ARTICLES
https://doi.org/10.1038/s41477-020-0659-6

nature plants

Check for updates

Genomic history and ecology of the geographic spread of rice

Rafal M. Gutaker[1], Simon C. Groen[1], Emily S. Bellis[2], Jae Y. Choi[1], Inês S. Pires[1,3], R. Kyle Bocinsky[4], Emma R. Slayton[5], Olivia Wilkins[1,6], Cristina C. Castillo[7,8], Sónia Negrão[9], M. Margarida Oliveira[3], Dorian Q. Fuller[7,8], Jade A. d'Alpoim Guedes[10], Jesse R. Lasky[2]✉ and Michael D. Purugganan[1,11]✉

Rice (*Oryza sativa*) is one of the world's most important food crops, and is comprised largely of *japonica* and *indica* subspecies. Here, we reconstruct the history of rice dispersal in Asia using whole-genome sequences of more than 1,400 landraces, coupled with geographic, environmental, archaeobotanical and paleoclimate data. Originating around 9,000 yr ago in the Yangtze Valley, rice diversified into temperate and tropical *japonica* rice during a global cooling event about 4,200 yr ago. Soon after, tropical *japonica* rice reached Southeast Asia, where it rapidly diversified, starting about 2,500 yr BP. The history of *indica* rice dispersal appears more complicated, moving into China around 2,000 yr BP. We also identify extrinsic factors that influence genome diversity, with temperature being a leading abiotic factor. Reconstructing the dispersal history of rice and its climatic correlates may help identify genetic adaptations associated with the spread of a key domesticated species.

The domestication of crop species marks a major transition in human–plant interaction, and has been responsible for the shift of humans from a hunter-gatherer to an agricultural species. There are about 24 areas in the world from which crop species originated, and attention has focused on the dynamics of the domestication process and the evolutionary genetics of crop origins and divergence[1]. By contrast, relatively little attention has been focused on the dispersal and diversification of crops from their centres of origin, and the accompanying evolution of adaptive traits that enable these domesticated species to establish themselves in different environmental and cultural contexts[2]. Reconstructing the patterns and timing of the spread of domesticated species can help us understand the climatic and other environmental factors that govern the expansion of their species range, as well as the relationship between crop dispersal and human migration and history.

Rice (*Oryza sativa* L.) is a major staple crop, providing more than 20% of calories for more than half of the human population. Domesticated rice encompasses genetically distinct populations grown in sympatry, including major subgroups *japonica* and *indica* (sometimes recognized as subspecies), as well as geographically more restricted *circum*-aus, and *circum*-basmati rices[3,4]. It is mainly cultivated in monsoon Asia, but rice is distributed across a wide latitudinal range, spanning tropical and temperate zones of Asia, probably requiring local water, temperature and photoperiod adaptation. Rice is grown in lowland ecosystems under paddy, deepwater or seasonal flood conditions, as well as in upland rainfed areas[5].

Archaeological evidence[6–8] indicates that cultivation of *japonica* rice began around 9,000 yr BP in the lower Yangtze Valley, whereas proto-*indica* rice cultivation started more than 5,000 yr BP in the lower Ganges valley[9]. Archaeological[10] and most population-genetic analyses[11–13] suggest that important domestication alleles have a single origin in *japonica* rice in East Asia. The spread of *japonica* to South Asia about 4,000 yr BP led to introgression of domestication alleles into proto-*indica* or local *Oryza nivara* populations and the emergence of *indica* rice[11–13].

While the origins of rice have been the focus of intensive study, less attention has been paid to its spread after domestication. From the Yangtze and Ganges Valleys, respectively, *japonica* and *indica* dispersed across much of Asia over the last five millennia, providing sustenance for emerging Neolithic communities in East, Southeast and South Asia[14]. Archaeological data show the general directionality of rice dispersal[9,15]; however, the details of dispersal routes and times and the environmental forces that shaped dispersal patterns remain unknown. In this study we undertake population-genomic analyses to examine environmental factors associated with the geographic distribution of rice diversity, and reconstruct the ancient dispersal of rice in Asia. Together with archaeobotanical, paleoclimatic and historical data, genomic data allow a robust reconstruction of the dispersal history of *O. sativa*.

Results

Structure of rice genomic diversity. To investigate the pattern and timing of dispersal of rice, we obtained whole-genome resequencing data from rice landraces or traditional varieties across a wide geographical distribution in Asia. Landraces, unlike elite cultivars, are associated with sustained cultivation in specific geographic localities and cultural contexts, and usually exhibit local adaptations. Our sample set includes 1,265 samples from the Rice 3 K Genome Project[3,16] and an additional 178 landraces sequenced for this study (Supplementary Table 1). The panel consists of 833 *indica*,

[1]Center for Genomics and Systems Biology, New York University, New York, NY, USA. [2]Department of Biology, Pennsylvania State University, University Park, PA, USA. [3]Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa, Lisbon, Portugal. [4]Crow Canyon Archaeological Center, Cortez, CO, USA. [5]Carnegie Mellon University Libraries, Pittsburgh, PA, USA. [6]Department of Biological Sciences, University of Manitoba, Winnipeg, Manitoba, Canada. [7]Institute of Archaeology, University College London, London, United Kingdom. [8]School of Cultural Heritage, North-West University, Xi'an, China. [9]School of Biology and Environmental Science, University College Dublin, Dublin, Ireland. [10]Department of Anthropology and Scripps Institution of Oceanography, University of California, San Diego, CA, USA. [11]Institute for the Study of the Ancient World, New York University, New York, NY, USA. ✉e-mail: lasky@psu.edu; mp132@nyu.edu

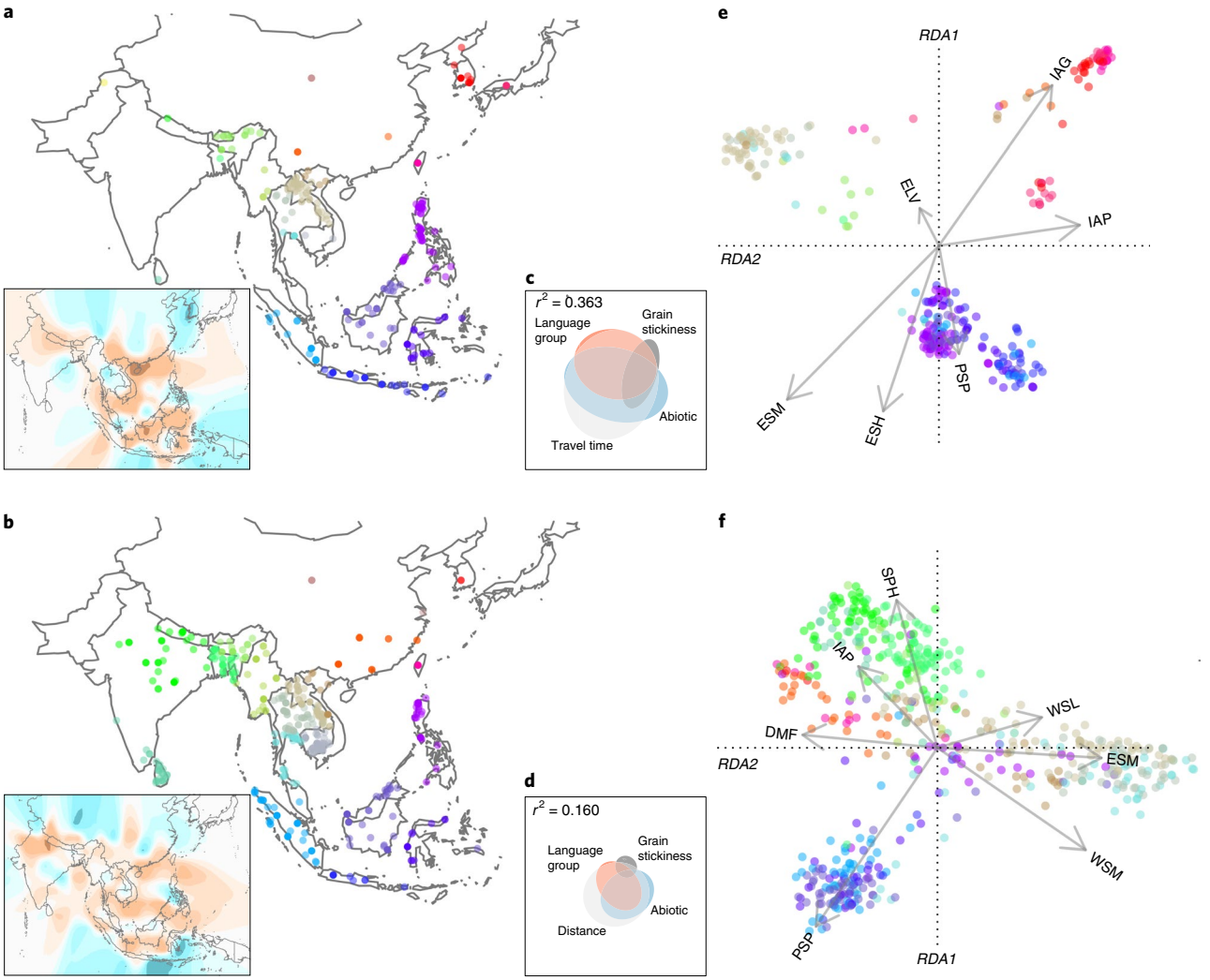

Fig. 1 | Factors underlying geographic distribution of genomic diversity in *japonica* and *indica*. **a**,**b**, Maps of collection sites for *japonica* (**a**) and *indica* (**b**) landraces used in this study. Colours represent regions of origin. Insets show effective migration surfaces representing migration barriers (orange) and channels (cyan). **c**,**d**, Genomic diversity is best explained by a combination of four factors represented in Euler Plots for *japonica* (**c**) and *indica* (**d**): travel time (migration resistance) or geographic distance, abiotic variables (temperature, moisture and soil characteristics), linguistic group and culinary properties (stickiness). Fields of squares represent total genomic variation, while elliptic shapes represent genomic variation explained by a particular group of variables calculated using variance partitioning with RDA ordination (*japonica*, $n = 317$; *indica*, $n = 656$). **e**,**f**, Genotypes of *japonica* (**e**) and *indica* (**f**) projected on the first two canonical axes of RDA. Arrows represent environmental predictors that strongly correlate with a maximal proportion of variation in linear combinations of SNPs. IAG, interannual coefficient of GDD variation; ESH, end of growing season heat; ELV, elevation; PSP, pre-growing season precipitation; ESM, early growing season minimum temperature; IAP, interannual coefficient of precipitation variation; DMF, distance to major freshwater source; WSM, whole growing-season minimum temperature; WSL, whole growing-season length; SPH, soil pH.

372 *japonica*, 165 *circum*-aus, 42 *circum*-basmati and 31 unclassified samples. We identified around 9.78 million single nucleotide polymorphisms (SNPs) with 9.63× mean coverage (s.d. = 5.03×), which we used in subsequent analyses (Supplementary Fig. 1).

Analysis of molecular variance (AMOVA) indicated that subspecies affiliation explained more than 36% of the total variation (AMOVA, permutation $P < 0.001$)[17], congruent with results from multidimensional scaling of genomic distances (Supplementary Fig. 2a). Only *japonica* and *indica* have wide geographic distributions (Fig. 1a,b and Supplementary Fig. 3), and AMOVA of these two subspecies ($n = 1{,}205$) revealed that genomic variance is explained by subspecies ($r^2 = 0.32$, permutation $P < 0.001$), country of origin ($r^2 = 0.11$, $P < 0.001$) and their interaction ($r^2 = 0.06$, $P < 0.001$). Landraces with mixed ancestry ($n = 154$) were excluded using silhouette scores[18] (Supplementary Fig. 2b); hereafter, we analysed these two subspecies independently.

We find support for isolation-by-distance (IBD) in *japonica* ($r^2 = 0.294$, $P < 0.001$) and *indica* ($r^2 = 0.265$, $P < 0.001$) (Supplementary Fig. 4). Geographic distance provides much lower explanation for genetic distance in the Malay Archipelago (that is, the islands of Southeast Asia) compared with mainland Asia, suggesting a stronger effect of local migration barriers on archipelago IBD (Supplementary Fig. 5). Effective migration surfaces[19] identified geographic barriers for dispersal over the Himalayan and Hengduan Mountains which separate China from South and Southeast Asia, respectively (with the caveat of sparse sampling north of the Himalayas), and the South China Sea, which reduces movement between Borneo and Philippines and mainland Southeast Asia (Fig. 1a,b and Supplementary Fig. 6).

To improve on the IBD model, we took into account actual barriers to travel between locations rather than simple geographic distances; for human-dispersed species such as crops, genetic

distances may correlate better with travel resistance, which is meant to capture cost in time and effort for human migration. Indeed, some migration barriers for rice coincide with those for humans[20]. An isolation-by-resistance model, using estimated human-associated land and marine travel times[21], is a better explanation than the IBD model for *japonica* landrace genetic distances based on Akaike information criterion (AIC) (archipelago $\Delta AIC = -34$, mainland $\Delta AIC = -17$), but not for *indica* (archipelago $\Delta AIC = +51$, mainland $\Delta AIC = +611$) (Supplementary Fig. 5).

Factors associated with spatial genomic structure. We used redundancy analysis (RDA) to partition genomic variance[22] associated with 22 different variables that include climatic and edaphic conditions, as well as interactions with humans and wild relatives (Supplementary Table 1). We assume that, while environments in localities fluctuate over time, current genome diversity may be determined both by current environment as well as long-term evolutionary history. SNP variation is better explained by our predictors for *japonica* (adjusted $r^2 = 0.363$; Fig. 1c) than *indica* (adjusted $r^2 = 0.164$; Fig. 1d). Associations between predictor sets and SNPs are substantially collinear with each other. For *japonica* and *indica*, travel time and geographic distance, respectively, explain most SNP variation (adjusted $r^2 = 0.326$ and $r^2 = 0.146$), followed by abiotic conditions, language groups (as proxy for cultural preferences associated with language barriers), grain stickiness (as proxy for conscious cultural preferences specifically) and genetic composition of proximal wild-rice populations (Fig. 1c,d and Supplementary Fig. 7). Among abiotic variables for *japonica*, temperature explains the greatest portion of SNP variation (adjusted $r^2 = 0.180$), followed by moisture ($r^2 = 0.086$) and soil characteristics ($r^2 = 0.081$). Similarly, temperature explains the most SNP variation in *indica* ($r^2 = 0.064$), followed by soil characteristics ($r^2 = 0.038$) and moisture ($r^2 = 0.036$) (Supplementary Fig. 7), although these factors have weaker explanatory power in *indica* compared with *japonica*.

The first two RDA axes of environment-associated SNP variation[23,24] separated *japonica* landraces consistent with geography (Fig. 1e), recapitulating results using total SNP variation (Supplementary Fig. 8a). Temperate *japonica* landraces from northern latitudes are most strongly identified by alleles associated with high coefficient of interannual variation in growing degree days (GDD) and low minimum temperatures early in the growing season (Fig. 1e and Supplementary Fig. 9a). Temperate landraces from upland rainfed ecosystems are further characterized by alleles associated with interannual variation in precipitation (Fig. 1e).

For *indica*, the first two axes also grouped individuals by their geographic origins (Fig. 1f and Supplementary Fig. 8b). Similar to *japonica*, *indica* landraces from the Malay Archipelago contain alleles associated with high precipitation prior to the growing season. Mainland Southeast Asian genotypes are characterized by alleles associated with warm minimum growing season temperatures and presence of nearby freshwater sources (Fig. 1f and Supplementary Fig. 9b). The associations with presence of nearby freshwater sources are in contrasts to *indica* from China and most of India, where irrigation is common and there is less reliance on natural water sources[25] (Supplementary Table 1). Finally, genotypes in south India are identified by alleles associated with interannual variation in precipitation.

Discrete subpopulations within *japonica* and *indica*. To model rice dispersal patterns, we first had to identify distinct geographical populations of *O. sativa*. To accomplish this, we clustered landraces on the basis of genomic distances by partitioning around medoids (PAM)[26], identifying the number of subpopulations (k) and subsequently applied a silhouette-based procedure (see Methods) to identify the number of discrete subpopulations (k_d). This discretization procedure removed genetic gradients between subpopulations (Fig. 2a,b and Supplementary Figs. 10 and 11). We compared PAM clusters with those from the ADMIXTURE algorithm[27]. Silhouette filtering removed individuals with spurious subpopulation assignments (Supplementary Figs. 12 and 13). In general, the clustering fit using silhouette scores is greater for *japonica* than *indica* (Supplementary Fig. 14). We find consistently higher fixation index (F_{ST}) values among *japonica* subpopulations (Supplementary Fig. 15), suggesting fewer past migrations compared with *indica* and/or older establishment of its population structure. Finally, subpopulations of both subspecies clearly correspond with geography (Fig. 2c,d and Supplementary Figs. 10 and 11), suggesting that contemporary rice landraces retain genomic signals of past dispersal across Asia.

Relationships between *japonica* subpopulations. To examine the pattern of rice dispersal, we modelled subpopulation relationships using the admixture-graph framework[28]. We used discrete subpopulations to reconstruct graphs representing ancient relationships between subpopulations that are not affected by allele frequency shifts from more recent migrations and admixtures, and we analysed *japonica* and *indica* separately.

We reconstructed relationships between *japonica* subpopulations at $k_d = 2$ to 9, considering graphs with population f-statistic z-scores below 3. Throughout all k_d levels, we find two similar and consistent graph topologies (Fig. 2e and Supplementary Fig. 16), which we used to infer dispersal routes of *japonica*. As expected[3,4], at $k_d = 2$ we observe divergence between lowland temperate varieties in northeast Asia (Korea, Japan, China and Taiwan) and tropical varieties from the Malay Archipelago (Malaysia, Philippines and Indonesia). At $k_d = 3$, we find a major lineage of tropical upland *japonica* in mainland Southeast Asia as sister group to Malay Archipelago landraces or from admixture with an ancestral temperate lineage (Supplementary Figs. 10 and 16). At higher k, these mainland Southeast Asian upland landraces always incorporate admixture from an ancestral temperate *japonica* population (see below).

At $k_d = 4$ we observe separation of primarily Indonesian from Philippine and Bornean landraces. Subsequently, at $k_d = 5$, upland temperate *japonica* in northeast Asia emerges as an admixture between lowland temperate and upland tropical varieties. Further increase of k_d enables separation of distinct Malay Archipelago subpopulations: a small subpopulation associated with the Philippines splits first, followed by a subpopulation in the Indonesian island of Java. Subsequent divisions among Malay Archipelago subpopulations are not fully resolved (Supplementary Fig. 16). Nevertheless, at $k_d = 8$, we identify a Bhutanese subpopulation closely related to upland Laotian landraces, that may represent a relict descendant population of the first early split in tropical *japonica*.

The rise of temperate *japonica*. Combining genomic, geographic, archaeological and paleoenvironmental data, we reconstructed routes and timing of the ancient dispersal of rice in Asia. *Japonica* represents the first domesticated *O. sativa*[11–13], and its tropical form was cultivated in eastern China between the Yangtze and the Huang He (Yellow) river valleys[15]. This occurred during the Holocene Climate Optimum (HCO), a period of increased monsoon activity and warmer temperatures between approximately 9,000 and 4,000 yr BP[29,30]; this coincides with the rise in frequency of non-shattering rice from around 20% just after 8,000 yr BP to fixation at about 5,000 yr BP[7,8].

The first major population divergence in *japonica* separates temperate from tropical landraces (Supplementary Figs. 10 and 16). Using sequentially Markovian coalescence (SMC++), we estimated a cross-coalescence split time between temperate and tropical *japonica* at approximately 5,000 to 1,500 yr BP, with 75% of estimates between approximately 4,100 to 2,500 years ago (Fig. 3a and Supplementary Fig. 17). Using dated archaeobotanical rice

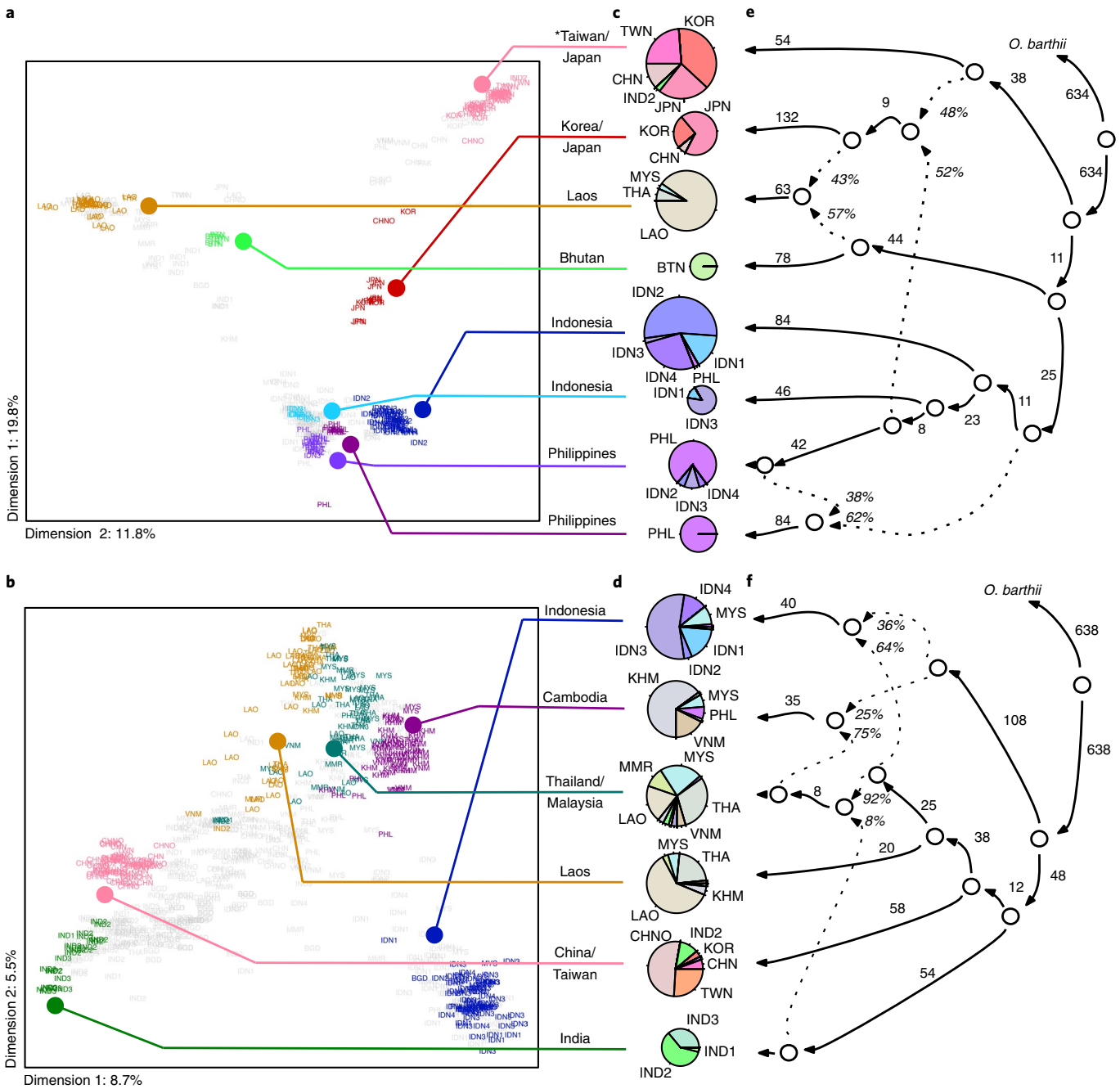

Fig. 2 | Subpopulations of *japonica* and *indica* rice. a,**b**, All *japonica* (**a**) and *indica* (**b**) landraces projected onto the first two dimensions after multidimensional scaling of genomic distances. **a**, The *japonica* genotypes were clustered using *k*-medoids ($k=9$ subpopulations) and filtered using silhouette parameters, which resulted in $k_d=8$ discrete subpopulations (coloured labels). Asterisk denotes a subpopulation cultivated in irrigated lowland conditions. **b**, *indica* genotypes were clustered using *k*-medoids ($k=7$ subpopulations) and filtered resulting in $k_d=6$ discrete subpopulations (coloured labels). **c**,**d**, Pie charts representing the geographical composition of each discrete subpopulation of *japonica* (**c**) and *indica* (**d**) subgroups. Chart diameter is proportional to the number of individuals in each subpopulation. **e**, Admixture graph for $k=9$, $k_d=8$ *japonica* subpopulations, rooted with *Oryza barthii* as an outgroup. This graph represents topology that is consistent between models for all lower values of *k*. **f**, Best admixture graph for $k=7$, $k_d=6$ *indica* subpopulations, rooted with *O. barthii* as an outgroup. Although this represents the best model, it is not consistent with other topologies at lower values of *k*, probably because of the complex history of *indica*. **e**,**f**, Solid lines with arrowheads represent uniform ancestries (attached numbers show scaled drift parameter f_2) and dashed lines represent mixed ancestries (percentage values indicate estimated proportion of ancestry). KOR, Korea; TWN, Taiwan; CHN, China; CHNO, China (centroid); IDN, Indonesia; JPN, Japan; MYS, Malaysia; THA, Thailand; LAO, Laos; BTN, Bhutan; PHL, Phillippines; IND, India; KHM, Cambodia; VNM, Vietnam; MMR, Myanmar. Numbers 1–4 represent distinct regions within countries.

remains[15], we note that rice agriculture spread northward and eastward along the Huang He river[31] and westward into the Chengdu Plains and the southwest China Highlands between approximately 5,000 to 4,000 yr BP[32–34] (Fig. 3b and Supplementary Fig. 18). During a minor climatic cooling event at around 5,000 yr BP, rice appears maladapted in parts of eastern China[35]. In the Shandong Peninsula,

rice disappeared by 5,000 yr BP and briefly re-emerged 4,500 yr BP as a short-grained variety similar to contemporary temperate *japonicas*[36]. The '4.2k event', a global temperature decrease that followed the HCO around 4,200 yr BP[29,30], resulted in waning rice agriculture in east China and strong pressure for *japonica* to adapt to a temperate environment[36]. Congruent with this, we observe that the highest density of estimated temperate *japonica* split times starts at about 4,100 yr BP (Fig. 3a and Supplementary Fig. 17).

Temperate adaptation created the opportunity for northeastern dispersal of *japonica* in Asia. From our demographic analysis of temperate *japonica* we note a five- to tenfold population (N_e) reduction between approximately 3,500 to 3,000 yr BP (Fig. 3c and Supplementary Fig. 19), which we interpret as a founder bottleneck during expansion to its new temperate niche. Indeed, this is consistent with archaeological dates for the introduction of rice agriculture to Korea[37,38] and Japan following decrease in rice remains in eastern China (Supplementary Fig. 18).

The southward spread of *japonica*. Throughout the HCO, tropical *japonica* was cultivated in eastern China; however, its contemporary descendants are grown predominantly in Southeast Asia[3], and we indeed find that Southeast Asian subpopulations descend from the tropical lineage. Demography reconstruction at $k_d = 2$–4 shows that the tropical *japonica* lineage experienced an approximately 50–100-fold N_e contraction between about 4,500 to 4,000 yr BP, and partial N_e recovery starting at around 2,500 yr BP (Fig. 3d and Supplementary Fig. 19). The population contraction in tropical *japonica* is contemporaneous with the 4.2k event, raising the possibility that cooling explains the collapse of tropical rice cultivation in East Asia and its southern relocation. This coincides with the arrival of rice in the far south of China around 4,500 yr BP and a shift to rainfed, upland cultivation[39].

Gradients of heat accumulation are highly associated with geographic distribution of *japonica* genomic diversity (Fig. 1e). On the basis of reconstruction of Holocene temperatures[40], we show that despite substantial temperatures changes, the spatial heat accumulation gradients, measured as GDD, remained stable over the last 5,500 years (Supplementary Fig. 20), suggesting that environment-associated genomic variation in *japonica* was influenced by spatial gradients in the past. To elucidate if tropical *japonica* could be successfully cultivated during the post-HCO period, we constructed a thermal-niche model[41], which estimates the probability of tropical rice cultivation in different areas during the post-HCO period (Fig. 3e and Supplementary Fig. 21). Survival probabilities of tropical *japonica* between approximately 4,400 and 3,500 yr BP dropped substantially in eastern China and high-altitude southwestern China (survival probability < 50%) compared with Southeast Asia (survival probability > 90%) (Fig. 3e and Supplementary Video 1). Indeed, after the cooling period we observe high densities of archaeological rice remains in Southeast Asia (Fig. 3b and Supplementary Fig. 18).

After the HCO, rice dispersed from China to Southeast Asia into Laos and Bhutan, and through maritime routes to the Philippines, Malaysia and Indonesia[15]. In our admixture-graph analysis, we find an early split in the tropical lineage that separates Bhutan and Laos upland rice from rice in the Malay Archipelago (Fig. 2e). From coalescence analyses we observe a population contraction of approximately 50–100-fold in the remote upland (Bhutan) rice

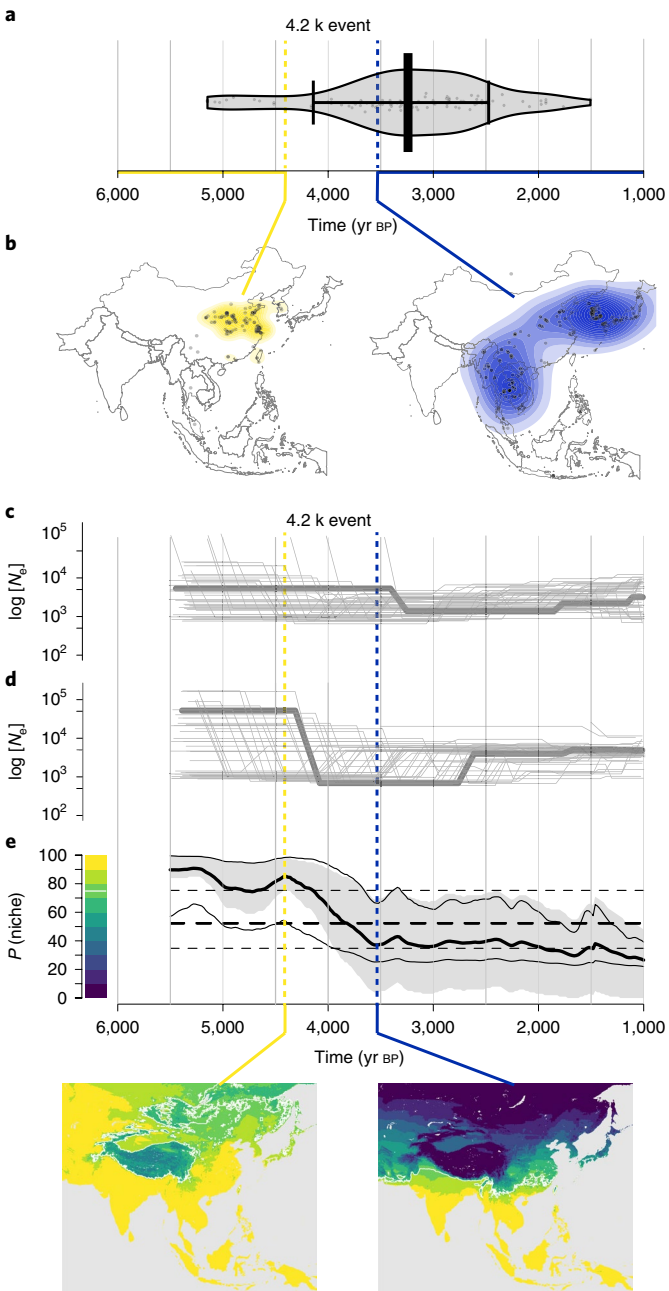

Fig. 3 | Demographic, paleoenvironmental and archaeological context of temperate *japonica* rice emergence. a, The distribution of temperate–tropical split times estimated from cross-coalescence analysis ($n = 100$ random pairs) of temperate and tropical individuals. The bar represents mean and bands represent 75% interquartile range. **b**, Maps indicating geographic locations and densities of archaeological sites with rice macro remains. Left: cumulative archaeobotanical evidence from 9,000–4,400 yr BP. Right: cumulative archaeobotanical evidence from 3,500–1,000 yr BP. **c**,**d**, Effective population sizes over time in tropical (**c**) and temperate (**d**) *japonica* subpopulations. Thin lines represent demographic histories ($n = 50$ random individuals) and bold lines represent joint models. **e**, Per cent probability of tropical rice being in the thermal niche (assuming requirement of 2,900 GDD at 10 °C base) over time. The thick black line represents mean and the grey shaded area represents 25% to 75% probability of being in the thermal niche ($n = 477{,}708$ cells). The thin black lines are the mean probabilities of being in the thermal niche when modelled using the 1σ uncertainty intervals as provided by the Northern Hemisphere temperature reconstruction ($n = 73$ datasets)[40]. The dashed horizontal lines are the long-term average probability of being in the thermal niche for the mean (thick dashed line) and 1σ uncertainty interval (thin dashed lines) reconstructions. The maps show the geographic distribution of niche probabilities. Left: before climate cooling (4,400 yr BP); right: after climate cooling (3,500 yr BP).

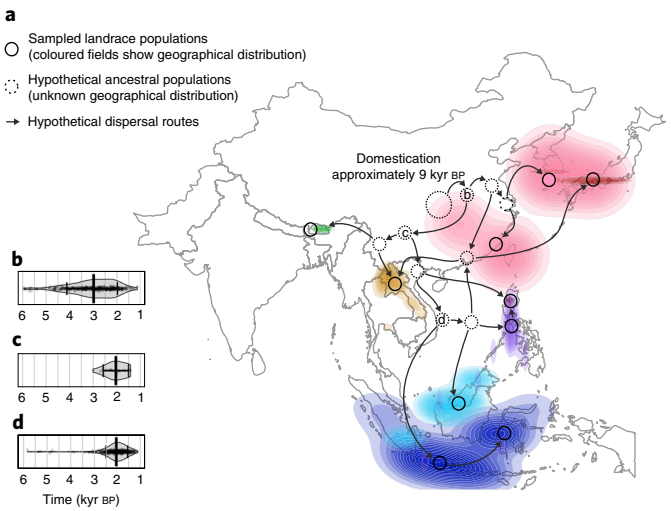

Fig. 4 | Proposed dispersal map of *japonica* rice in Asia. a, Map generated for *japonica*, $k_d = 8$ discrete subpopulations. The geographic distributions of subpopulations are represented as coloured, two-dimensional Kernel density fields. Bold circles represent leaves in the admixture graphs and are mapped close to the centres of subpopulation distributions. Dashed circles represent hypothetical ancestral subpopulations inferred from splits in best-matching admixture graphs; their precise geographic placement is uncertain. Arrows indicate hypothetical routes of dispersal. The distribution of split times between non-admixed subpopulations (**b**–**d**, as indicated on the map) was created from cross-coalescence estimates summarized over all k_d levels and presented as violin plots. The bar represents mean and bands represent 75% interquartile range. **b**, Split times between tropical and temperate linages ($n = 518$ random pairs). **c**, Split times between Bhutanese and Southeast Asian lineages ($n = 26$ random pairs). **d**, Split times between Philippine and Indonesian linages ($n = 483$ random pairs).

population between around 4,000 and 3,000 yr BP (Supplementary Fig. 19), which may arise from a bottleneck associated with population movements into these new areas. Emergence of upland rice in Laos and Bhutan (Fig. 4) coincides in time and space with widespread establishment of rainfed rice agriculture in mainland Southeast Asia around 4,000 yr BP[14,42] and dispersal of metallurgy traditions from Bronze Age Yunnan around 3,500 yr BP southwards to Thailand by approximately 3,000 yr BP[43,44]. Subsequent agricultural intensification of rice production took place from around 2,500 to 1,500 yr BP and included evolution of irrigation systems in present-day Thailand[45]. Consistent with these, ancient human DNA studies in Southeast Asia report two farmer-associated migration events from East Asia, one at least 4,000 years ago and a second before 2,000 yr BP[46,47].

Our analysis also shows a decrease of approximately five- to tenfold of N_e in the Malay archipelago between about 3,000 and 2,500 yr BP and, on the basis of cross-coalescence analyses, that divergence between mainland and Malay Archipelago rice occurred between around 3,000 to 1,500 yr BP (75% of estimates fall between about 2,500 and 1,600 yr bp) (Fig. 4 and Supplementary Fig. 17). Distinct island populations in the Malay Archipelago diverged at around a similar timeframe, in an interval from about 3,000 to 1,000 yr BP (75% of estimates fall between about 2,500 and 1,500 yr BP). This period coincides with dispersal of Dong Son drums in the Malay Archipelago (around 2,400 yr BP)[44,48], and suggests maritime dispersal of rice from a North Vietnam hub within the Austronesian trading sphere, which stretched between Taiwan and the Malay Peninsula[49,50]. Ancient DNA studies also suggest a wave of Austronesian human expansion into island Southeast Asia around 2,000 yr BP[46], which agrees with our estimates of *japonica* movement into the area. Interestingly, upland temperate *japonica* in Japan appears to be an admixed population of local lowland temperate rice and upland tropical rice from the Malay Archipelago, which may have moved northwards through Taiwan and perhaps the Ryukyu Islands around 1,200 yr BP[51].

Relationships and dispersal of *indica* subpopulations. We reconstructed relationships between *indica* subpopulations with $k_d = 2$ to 6. Divergence between Sino-Indian and Southeast Asian *indica* is present in all graph topologies beginning at $k_d = 2$. At $k_d = 3$ we observe separation of mainland and archipelago Southeast Asian subpopulations, and at $k_d = 4$ we observe separation of Indian from Chinese landraces (Supplementary Fig. 22). With $k_d = 5$ and $k_d = 6$, we note differentiation of mainland Southeast Asian landraces into subpopulations associated with Laos, Thailand and Cambodia (Fig. 2f). Notably, a subpopulation associated primarily with Cambodia and another in Indonesia, share ancestry with the main Laos–Thailand Southeast Asian lineage as well as an early ancestral *indica* population. Further increase of k_d also increases the number of admixture events in the model to four, which renders further exhaustive graph topology searches unfeasible.

We observed high diversity of graph topologies in *indica*, probably due to weak population structure and elevated gene flow (Supplementary Figs. 14 and 15), which also explains low silhouette scores and low associations with local environments. These characteristics of *indica* subpopulations are probably the reason behind difficulties with reconstruction of *indica* dispersal routes. Given the complexity in multiple reconstructed admixture-graph topologies, we can only confidently date separation of Chinese and Indian *indica*, which is unaffected by admixture. Our analysis estimates this divergence at around 2,500 to 1,100 yr BP (75% of estimates fall between around 2,000 and 1,400 yr BP) (Fig. 5 and Supplementary Fig. 17). Possible routes for *indica* dispersal from India to China could be the Silk Road or more direct passage to southwest China across the Hengduan mountains. The timing agrees with written reports of the introduction of Buddhism from India to China at around 1,950 yr BP[52], but is later than the earliest putative finds of *indica* rice in China[53]. The close relationship between Indian and Chinese subpopulations is mirrored by higher proportions of irrigated varieties in both regions; by contrast, Southeast Asian varieties are more frequently rainfed[25].

The dispersal of *indica* to Southeast Asia (for example, to Thailand and Cambodia) was from either India or China (Fig. 5 and Supplementary Fig. 23). Archaeobotanical studies suggest that *indica* arrived in central Thailand around 1,800 yr BP[45], at a time when Asian trade routes were well established[14]. Late adoption of *indica* in Southeast Asia is hypothesized to be due to early availability of *japonica* in this region[14]. There is no earlier archaeological evidence for *indica* cultivation in Southeast Asia, and thus it comes as a surprise that *indica* mainland subpopulations suffered marked population size reduction between about 5,000 and 3,500 yr BP (Supplementary Fig. 24). It is even more puzzling that a bottleneck in *indica* subpopulation in Indonesia occurred between about 6,000 and 5,000 yr BP, suggesting complex origins, perhaps through post-domestication introgression with local wild ancestors or managed pre-domesticated varieties (Supplementary Fig. 23).

Discussion

Rice domestication in the Yangtze Valley had an enormous impact on the peoples of East, Southeast and South Asia. In the first 4,000 years of its history, *japonica* rice cultivation was largely confined to China, and its dispersal and diversification did not occur until the global 4.2k cooling event. This abrupt climate change event was characterized by a global reduction in humidity and temperature; for example, average Northern Hemisphere temperatures shifted

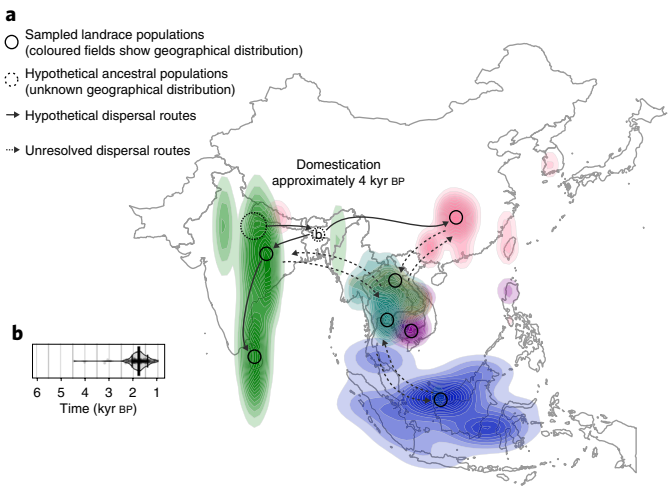

Fig. 5 | Proposed dispersal map of *indica* rice in Asia. a, Map generated for *indica*; $k_d = 6$ discrete subpopulations. The geographic distributions of subpopulations are represented as coloured, two-dimensional Kernel density fields. Bold circles represent leaves in the admixture graphs and are mapped close to the centres of subpopulation distributions. A dashed circle represents consistent split; its geographic position is uncertain. Solid arrows indicate hypothetical routes of dispersal and dotted arrows indicate possible routes that remain unresolved from admixture graphs. **b**, The distribution of split times between non-admixed subpopulations (as indicated in **a**) was created from cross-coalescence estimates summarized over all k_d levels ($n = 242$ random pairs) and presented as violin plots. The bar represents mean and bands represent 75% interquartile range.

from anomalies that were 0.4 °C above present day temperatures, to 0.2 °C cooler than the present[40].

This change had widespread consequences: it is believed to have caused the breakdown of rice agriculture in East Asia[29,36], turnover of cattle ancestry in the Near East[54] and the collapse of civilizations from Mesopotamia[55] to China[56]. We find, from our genomic and paleoclimate modelling, that the 4.2 k event coincides with the rise of temperate *japonica* and the dispersal of rice agriculture[14,15,42] and farmer communities[46,47] southwards into Southeast Asia. Correlation between changing climate and rice distribution raises the possibility of a causal relationship, and indeed we find temperature is a key environmental factor patterning contemporary rice genomic diversity.

The movement of *japonica* rice to island Southeast Asia took place later, after about 2,500 yr BP, as rice populations established themselves in Indonesia, Borneo and the Philippines. The islands of Southeast Asia were connected to each other and to the mainland at this time by extensive trade networks that were associated with the movement of goods and peoples in the region[49,50]. Our study suggests that these trading networks may have facilitated the establishment of rice agriculture in the Malay Archipelago, consistent with archaeological studies that suggest a late arrival of rice to the islands[15].

In South Asia, *indica* rice began to be domesticated at around the time of the 4.2k event; it subsequently spread into China and Southeast Asia. The spread of *indica* rice occurred much later than that of *japonica* rice, and extensive gene flows between geographic populations appears to have occurred, resulting in weaker between-population differentiation. Despite its current importance as the dominant rice subspecies grown in Asia, the details of the dispersal of *indica* remain obscure and will need further investigation.

The ability to infer dispersal patterns of rice arises from the availability of extensive landrace populations, whole-genome sequences representing global diversity[16] and population genomic approaches, as well as environmental, archaeobotanical and paleoclimate data. Reconstructing the history of domesticated species provides insight into the evolutionary process, nature of human–plant co-evolutionary dynamics, and extrinsic landscape, environmental and cultural factors that drive crop dispersal. Armed with knowledge of the pattern of rice dispersal and environmental features that influenced this migration, it may be possible to examine the evolutionary adaptations of rice as it spread to new environments, which could allow us to identify traits and genes to help future breeding efforts.

Methods

Landrace status. We considered 2,466 domesticated Asian rice (*O. sativa* L.) accessions from the International Rice Genebank Collection at the International Rice Research Institute that were included in the 3K-RG project[3,16], as well as an additional 178 accessions that were resequenced at New York University (Supplementary Table 1).

The definitions of landrace are very complex[57,58] and difficult to apply in practice during material collections. In our work we relied on the fact that landraces contain the signal of association with local geographic, environmental and cultural context. We used the following criteria: (1) we pre-selected 'candidate' landraces from available annotation, and (2) filtered them on the basis of their joint genetic and geographic clustering.

Accession passport data were obtained from the International Rice Information System (IRIS) (http://iris.irri.org/)[59]. We considered sample status of each accession and removed 'improved variety', 'wild' and 'weedy' accessions, while keeping 'traditional variety/landrace' accessions. We also kept 'breeding/inbred line' accessions if these were pure lines directly derived from 'traditional varieties/landraces' or were classic breeding lines from before the Green Revolution in the 1960s. From this set we removed any genetic clusters that were represented by individuals collected in in countries that do not share contiguous borders.

Geolocations and cultivation systems. Landrace geo-references were obtained from Genesys (https://www.genesys-pgr.org/welcome). For some landraces, instead of precise geo-coordinates, country- or region-level centroids were given (Supplementary Table 1). This problem was particularly relevant for landraces from China and Japan. Data on agro-ecosystems in which accessions are cultivated were obtained from IRIS[59]. On the basis of their agro-ecosystem of origin, accessions were divided into six cultivation types: 'irrigated', 'rainfed lowland', 'deepwater', 'upland', 'tidal wetland' and 'swamp'[60].

Accession growing season(s) in its local environment were estimated by considering information on cultivation type with prevalent rice growing-season months at the collection location. The latter information was obtained from the Rice Almanac[61] and Rice Atlas[62]. An accession from an 'irrigated' agro-ecosystem was assumed to be grown in all growing seasons if there were multiple seasons in its location of origin, since sufficient irrigation can presumably be provided in 'off' or 'dry' seasons. Accession from other agro-ecosystems were assumed grown only in the 'main' or 'wet' growing season as indicated in the Rice Almanac[61] and the Rice Atlas[62]. An accession's growing-season months were further specified if additional metadata on growing season were available from the IRIS database[59].

Biotic variables. It has been suggested that wild relatives of rice, particularly *Oryza rufipogon* and *O. nivara*, hybridized with cultivated rice in the past[11–13] altering the genomic composition of local subpopulations. We therefore considered the wild gene pool available to each candidate landrace. To this end, we used published ancestry composition data for rice wild relatives[63]. For each of our candidate landraces we took the ten geographically closest wild individuals and calculated means for six ancestry probabilities from fastStructure analysis at $k = 6$ (ref. [63]). Resulting ancestry probabilities do not represent any biological individuals, but rather a most probably hypothetical wild relative in the area of rice cultivation.

We also considered rice–human relationships not covered by geographic distance or resistance dispersal. One important property of rice grains in a culinary cultural context is stickiness, which is determined by the *waxy* gene[64]. As a proxy for conscious cultural preferences, we genotyped *waxy* alleles from genome-wide data. We also considered effects of unconscious cultural preferences on distribution of rice genomic diversity by accounting for the language family of nearby human populations. This aimed at modelling the different cultural preferences for rice traits, different cultural agronomic practices, or barriers to gene flow due to human interactions. We downloaded the linguistic map from the Glottolog database[65], and for each candidate landrace we queried the geographically closest spoken language.

Abiotic variables. We collated data for a suite of climate-related variables at the geo-location of each landrace using the EXTRACT function of the R package RASTER[66] v.2.8-19. Six temperature variables (average coldest temperature throughout growing season(s), average coldest temperature in first two months of growing season(s), mean temperature over growing season(s), average high temperature for last two months of growing season(s), GDD in growing season(s)

and interannual coefficient of variation of GDD) and three precipitation variables (accumulated precipitation in two months before the growing season(s), mean precipitation throughout growing season(s) and interannual coefficient of variation of precipitation) were derived from climatological data from 'climatologies at high resolution for the earth's land surface areas' (CHELSA v.1.2), which provides monthly and mean-annual precipitation and temperature data at 30 arcsec resolution for the time period 1979–2013[67]. For calculations of GDD, we used monthly means as proxies of average daily air temperatures for months in the growing season with a mean above a base temperature of 10 °C.

We included two variables that reflect evapotranspiration processes during the growing season(s): potential evapotranspiration (PET) and ratio of PET to mean precipitation. PET variables were based on monthly values from the CGIAR Consortium for Spatial Information Global-PET Database (http://www.cgiar-csi.org)[68,69].

We also included distance from the geo-location of each landrace to the nearest lake or river based on a previous global analysis of human population distance to freshwater[70]. Elevation above sea level was obtained from WorldClim[71].

Among edaphic variables, we included soil salinity (measured as electric conductivity), pH and sodicity (exchangeable sodium percentage) from the Harmonized World Soil Database v.1.2 (http://webarchive.iiasa.ac.at/Research/LUC/External-World-soil-database/HTML/index.html?sb=1). Information on soil total nitrogen density was extracted from the Global Gridded Surfaces of Selected Soil Characteristics dataset[72]. To capture the soil moisture potentially available for plant growth, we used plant extractable water capacity of soil[73] and depth to water table[74].

Sequencing data. Sequencing data for individuals that were marked as candidate landraces (see landrace status section in Methods) from the 3K-RG project were downloaded in fastq format from the Short Read Archive (SRA) using the FASTQ-DUMP tool v.2.8.2 with the option to split reads into forward, reverse and trimmed.

We generated sequencing data for an additional 178 landraces. Leaf samples were ground using mortar and pestle in liquid nitrogen. DNA was extracted using the Qiagen DNeasy Plant Mini Kit following the manufacturer's protocol (Qiagen). Yields ranged between 3 ng μl^{-1} and 102 ng μl^{-1}. Extracted DNA from each sample was prepared for Illumina genome sequencing using the Illumina Nextera DNA Library Preparation Kit. Sequencing was done on the Illumina HiSeq 2500 in HighOutput Mode v3 with 2 × 100 bp read configuration, at the New York University Genomics Core Facility. Sequencing data these accessions are available from the SRA under Bioproject accession numbers PRJNA422249 and PRJNA557122.

Alignment and genotyping. We used Nextflow[75] v.0.25.1.4460 to build a pipeline for calling SNPs in our dataset (https://github.com/grafau/NextGatkSNPs). All steps necessary to obtain our SNP set are described below. Sequencing data in fastq format for each run of candidate landraces were mapped against the reference genome of *indica* variety Shuhui498 v.1.0[76] using the global aligner BWA v.0.7.15 in 'mem' mode[77] and sorted using PICARD v.2.15.0. Sequences for the same sample, but from different runs, were merged and amplification duplicates were removed using PICARD. The resultant sam format files were validated and indexed, producing bam format files.

Bam files were used to call haplotypes in GATK v.3.8[78] with the HAPLOTYPECALLER function in 'discovery' mode and set to produce gvcf format files. Subsequently, gvcf files were validated and combined into 8 batches with GATK, each batch containing approximately 200 landraces. These combined gvcf files were compressed and indexed using BGZIP and TABIX, respectively[79]. Contents of combined gvcf files were divided into 12 chromosomes and each chromosome file was genotyped for all eight batches together using the GENOTYPEGVCFS function of GATK to produce the raw set of SNPs segregating among rice landraces.

SNP filtering. The raw set of SNPs was subject to a series of filtering steps. First, we kept only biallelic SNPs. Subsequently, we applied five filtering criteria: qualities normalized by depth (QD), mapping quality (MQ and MQRankSum), read position bias from Wilcoxon's test (ReadPosRankSum) and strand bias from Fisher's test (FS). Filtering thresholds for these criteria were trained dynamically using the VARIANTRECALIBRATOR function of GATK, referencing a true-positive set of SNPs that were discovered independently in the 3K-RG project[3], and in the rice diversity panel that was genotyped with a high-density SNP array[80]. We applied the dynamic filter to our raw set of SNPs using the APPLYRECALIBRATION function of GATK, conservatively set to recover 90% of true positives.

To obtain an estimate for expected heterozygosity in rice populations, we calculated inbreeding coefficients in all landraces of *circum*-aus, *indica* and *japonica* groups. Coefficients were calculated as medians of ratios, where each ratio is equal to observed heterozygosity divided by expected heterozygosity for each SNP with >5% minor allele frequency (only ratios smaller than 1 were taken into account). We then compared observed heterozygosity to expected heterozygosity for each SNP, given the inbreeding coefficient, and carried out a chi-squared test to filter out SNPs with excess heterozygosity. We performed this step for all landraces and for each subgroup separately. We interpret excessively heterozygous sites as mismapped reads in chromosomal regions with structural variants that are present in the resequencing data but absent in the reference genome.

Next, we transformed vcf files into bed format files using PLINK v.1.90b4[81,82], and kept only candidate landraces that were collected in Asia. From this set, we filtered out any SNP that had a genotyping rate lower than 80% with PLINK. This step was carried out independently for all landraces, and for *indica* and *japonica* subgroups separately. For some analyses (Supplementary Fig. 1), SNP sets were subject to additional two-step linkage-disequilibrium pruning. The first step was carried out with the INDEP-PAIRWISE function in windows of 10 kb with variant shift = 1 and $r^2 = 0.8$. The second step was carried out with the same function in windows of 50 variants.

Landrace subsetting. We used metadata retrieved from the online platform Genesys (for details, see landrace status section in Methods) to annotate each candidate landrace with country of origin and subgroup designation (*indica*, *japonica*, *circum*-aus or *circum*-basmati) provided by the 3K-RG[3] and rice diversity panel[80] projects. We then carried out factorial analysis of molecular variance in R v.3.4.1[83] using the ADONIS function from the VEGAN package[84]. Subsequently, we split the dataset into four subsets corresponding to four subgroups (each SNP set was subject to heterozygosity and genotyping rate filtering, see section: SNP filtering). We used pairwise genomic distances among all landraces to calculate silhouette scores[18] for each landrace given its subgroup affiliation. We then filtered out landraces with silhouette scores below 0.2, as this might indicate admixture between subgroups or mislabelling. All analyses described in this section were carried out after phasing imputation on the SNP sets with BEAGLE v.5.0[85].

Migration barriers. We estimated effective migration surfaces using the EEMS tool[19]. We chose map-outline coordinates that stretch from Pakistan to Japan and Papua New Guinea using an online tool (http://www.birdtheme.org/useful/v3tool.html) and specified a triangular grid with 200 demes for Voronoi tessellation. The best-fitting model was acquired from converging three independent runs of five million Monte Carlo Markov chain iterations, of which the first two million burn-in runs were discarded. Surfaces were plotted in R v.3.4.1[83] using the EEMS.PLOT function from the REEMSPLOTS package[19] and mapped with Mercator projection.

Fastest travel time between each pair of geo-referenced accessions was estimated using least-cost paths analysis in R v3.5.1 with the package GDISTANCEv1.1[86]. Travelling speed over land given the slope between adjacent grid cells was calculated according to Tobler's hiking function[87], on the basis of elevation data at 30 arcsec resolution from WorldClim v.1.4. For travel over sea, we assumed a constant speed of 3 knots under sail[21,88,89]. GPS coordinates for each landrace accession were rounded to the nearest 0.1° to reduce the computation time needed. Pairwise resistance-distance matrices were populated separately for *indica* and *japonica* accessions.

Spatial correlations. We tested whether genetic distance between landraces could be explained better by geographic distance or estimated travel time between geo-locations of origin. We first filtered out landraces in China, because they all were annotated with low resolution geo-coordinates that mapped to country and regional centroids. We employed a linear mixed model with maximum-likelihood estimation from Clarke et al.[90] and used AIC[91] to select between geographic-distance and travel-time models. This linear mixed model includes spatial random effects to account for non-independence among nearby samples. The use of AIC with such a mixed model has been shown to offer the greatest accuracy in identifying the true isolation model under a wide range of scenarios[92]. We implemented our mixed model and AIC calculations with the RESISTANCEGA package[93,94] in R v.3.6.0[83]. Proportions of variance explained (r^2) were calculated with the LM function and *P* values were calculated using Mantel tests and permutations implemented in VEGAN[84].

Processes driving gene flow may have been very different in mainland Asia versus the Malay Archipelago. Additionally, the travel-time model we developed was new, and therefore had an uncertain ability to capture different travel mechanisms. Thus, we stratified analyses into two main groups each for both *japonica* and *indica*: a group of 'mainland' landraces and a group of 'archipelago' landraces. Mainland landraces were defined to include those north of 9.7° N latitude and west of 110° E longitude, thus excluding the relatively small number of isolated mainland landraces to the east (for example, eastern China). Archipelago landraces included those from the Malay Archipelago and the Malay Peninsula, but not from the islands to the north (that is, Taiwan or Japan).

Redundancy analyses. RDA are eigen analyses for multivariate responses and multivariate predictors that maximize the proportion of variation explained in the responses. We used RDA to identify sets of variables important for explaining SNP variation in landraces and for identifying specific biotic and abiotic variables explaining the most genome-wide SNP variation. To incorporate pairwise geographic distance or travel time into our RDA, we converted distance matrices into spatial weighting matrices and then a reduced-dimension set of orthogonal

variables (Moran's eigenvector maps (MEMs))[95]. MEMs are eigenvectors of the pairwise spatial weighting matrix among samples. We optimized both geographic-distance and travel-time matrices using a subset of 10,000 randomly chosen SNPs for response variables in RDA, optimizing separately for *japonica* and *indica*.

Weighting matrices among unique landrace collection locations (Chinese accessions were all filtered out) were generated using the ADESPATIAL package[96] in R. We used two algorithms, Gabriel graph and distance-based graph, to generate three candidate connectivity matrices. The Gabriel graph results primarily in connections among neighbouring sites. A distance-based graph connects sites closer than a given threshold, for which we used two values: minimum distance required to connect all points (that is, the largest distance of a minimum spanning tree) and infinity (resulting in a fully connected graph[95]). With each of these three connectivity matrices, we generated two spatial weighting matrices using two distance-decay functions: linear (weight between two sites $= 1 - D/D_{max}$, where D is distance between sites and D_{max} is maximum distance among all sites) or concave up (weight between two sites $= D^{-0.01}$). These connectivity and weighting algorithms resulted in six diverse MEM sets, differing largely in levels of spatial autocorrelation and structure among MEM eigenvectors. We used the Bauman et al.[95] forward selection of MEM eigenvectors algorithm to optimize number of eigenvectors (restricted to those with positive eigenvalues) included in RDA for each MEM set. Optimization is based on adjusted r^2 (which are penalized or adjusted for number of explanatory values), and the MEM set with greatest adjusted r^2 is defined as the optimal set. In the RDA presented in the main text, we used weighting matrices based on geographic distance for *indica* and based on travel time for *japonica*, because model selection favoured these distance measures. For *indica* and geographic distance, optimization selected 25 MEM eigenvectors from the connectivity matrix on the basis of connecting all sites within the threshold distance required to connect all points in a single graph and using weighting that was a linear function of distance. For *japonica* and travel time, optimization selected the same connectivity matrix and distance-weighting algorithms, with 33 eigenvectors. These eigenvectors were included in the RDA described below on *japonica* and *indica* whole-SNP dataset.

We then conducted RDA with variance partitioning[22] to quantify proportion of genome-wide SNP variation explained by each of four categories of covariates: abiotic variables, geographic-isolation MEMs, *waxy* allelic status and language family. Variance partitioning estimates proportion of SNP variance that is explained by variables in each category and by collinearity among variables. To identify specific abiotic variables associated with genome-wide divergence among landraces, we also conducted RDA using only abiotic gradients for *indica* and *japonica*. For visualization, specific abiotic variables highlighted in Fig. 1 indicate those loading most strongly in each direction along each RDA canonical axis as well as those loading most strongly on each diagonal (identified by multiplying the loadings on the first two canonical axes). All RDAs (including variance partitioning) were conducted using VEGAN[84].

Clustering and discretization. Clustering was visualized using multidimensional-scaling methods. Genetic distances among and within each rice subgroup were calculated between all pairs of candidate landraces using PLINK v.1.9[81,82] with formulation: 1 − IBS, where IBS is identity by state. After importing the distance matrix into R v.3.4.1[83] the CMDSCALE function was used to calculate eigenvectors[97], which were plotted in three dimensions. The variance explained by each dimension was calculated as the dimension's eigenvalue divided by sum of all positive eigenvalues.

Formal clustering of landraces within *japonica* and *indica* was carried out on the basis of pairwise genetic matrices with the PAM method[98] implemented as the PAM function in the CLUSTER package for R v.3.4.1[83]. Subsequently, clusters were filtered with our DISCRETIZE algorithm implemented in R. The algorithm first removes individuals with negative silhouette scores. Second, for each cluster it designates a pairing partner, which is another cluster with the least-distant medoid. DISCRETIZE simulates individuals that are admixed between the two paired clusters with requested ancestry proportions by computing weighted-mean distance between paired medoids and all other individuals (here we simulated individuals with 0.5–0.5, 0.4–0.6 and 0.6–0.4 admixture proportions). For all simulated individuals, our algorithm computes silhouette scores and keeps the highest value as threshold for filtering. Individuals are clustered with PAM and filtered on the basis of each cluster's silhouette threshold. This process is repeated iteratively until no more individuals are filtered out. A script written in R that can perform these analyses is publicly available (https://github.com/grafau/discretize).

Clustering and discretization was carried out independently for a number of clusters, k, that varied from 2 to 12. Discrete clusters are considered subpopulations and their members are considered landraces conditional on a colocalized geographic distribution within each discrete cluster. We investigated composition of clusters with regard to region of origin to determine whether each fulfilled our latter criterion for landrace status. One *indica* cluster exhibited poor geographic colocalization and was therefore removed from all analyses (see landrace status section in Methods). In order to visualize the geographic provenance of each discrete cluster, we plotted the two-dimensional distribution (for latitude and longitude) of landraces using the GEOM_DENSITY2D and STAT_DENSITY2D functions from the GGPLOT package[99] onto a map of Asia in R v.3.4.1[83].

Admixture-graph reconstruction. We reconstructed admixture graphs for *japonica* and *indica* subpopulations defined by the DISCRETIZE algorithm. Individual lists are available in Supplementary Table 1. Reconstruction attempts were carried out independently for varying numbers of subpopulations, with k_d ranging from 2 to 9, using 19 accessions of *O. barthii* as outgroup. We aimed to show that our conclusions are supported independently of the chosen number of populations (k_d). The CONVERTF function from ADMIXTOOLS was used to produce eigenstrat data files, and the QPGRAPH function was used to evaluate whether models fit the data. Models were taken from ADMIXTUREGRAPH package[100] in R v.3.4.1[83] and transcribed into the format accepted by ADMIXTOOLS[28].

From $k_d = 2$ to $k_d = 5$ (3 to 6 subpopulations including outgroup), we explored the entire space of possible models with 0, 1 and 2 migrations and reported all models with f_4-statistic z-scores <3.0 (Supplementary Figs. 16 and 22). For $k_d = 7$ to 9, we first explored all possible models with 6 subpopulations and 0, 1 and 2 migrations, keeping only those with f_4-statistic z-scores <3.0. For each model we kept, we attached an additional subpopulation in all possible nodes using ADMIXTUREGRAPH and tested the resulting models in ADMIXTOOLS, again keeping only models with f_4-statistic z-scores <3.0. We progressively added subpopulations until no more were present or until no models with f_4-statistic z-scores <3.0 were found. In the latter case, we kept all models with f_4-statistic z-scores lower than 10.0. We then added an additional admixture event in all possible nodes using ADMIXTUREGRAPH and tested resultant models in ADMIXTOOLS, keeping only models with f_4-statistic z-scores <3.0. The number of possible models fulfilling this criterion was large, so we summarized their topologies in three different 'topology groups' and showed representative models characterized by the best z-scores together with total number of models in these topology groups (Supplementary Figs. 16 and 22).

Demography and split time reconstruction. To better understand past demographics of rice, we attempted to reconstruct past effective population sizes using the SMC++ method[101]. Reconstructions were carried out independently for a varying number of subpopulations, with k_d ranging from 2 to 8. We aimed to show that our conclusions are supported independently of the chosen number of populations (k_d). We selected a variety of 'distinguished pairs' for each subpopulation by sampling 50 individuals without replacement and pairing them with 50 individuals sampled with replacement. We kept this number close to the mean number of individuals per subpopulation. In subpopulations with fewer than 50 individuals assigned, we sampled all of the individuals and paired each with individuals sampled with replacement. We then partitioned vcf files into smc haploblock files for each distinguished pair, further partitioned for each chromosome, and masked the homozygous pericentromeric regions[102]. Subsequently, we used a polarization error of 0.5 and mutation rate[103] of 6.5×10^{-9} in the ESTIMATE function of SMC++ to estimate past effective population sizes. Results were scaled in time using an estimate of 1 yr as generation time and plotted on a linear timescale. We also used these demographies in calculating split times between subpopulations in a cross-coalescent framework of SMC++. The distinguished pairs were determined as described above, with the difference that each individual of the pair belonged to a different subpopulation.

Archaeological and paleoenvironmental context. Using a comprehensive database of rice archaeological records[15] in 1,000 yr intervals, we plotted two-dimensional distributions (for latitude and longitude) using GEOM_DENSITY2D and STAT_DENSITY2D functions from GGPLOT[99] onto a map of Asia in R v.3.4.1[83].

To predict how changing temperatures might have affected distribution of different types of rice (*indica* and temperate and tropical varieties of *japonica*) we used a global record of Holocene temperatures[40] to reconstruct GDD following the methods of d'Alpoim and Bocinsky[41]. We derived daily modern temperatures from Global Historical Climatology Network weather stations across East, South and Central Asia[104]. To account for spatial heterogeneity in how stations at different altitudes respond to climatic change, we used variance matching and modulated maximum and minimum mean weather-station climatology by standard deviations (SDs) derived from Marcott et al.[40]. This was carried out for each year in the Marcott record. The niche of different types of landraces was established by thresholding annual GDD, a measure of accumulated units of heat required by plants to complete their life cycle. We then used indicator kriging to spatially interpolate these niches across the ETOPO5 5 arcmin (approximately 10 km) resolution elevation model[105]. The full research compendium that contains all the code and data necessary to reproduce this analysis is available at https://github.com/bocinsky/gutaker2020.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Raw FASTQ reads for 178 accessions whose genomes were resequenced for this study have been deposited in the SRA under Bioproject accession numbers PRJNA422249 and PRJNA557122. Sources for all downloaded data are referred to in the Supplementary Information.

Code availability

Code repositories are available at: https://github.com/bocinsky/gutaker2020, https://github.com/grafau/discretize, https://github.com/grafau/NextGatkSNPs and https://github.com/em-bellis/riceTravelTime.

Received: 14 November 2019; Accepted: 2 April 2020;
Published online: 15 May 2020

References

1. Purugganan, M. D. & Fuller, D. Q. The nature of selection during plant domestication. *Nature* **457**, 843–848 (2009).
2. Meyer, R. S. & Purugganan, M. D. Evolution of crop species: genetics of domestication and diversification. *Nat. Rev. Genet.* **14**, 840–852 (2013).
3. Wang, W. et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* **557**, 43–49 (2018).
4. Glaszmann, J. C. Isozymes and classification of Asian rice varieties. *Theor. Appl. Genet.* **74**, 21–30 (1987).
5. Fuller, D. Q. et al. The contribution of rice agriculture and livestock pastoralism to prehistoric methane levels: an archaeological assessment. *Holocene* **21**, 743–759 (2011).
6. Fuller, D. Q. & Qin, L. Water management and labour in the origins and dispersal of Asian rice. *World Archaeol.* **41**, 88–111 (2009).
7. Fuller, D. Q. et al. The domestication process and domestication rate in rice: spikelet bases from the Lower Yangtze. *Science* **323**, 1607–1610 (2009).
8. Allaby, R. G., Stevens, C., Lucas, L., Maeda, O. & Fuller, D. Q. Geographic mosaics and changing rates of cereal domestication. *Philos. Trans. R. Soc. Lond. B* **372**, 20160429 (2017).
9. Silva, F. et al. A tale of two rice varieties: modelling the prehistoric dispersals of *japonica* and proto-*indica* rices. *Holocene* **28**, 1745–1758 (2018).
10. Fuller, D. Q. Pathways to Asian civilizations: tracing the origins and spread of rice and rice cultures. *Rice* **4**, 78–92 (2011).
11. Huang, X. et al. A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490**, 497–501 (2012).
12. Choi, J. Y. & Purugganan, M. D. Multiple origin but single domestication led to *Oryza sativa*. *G3* **8**, 797–803 (2018).
13. Choi, J. Y. et al. The rice paradox: multiple origins but single domestication in Asian rice. *Mol. Biol. Evol.* **34**, 969–979 (2017).
14. Fuller, D. Q., Castillo, C. C. & Murphy, C. in *The Routledge Handbook of Archaeology and Globalization* (ed. Hodos, T.) 711–729 (Routledge, 2016).
15. Silva, F. et al. Modelling the geographical origin of rice cultivation in Asia using the rice archaeological database. *PLoS ONE* **10**, e0137024 (2015).
16. Li, J.-Y., Wang, J. & Zeigler, R. S. The 3,000 rice genomes project: new opportunities and challenges for future rice research. *Gigascience* **3**, 2047–217X–3–8 (2014).
17. Excoffier, L., Smouse, P. E. & Quattro, J. M. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**, 479–491 (1992).
18. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
19. Petkova, D., Novembre, J. & Stephens, M. Visualizing spatial population structure with estimated effective migration surfaces. *Nat. Genet.* **48**, 94–100 (2016).
20. Peter, B. M., Petkova, D. & Novembre, J. Genetic landscapes reveal how human genetic diversity aligns with geography. *Mol. Biol. Evol.* **37**, 943–951 (2020).
21. Slayton, E. R. *Seascape Corridors: Modeling Routes to Connect Communities Across the Caribbean Sea*. (Sidestone Press, 2018).
22. Peres-Neto, P. R., Legendre, P., Dray, S. & Borcard, D. Variation partitioning of species data matrices: estimation and comparison of fractions. *Ecology* **87**, 2614–2625 (2006).
23. Lasky, J. R. et al. Genome–environment associations in sorghum landraces predict adaptive traits. *Sci. Adv.* **1**, e1400218 (2015).
24. Lasky, J. R. et al. Characterizing genomic variation of *Arabidopsis thaliana*: the roles of geography and climate. *Mol. Ecol.* **21**, 5512–5529 (2012).
25. Haefele, S. M., Nelson, A. & Hijmans, R. J. Soil quality and constraints in global rice production. *Geoderma* **235-236**, 250–259 (2014).
26. Kaufmann, L & Rousseeuw, P. J. in *Reports of the Faculty of Technical Mathematics and Informatics* Vol. 87 (Delft University of Technology, 1987).
27. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
28. Patterson, N. et al. Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
29. An, C.-B., Tang, L., Barton, L. & Chen, F.-H. Climate change and cultural response around 4000 cal yr B.P. in the western part of Chinese Loess Plateau. *Quat. Res* **63**, 347–352 (2005).
30. Walker, M. J. C. et al. Formal subdivision of the Holocene series/epoch: a discussion paper by a working group of INTIMATE (integration of ice-core, marine and terrestrial records) and the subcommission on Quaternary stratigraphy (International Commission on Stratigraphy). *J. Quat. Sci* **27**, 649–659 (2012).
31. Lanehart, R. E. et al. Dietary adaptation during the Longshan period in China: stable isotope analyses at Liangchengzhen (southeastern Shandong). *J. Archaeol. Sci.* **38**, 2171–2181 (2011).
32. Guedes, J. D., Jiang, M., He, K., Wu, X. & Jiang, Z. Site of Baodun yields earliest evidence for the spread of rice and foxtail millet agriculture to south-west China. *Antiquity* **87**, 758–771 (2013).
33. Guedes, J. D. & Butler, E. E. Modeling constraints on the spread of agriculture to Southwest China with thermal niche models. *Quat. Int.* **349**, 29–41 (2014).
34. Dal Martello, R. et al. Early agriculture at the crossroads of China and Southeast Asia: archaeobotanical evidence and radiocarbon dates from Baiyangcun, Yunnan. *J. Archaeol. Sci. Rep.* **20**, 711–721 (2018).
35. Fuller, D. Q., Weisskopf, A. R. & Castillo, C. Pathways of rice diversification across Asia. *Archaeol. Int.* **19**, 84–96 (2016).
36. d'Alpoim Guedes, J., Jin, G. & Bocinsky, R. K. The impact of climate on the spread of rice to north-eastern China: a new look at the data from Shandong province. *PLoS ONE* **10**, e0130430 (2015).
37. Crawford, G. W. & Lee, G.-A. Agricultural origins in the Korean Peninsula. *Antiquity* **77**, 87–95 (2003).
38. Ahn, S.-M. The emergence of rice agriculture in Korea: archaeobotanical perspectives. *Archaeol. Anthropol. Sci.* **2**, 89–98 (2010).
39. Yang, X. et al. New radiocarbon evidence on early rice consumption and farming in South China. *Holocene* **27**, 1045–1051 (2017).
40. Marcott, S. A., Shakun, J. D., Clark, P. U. & Mix, A. C. A reconstruction of regional and global temperature for the past 11,300 years. *Science* **339**, 1198–1201 (2013).
41. d'Alpoim Guedes, J. & Bocinsky, R. K. Climate change stimulated agricultural innovation and exchange across Asia. *Sci. Adv.* **4**, eaar4491 (2018).
42. Castillo, C. C., Fuller, D. Q., Piper, P. J., Bellwood, P. & Oxenham, M. Hunter-gatherer specialization in the late Neolithic of southern Vietnam—the case of Rach Nui. *Quat. Int.* **489**, 63–79 (2018).
43. Higham, C. F. W. Debating a great site: Ban Non Wat and the wider prehistory of Southeast Asia. *Antiquity* **89**, 1211–1220 (2015).
44. Higham, C. *The Bronze Age of Southeast Asia* (Cambridge Univ. Press, 1996).
45. Castillo, C. C. et al. Social responses to climate change in Iron Age north-east Thailand: new archaeobotanical evidence. *Antiquity* **92**, 1274–1291 (2018).
46. McColl, H. et al. The prehistoric peopling of Southeast Asia. *Science* **361**, 88–92 (2018).
47. Lipson, M. et al. Ancient genomes document multiple waves of migration in Southeast Asian prehistory. *Science* **361**, 92–95 (2018).
48. Calò, A. *The Distribution of Bronze Drums in Early Southeast Asia: Trade Routes and Cultural Spheres*. (Archaeopress, 2009).
49. Castillo, C. C., Bellina, B. & Fuller, D. Q. Rice, beans and trade crops on the early maritime Silk Route in Southeast Asia. *Antiquity* **90**, 1255–1269 (2016).
50. Hung, H.-C. et al. Ancient jades map 3,000 years of prehistoric exchange in Southeast Asia. *Proc. Natl Acad. Sci. USA* **104**, 19745–19750 (2007).
51. Takamiya, H., Hudson, M. J., Yonenobu, H., Kurozumi, T. & Toizumi, T. An extraordinary case in human history: prehistoric hunter-gatherer adaptation to the islands of the Central Ryukyus (Amami and Okinawa archipelagos), Japan. *Holocene* **26**, 408–422 (2016).
52. Zürcher, E. in *The Buddhist conquest of China* (Brill, 1972).
53. Deng, Z. et al. From early domesticated rice of the middle Yangtze basin to millet, rice and wheat agriculture: archaeobotanical macro-remains from Baligang, Nanyang Basin, central China (6700–500 BC). *PLoS ONE* **10**, e0139885 (2015).
54. Verdugo, M. P. et al. Ancient cattle genomics, origins, and rapid turnover in the Fertile Crescent. *Science* **365**, 173–176 (2019).
55. Gibbons, A. How the Akkadian empire was hung out to dry. *Science* **261**, 985 (1993).
56. Wang, J. et al. The abrupt climate change near 4,400 yr BP on the cultural transition in Yuchisi, China and its global linkage. *Sci. Rep.* **6**, 27723 (2016).
57. Harlan, J. R. Our vanishing genetic resources. *Science* **188**, 617–621 (1975).
58. Villa, T. C. C., Maxted, N., Scholten, M. & Ford-Lloyd, B. Defining and identifying crop landraces. *Plant Genet. Resour.* **3**, 373–384 (2005).
59. McLaren, C. G., Bruskiewich, R. M., Portugal, A. M. & Cosico, A. B. The International Rice Information System. A platform for meta-analysis of rice crop data. *Plant Physiol.* **139**, 637–642 (2005).
60. Huke, R. E. & Huke, E. H. *Rice Area by Type of Culture: South, Southeast, and East Asia: A Revised and Updated Data Base* (International Rice Research Institute, 1997).

61. Maclean, J., Hardy, B. & Hettel, G. *Rice Almanac: Source Book for One of the Most Important Economic Activities on Earth* 4th edn (International Rice Research Institute, 2013).
62. Laborte, A. G. et al. RiceAtlas, a spatial database of global rice calendars and production. *Sci. Data* **4**, 170074 (2017).
63. Kim, H. et al. Population dynamics among six major groups of the *Oryza rufipogon* species complex, wild relative of cultivated Asian rice. *Rice* **9**, 56 (2016).
64. Hirano, H. Y., Eiguchi, M. & Sano, Y. A single base change altered the regulation of the *Waxy* gene at the posttranscriptional level during the domestication of rice. *Mol. Biol. Evol.* **15**, 978–987 (1998).
65. Hammarström, H., Forkel, R. & Haspelmath, M. *Glottolog 4.0* (2019); https://doi.org/10.5281/zenodo.3260726
66. Hijmans, R. J. & van Etten, J. raster: Geographic data analysis and modeling v.2 (2014).
67. Karger, D. N. et al. Climatologies at high resolution for the earth's land surface areas. *Sci. Data* **4**, 170122 (2017).
68. Zomer, R. J. et al. *Trees and Water: Smallholder Agroforestry on Irrigated Lands in Northern India*. (International Water Management Institute, 2007).
69. Zomer, R. J., Trabucco, A., Bossio, D. A. & Verchot, L. V. Climate change mitigation: A spatial analysis of global land suitability for clean development mechanism afforestation and reforestation. *Agric. Ecosyst. Environ.* **126**, 67–80 (2008).
70. Kummu, M., de Moel, H., Ward, P. J. & Varis, O. How close do we live to water? A global analysis of population distance to freshwater bodies. *PLoS ONE* **6**, e20578 (2011).
71. Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. & Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **25**, 1965–1978 (2005).
72. Global Soil Data Task Group. *Global gridded surfaces of selected soil characteristics (IGBP-DIS)* (2002); https://doi.org/10.3334/ORNLDAAC/569
73. Dunne, K. A. & Willmott, C. J. Global distribution of plant-extractable water capacity of soil. *Int. J. Climatol.* **16**, 841–859 (1996).
74. Fan, Y., Li, H. & Miguez-Macho, G. Global patterns of groundwater table depth. *Science* **339**, 940–943 (2013).
75. Di Tommaso, P. et al. Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
76. Du, H. et al. Sequencing and de novo assembly of a near complete *indica* rice genome. *Nat. Commun.* **8**, 15324 (2017).
77. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).
78. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–11.10.33 (2013).
79. Li, H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* **27**, 718–719 (2011).
80. McCouch, S. R. et al. Open access resources for genome-wide association mapping in rice. *Nat. Commun.* **7**, 10532 (2016).
81. 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).
82. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
83. R Core Team. *R: A Language and Environment for Statistical Computing* http://www.R-project.org/ (R Foundation for Statistical Computing, 2013).
84. Oksanen, J. *Vegan: An Introduction to Ordination* https://cran.r-project.org/web/packages/vegan/vignettes/intro-vegan.pdf (2015).
85. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
86. van Etten, J. R package gdistance: distances and routes on geographical grids. *J. Stat. Softw.* **76**, v76i13 (2017).
87. Tobler, W. *Three Presentations on Geographical Analysis and Modeling: Non-isotropic Geographic Modeling; Speculations on the Geometry of Geography; And Global Spatial Analysis* (National Center for Geographic Information and Analysis, 1993).
88. White, D. A. & Surface-Evans, S. L. *Least Cost Analysis of Social Landscapes: Archaeological Case Studies* (Univ. Utah Press, 2012).
89. Irwin, G., Bickler, S. & Quirke, P. Voyaging by canoe and computer: experiments in the settlement of the Pacific Ocean. *Antiquity* **64**, 34–50 (1990).
90. Clarke, R. T., Rothery, P. & Raybould, A. F. Confidence limits for regression relationships between distance matrices: estimating gene flow with distance. *J. Agric. Biol. Environ. Stat.* **7**, 361 (2002).
91. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* **19**, 716–723 (1974).
92. Shirk, A. J., Landguth, E. L. & Cushman, S. A. A comparison of regression methods for model selection in individual-based landscape genetic analysis. *Mol. Ecol. Resour.* **18**, 55–67 (2018).
93. Peterman, W. E. ResistanceGA: An R package for the optimization of resistance surfaces using genetic algorithms. *Methods Ecol. Evol.* **9**, 1638–1647 (2018).
94. Peterman, W. E., Connette, G. M., Semlitsch, R. D. & Eggert, L. S. Ecological resistance surfaces predict fine-scale genetic differentiation in a terrestrial woodland salamander. *Mol. Ecol.* **23**, 2402–2413 (2014).
95. Bauman, D., Drouet, T., Fortin, M.-J. & Dray, S. Optimizing the choice of a spatial weighting matrix in eigenvector-based methods. *Ecology* **99**, 2159–2166 (2018).
96. Dray, S. et al. adespatial: Multivariate Multiscale Spatial Analysis. R package v.0 (2019).
97. Mardia, K. V. Some properties of classical multi-dimesional scaling. *Comm. Stat. Theory Methods* **7**, 1233–1241 (1978).
98. Schubert, E. & Rousseeuw, P. J. in *Similarity Search and Applications. SISAP 2019. Lecture Notes in Computer Science* Vol 11807 (eds Amato, G. et al.) 171–187 (Springer, 2019).
99. Kahle, D. & Wickham, H. ggmap: spatial visualization with ggplot2. *R J.* **5**, 144–161 (2013).
100. Leppälä, K., Nielsen, S. V. & Mailund, T. admixturegraph: an R package for admixture graph manipulation and fitting. *Bioinformatics* **33**, 1738–1740 (2017).
101. Terhorst, J., Kamm, J. A. & Song, Y. S. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat. Genet.* **49**, 303–309 (2017).
102. Choi, J. Y. & Purugganan, M. D. Evolutionary epigenomics of retrotransposon-mediated methylation spreading in rice. *Mol. Biol. Evol.* **35**, 365–382 (2018).
103. 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).
104. *Global Historical Climatology Network-DAILY (GHCN-Daily) version 3* (NOAA National Climatic Data Center, 2012); https://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets/global-historical-climatology-network-monthly-version-3
105. Edwards, M. *Data Announcement 88-MGG-02: Digital Relief of the Surface of the Earth* (National Oceanic and Atmospheric Administration and National Geophysical Data Center, 1988).

Acknowledgements

We thank our colleagues for helpful discussions on this project. This work was supported in part by Zegar Family Foundation grant A16-0051-004 and US National Science Foundation Plant Genome Research Program grant IOS-1546218 to M.D.P., Portugal Fundação para a Ciência e a Tecnologia grant EXPL/BIA-BIC/0947/2012 to S.N., SFRH/BD/68835/2010 to I.S.P. and UID/Multi/04551/2013 to M.M.O., Gordon and Betty Moore Foundation and Life Sciences Research Foundation grant GBMF2550.06 to S.C.G., US National Science Foundation grant PRFB 1711950 to E.S.B., Natural Environment Research Council UK grant NE/N010957/1 to C.C.C. and D.Q.F., US National Science Foundation grant BCS-1632207 to J.A.d.G. and United States Department of Agriculture and National Institute of Food and Agriculture grant 2019-67009-29006 to J.R.L.

Author contributions

R.M.G. and M.D.P. conceived and designed the study with input from J.R.L. and S.C.G. J.Y.C., I.S.P. and O.W. generated sequencing data. M.D.P., S.N. and M.M.O. supervised laboratory work. R.M.G. assembled and processed the sequencing data. S.C.G. and E.S.B. assembled and processed the environmental data with input from J.R.L. J.R.L. led the spatial analyses with input from R.M.G. E.S.B. and E.R.S. carried out travel-time analyses with input from J.R.L. J.R.L. carried out R.D.A. analyses. R.M.G. carried out population-structure, admixture-graph and coalescence analyses. R.K.B. and J.A.d.G. conducted thermal-niche modelling. D.Q.F., C.C.C. and J.A.d.G. provided archaeological context. M.D.P., R.M.G. and J.R.L. wrote the manuscript with input from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41477-020-0659-6.

Correspondence and requests for materials should be addressed to J.R.L. or M.D.P.

Peer review information *Nature Plants* thanks Laura Botigué and Angélica Cibrián-Jaramillo and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Corresponding author(s): Michael D. Purugganan
Jesse R. Lasky

Last updated by author(s): Mar 27, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Illumina cBot Control Illumina HiSeq Control/RTA Illumina bcl2fastq
Data analysis	Custom codes: Sequence data processing - https://github.com/grafau/NextGatkSNPs Discretization - https://github.com/grafau/discretize Travel distances - https://github.com/em-bellis/riceTravelTime Niche modeling - https://github.com/bocinsky/gutaker2019 Published software: Nextflow v.0.25.1.4460 BWA v.0.7.15 PICARD v.2.15.0 GATK v.3.8 BGZIP TABIX v.0.2.5 BEAGLE v.5.0 EEMS PLINK v.1.90b4 SMC++ v.1.15.1 ADMIXTOOLS v.4.1 package RASTER v.2.8-19 [R v.3.5.1] package GDISTANCE v1.1 [R v.3.5.1] package REEMSPLOTS v.0.0.0.9 [R v.3.4.3] package VEGAN v.2.5.4 [R v.3.4.1]

package ADESPATIAL [R v.3.6.0]
package RESISTANCEGA [R v.3.6.0]
package STATS v.3.4.1 [R v.3.4.1]
package CLUSTER v.2.0.6 [R v.3.4.1]
package GGPLOT2 v.2.2.1 [R v.3.4.1]
package ADMIXTUREGRAPH v.1.0.2 [R v.3.4.1]

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Sequencing data :
Available under following SRA Bioproject accession numbers: PRJEB6180, PRJNA422249, and PRJNA557122
All Figures presented in this manuscript were constructed from these raw data
There are no known restrictions on these data availability

Metadata:
Available under following databases:
Coordinates http://iris.irri.org/,
Cultivation system https://www.genesys-pgr.org/welcome [v.2.3],
Environment http://chelsa-climate.org/ [v.1.2],
Eleveation https://www.worldclim.org/ [v.1.4],
Evapotranspiration http://www.cgiar-csi.org [v.2],
Soil http://webarchive.iiasa.ac.at/Research/LUC/External-World-soil-
592 database/HTML/index.html?sb=1 [v.1.2],
Language https://glottolog.org/ [v.4.1]
These raw data were used to generate Figure 1, Supplementary Figures 5, 7, 9 , and Suppelmentary Table 1

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We obtained whole genome re-sequencing data from rice landraces/traditional varieties across a wide geographical distribution in Asia. Landraces, unlike elite cultivars, are associated with sustained cultivation in specific geographic localities and cultural contexts, usually exhibiting local adaptations. We have utilized available data for environmental conditions at the sites where our genomes were sampled.
Research sample	We have used all available high-quality (~10x genomic coverage) sequencing data for rice landraces from Asia. Majority of those were taken from Wang et al. 2018 (Nature). Additional 178 landraces were sequenced to ~10x coverage for this study.
Sampling strategy	Genomes sequenced for this study were a random sample of landraces that showed highest diversity in International Rice Research Institute Genebank and cover the whole geographic distribution.
Data collection	We analyzed genomes of plant material stored in Genebank at International Rice Research Institute Genebank. Rice seeds were collected by multiple independent collectors working for IRRI. Metadata was collected by authors that are cited and referred in this manuscript in Materials and Methods.
Timing and spatial scale	We analyzed genomes of plant material stored in Genebank at International Rice Research Institute Genebank. IRRI started maintaining seed collection since 1962 and is continuing to expand it's collection. Samples used in this study were collected between 1962 and 2015.
Data exclusions	We have excluded all individuals that did not match our criteria of landraces/local variety (see Materials and Methods for details). In spatial analyses (RDA, EEMS, IBD) we have additionally excluded individuals that did not posses precise geolocation.
Reproducibility	Data acquisition can be reproduced by requesting plant material from International Rice Research Institute Genebank. Sample IDs are listed in Supplementary Table 1.

Randomization	Our study has no experimental component and there are no experimental groups in its design. Our study is a retrospective take on evolutionary history of rice. Randomization occurred at the point of collecting plant material by multiple independent collectors over many years to store in International Rice Research Institute Genebank.
Blinding	Genetic analyses performed in this study are sufficiently objective that investigator blinding is not required.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging