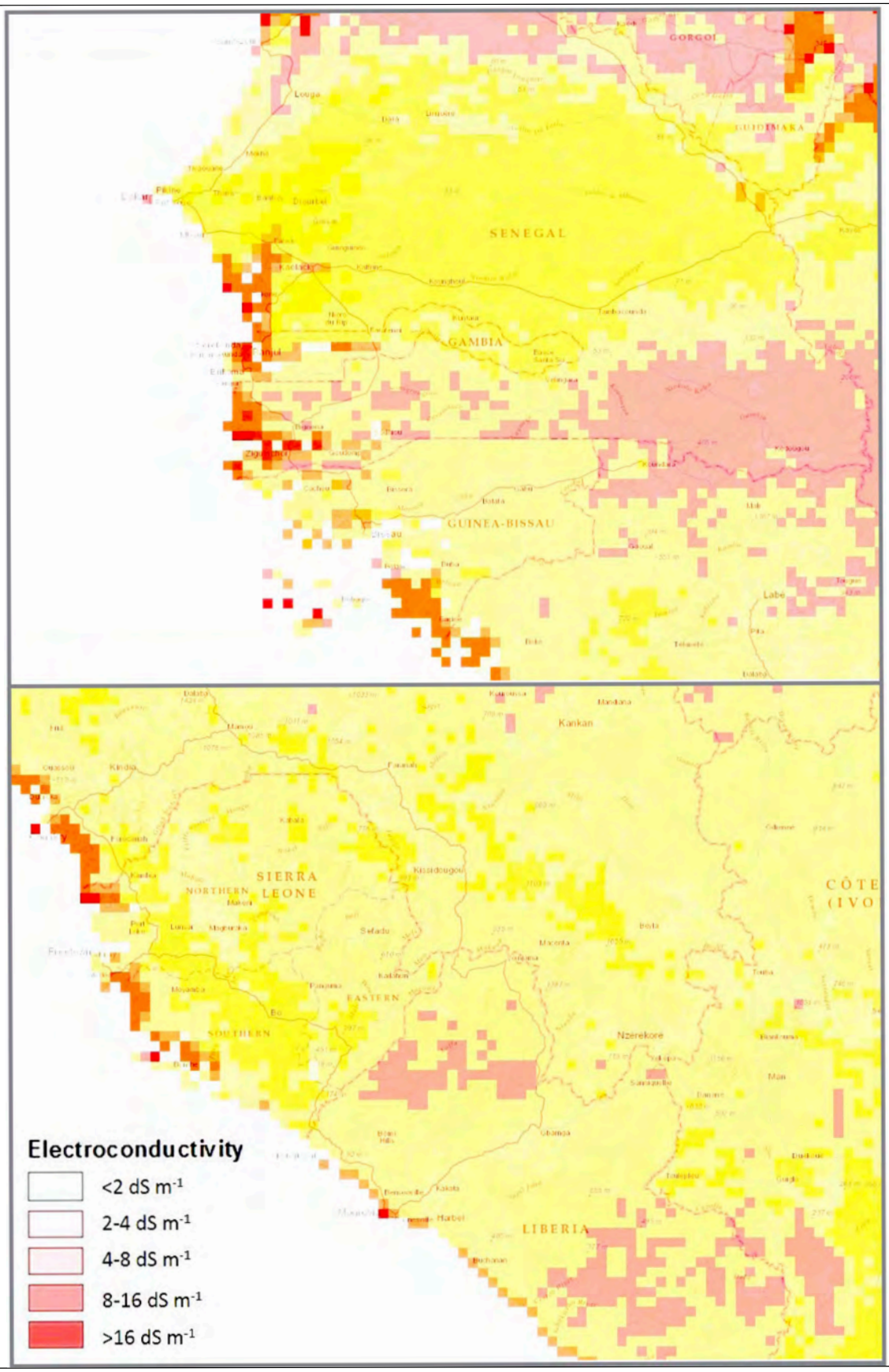
STRUCTURE results for four different numbers of populations (k). Accessions are sorted by geographic quadrant of the accession collection origins.



Supplementary Figure 5

Salt tolerance phenotypes of each West African population

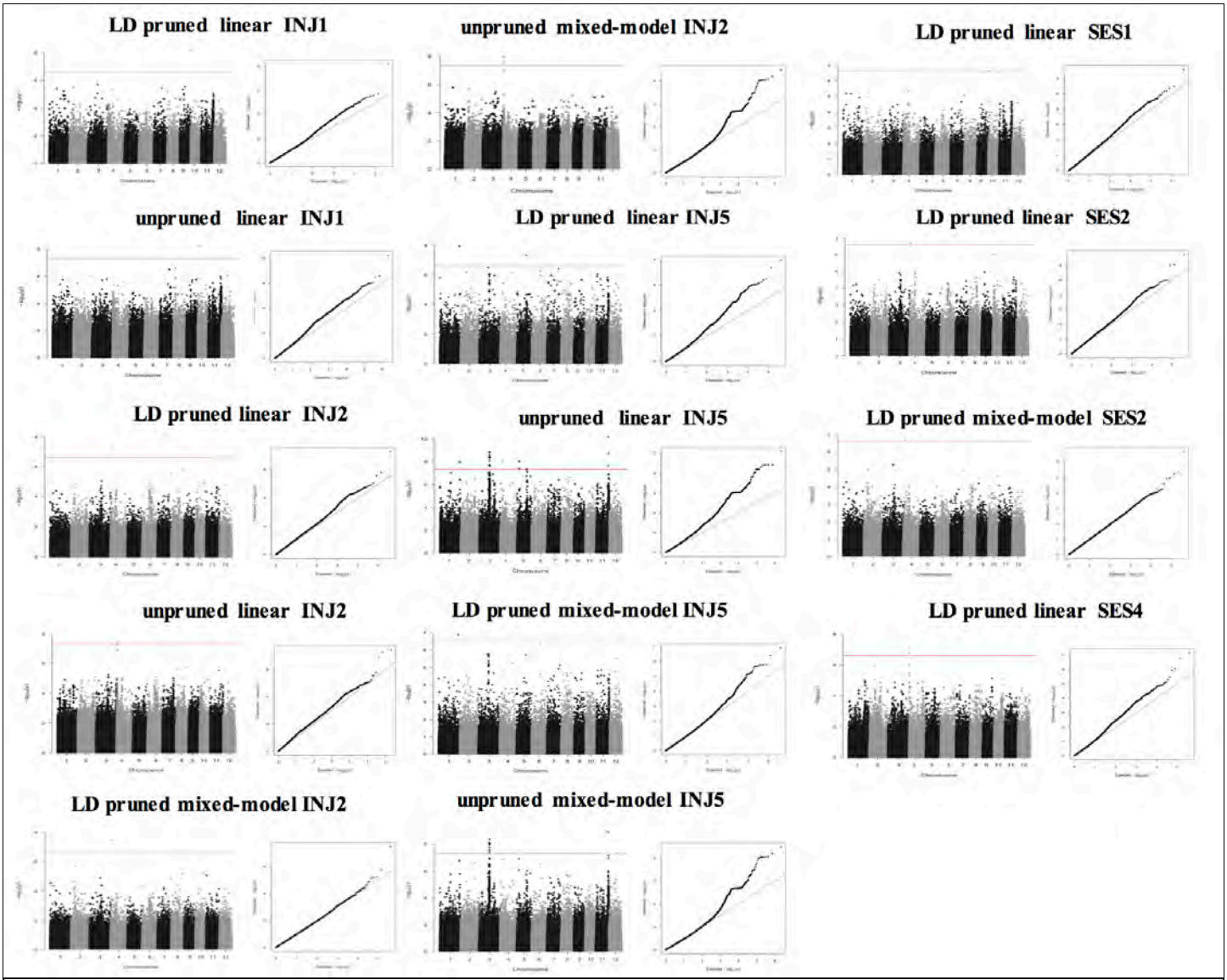
Box and whisker diagrams of salt stress phenotype values within accessions grouped by geographic quadrant.


SENEGAL
GAMBIA
GUINEA-BISSAU
SIERRA LEONE
LIBERIA
Kankan
Nzerekore
Electroconductivity
<2 dS m-1
2-4 dS m-1
4-8 dS m-1
8-16 dS m-1
>16 dS m-1

Supplementary Figure 6

Salinity maps of West African regions

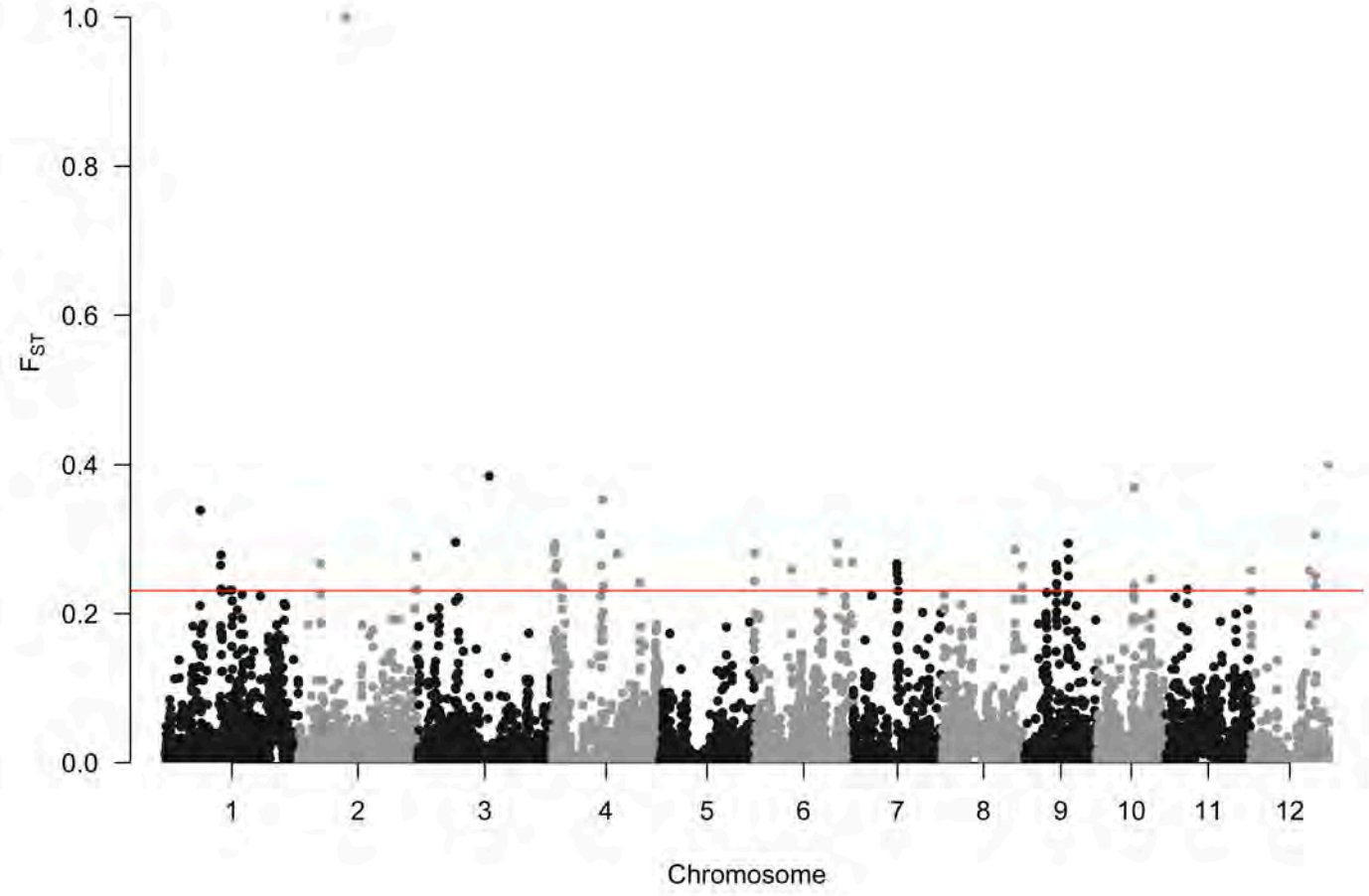
Bright yellow areas are land used for crop cultivation. In the arid Northwest (top) that spans Senegal and Guinea-Bissau, soil salinity is present far inland, and can be traced along rivers that connect to the Atlantic Ocean. Along the sub-tropical and tropical coast (bottom), salinity decreases from North to South along the coast, and inland water tables are less saline than in the arid region. No apparent high salinity levels, that would be associated with the presence of coastal rivers, exist inland.



Supplementary Figure 7

GWAS results for salt tolerance phenotypes in African rice

Manhattan and quantile-quantile plots of GWAS results found to have Bonferroni-significant associations. Red lines in Manhattan plots signify the Bonferroni threshold. Red lines in quantile-quantile plots signify concordance between observed and expected associations. Trait codes are in Supplementary Table 5.



Supplementary Figure 8

Genome-wide F_{ST} between NW and SW coastal populations

Plots of mean F_{ST} in 25-kb SNP windows across the 12 chromosomes for NW versus SW populations. The horizontal line indicates the 0.5% threshold for the outlier test.