Published version in *Nature Plants* 6*:* 492–502 (2020)

DOI: 10.1038/s41477-020-0659-6

Title:

Genomic history and ecology of the geographic spread of rice

Authors:

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]*, Michael D. Purugganan[1,11,12]*

[1]Center for Genomics and Systems Biology, New York University, New York, NY 10003 USA;

[2]Department of Biology, Pennsylvania State University, University Park, PA 16802 USA;

[3]Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa, Av. da República, 2780-157 Oeiras, Portugal;

[4]Crow Canyon Archaeological Center, Cortez, CO 81321, USA;

[5]Carnegie Mellon University Libraries, Pittsburgh, PA 15213-3890;

[6]Department of Biological Sciences, University of Manitoba, Winnipeg, MB R3T 2N2, Canada;

[7]Institute of Archaeology, University College London, London, WC1H 0PY, United Kingdom;

[8]School of Cultural Heritage, North-West University, Xi'an, Shanxi 710069, China;

[9]School of Biology and Environmental Science, University College Dublin, Belfield, Dublin 4, Ireland;

[10]Department of Anthropology and Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA 92093, USA;

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

[12]Institute for the Study of the Ancient World, New York University, New York, NY 10028.

*Correspondence to: mp132@nyu.edu, jrl35@psu.edu.

Summary:

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

Main Text:

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 where crop species originated, and attention has focused on the dynamics of the domestication process, and the evolutionary genetics of crop origins and divergence[1]. In contrast, relatively little attention has been focused on the dispersal and diversification of crops from their center(s) of origin, and the accompanying evolution of adaptive traits that allow 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 > 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, likely requiring local water, temperature and photoperiod adaptation. Rice is grown in lowland ecosystems under paddy, deep-water, or seasonal flood conditions, as well as in upland rainfed areas[5].

Archaeological evidence[6–8] indicates that cultivation of japonica rice began ~9,000 years before present (yBP) in the lower Yangtze Valley, while proto-indica rice cultivation started >5,000 yBP 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 ~4,000 years ago led to introgression of domestication alleles into proto-indica or local *O. 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 5 millennia, providing sustenance for emerging Neolithic communities in East, Southeast and South Asia[14]. Archaeological data shows the general directionality of rice dispersal[9,15]; the details of dispersal routes, times, and the environmental forces that shaped dispersal patterns, however, remain unknown. Here, 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 allows a robust reconstruction of the dispersal history of *Oryza sativa*.

Results

Structure of rice genomic diversity. To investigate the pattern and timing of dispersal of rice, 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. Our sample set includes 1,265 samples from the Rice 3K Genome Project[3,16] and additional 178 landraces sequenced for this study (Supplementary Table 1); the panel consists of 833 indica, 372 japonica, 165 *circum*-aus, 42 *circum*-basmati, and 31 unclassified samples. We identified ~9.78 million single nucleotide polymorphisms (SNPs) with 9.63x 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 >36% of the total variation (AMOVA, permutation $P < 0.001$)[17], congruent with results from multidimensional scaling (MDS) of genomic distances (Supplementary Fig. 2a). Only japonica and indica have wide geographic distributions (Fig. 1 a and b; 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); henceforth, 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 explains genetic distance much less in the Malay Archipelago (*i.e.* islands SE Asia) compared to 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 Himalayas), and the South China Sea which reduces movement between Borneo/Philippines and mainland Southeast Asia (Fig. 1a and b; Supplementary Fig. 6).

To improve on the IBD model, we decided to take into account actual travel times between locations rather than simple geographic distances; for human-dispersed species such as crops, genetic distances may correlate better with travel resistance, 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 (IBR) 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 (archipelago

ΔAIC = -34, mainland ΔAIC = -17), but not for indica (archipelago ΔAIC = +51, mainland ΔAIC = +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 unconscious cultural preferences arisen from language barriers), grain stickiness (as proxy for conscious cultural preferences), and genetic composition of proximal wild rice populations (Figs. 1c and d; 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 to 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 inter-annual variation in growing degree days (IAG), and low minimum temperatures early in the growing season (ESM; Fig. 1e; Supplementary Fig. 9a). Temperate landraces from upland rainfed ecosystems are further characterized by alleles associated with inter-annual variation in precipitation (IAP; Fig. 1e).

For indica, the first two axes also grouped individuals by their geographic origins (Fig. 1f; Supplementary Fig. 8b). Similar to japonica, indica Malay Archipelago landraces contain alleles associated with high precipitation prior to the growing season (PSP). Mainland Southeast Asian genotypes are characterized by alleles associated with warm minimum growing season temperatures (WSM) and presence of nearby freshwater sources (DMF Fig. 1f; Supplementary Fig. 9b). The latter contrasts with 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 inter-annual variation in precipitation (IAP).

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 based on genomic distances by partitioning-around-medoids (PAM)[26], identifying the number of subpopulations (k) and subsequently applied silhouette-based procedure (see Methods) to identify number of discrete subpopulations (k_d). This discretization procedure removed genetic gradients between subpopulations (Fig. 2a and 2b; Supplementary Figs. 10 and 11). We compared PAM clusters to 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 F_{ST} values among japonica subpopulations (Supplementary Fig. 15), suggesting fewer past migrations compared to indica, and/or older establishment of its population structure. Finally, subpopulations of both subspecies clearly correspond with geography (Fig. 2c and 2d; 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 analyzed japonica and indica separately.

We reconstructed relationships between japonica subpopulations at $k_d = 2$ to 9 considering graphs with population f-statistic z-scores <3. Throughout all k_d levels, we find two similar and consistent graph topologies (Fig. 2e; 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 allows 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, and 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 ~9,000 and 4,000 yBP[29,30]; this coincides with the rise in frequency of non-shattering rice from ~20% just after 8,000 yBP to fixation at ~5,000 yBP[7,8].

The first major population divergence in japonica separates temperate from tropical landraces (Supplementary Figs. 10 and 16). Using sequentially Markovian coalescent (SMC++), we estimated a cross-coalescence split time between temperate and tropical japonica at ~5,000 to 1,500 years ago, with 90% of estimates between ~4,100 to 2,500 years ago (Fig. 3a; Supplementary Fig. 17). Using dated archaeobotanical rice remains[15], we note that rice agriculture spread north- and eastward along the Huang He river[31] and westward into the Chengdu Plains and the Southwest China Highlands between ~5,000 to 4,000 yBP[32–34] (Fig. 3b; Supplementary Fig. 18). During a minor climatic cooling event at ~5,000 yBP, rice appears maladapted in parts of eastern China[35]. In the Shandong Peninsula, rice disappeared by 5,000 yBP and briefly re-emerged 4,500 yBP as a short-grained variety similar to contemporary temperate japonicas[36]. A global temperature decrease that followed the HCO at

~4,200 years ago, the '4.2k event'[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 ~4,100 years ago (Fig. 3a; Supplementary Fig. 17).

Temperate adaptation created opportunity for northeastern dispersal of japonica in Asia. From our demographic analysis of temperate japonica we note a ~5-10-fold Ne reduction between ~3,500 to 3,000 yBP (Fig. 3c; 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; its contemporary descendants however, 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 tropical japonica lineage experienced a ~50-100-fold population (Ne) contraction between ~4,500 to 4,000 yBP, and partial Ne recovery starting ~2,500 yBP (Fig. 3d, 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 ~4,500 yBP and a shift to rainfed, upland cultivation[39].

Gradients of heat accumulation are highly associated with geographic distribution of japonica genomic diversity (Fig. 1e). Based on reconstruction of Holocene temperatures[40] we show that despite substantial temperatures changes, the spatial heat accumulation gradients, measured as growing degree days (GDD), remained stable in 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; Supplementary Fig. 21). Survival probabilities of tropical japonica between ~4,400 and 3,500 yBP dropped dramatically in eastern China and high-altitude South China (survival probability < 50%) compared to Southeast Asia [survival probability > 90%](Fig. 3e; Supplementary Video 1). Indeed, after the cooling period we observe high densities of archaeological rice remains in Southeast Asia (Fig. 3b; 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 ~50-100-fold population contraction in the remote upland (Bhutan) rice population between ~4,000 and 3,000 yBP (Fig. 4; 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 coincides in time and space with widespread establishment of rainfed rice agriculture in mainland Southeast Asia, ~4,000 yBP[14,42] and dispersal of metallurgy traditions from Bronze Age Yunnan, ~3,500 yBP southwards to Thailand by ~3,000 yBP[43,44]. Subsequent agricultural intensification of rice production took place from ~2,500 to 1,500 yBP 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 yBP[46,47].

Our analysis also shows an ~5-10-fold Ne decrease in the Malay archipelago between ~3,000 and 2,500 yBP, and based on cross-coalescence analyses, divergence between mainland and Malay Archipelago rice occurred between ~3,000 to 1,500 years ago (90% of estimates in ~2,500 to 1,600 yBP) [Fig. 4; Supplementary Fig. 17]. Distinct island populations in the Malay Archipelago diverged at around a similar timeframe, in an interval from ~3,000 to 1,000 years ago (90% estimates fall between ~2,500 and 1,500 yBP). This period coincides with dispersal of Dong Son drums in the Malay Archipelago (~2,400 years ago)[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 ~2,000 years ago[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 ~1,200 yBP[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, while 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). Interestingly, 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, likely 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 likely the reason behind difficulties with indica dispersal routes reconstruction. 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 ~2,500 and 1,100 yBP (90% of estimates between ~2,000 and 1,400 yBP)[Fig. 5; 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 ~1,950 yBP[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; in contrast, Southeast Asian varieties are more often rainfed[25].

Indica dispersal to Southeast Asia (e.g., Thailand and Cambodia) were either from India or China (Fig. 5; Supplementary Fig. 23). From archaeobotanical studies, indica arrived in Central Thailand at ~1,800 years ago[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 hence it comes as a surprise that indica mainland subpopulations suffered dramatic population size reduction between ~5,000 and 3,500 yBP (Supplementary Fig. 24). It is even more puzzling that a bottleneck in indica subpopulation in Indonesia occurred between ~6,000 and 5,000 yBP, 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 moved from anomalies which were 0.4 degrees Celsius above present day, to 0.2 degrees cooler than 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 for 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 years 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 are 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].

Indica rice began to be domesticated in South Asia at around the time of the 4.2k event, and spread later into China and Southeast Asia. The spread of indica rice occurred much later than 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 remains 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.

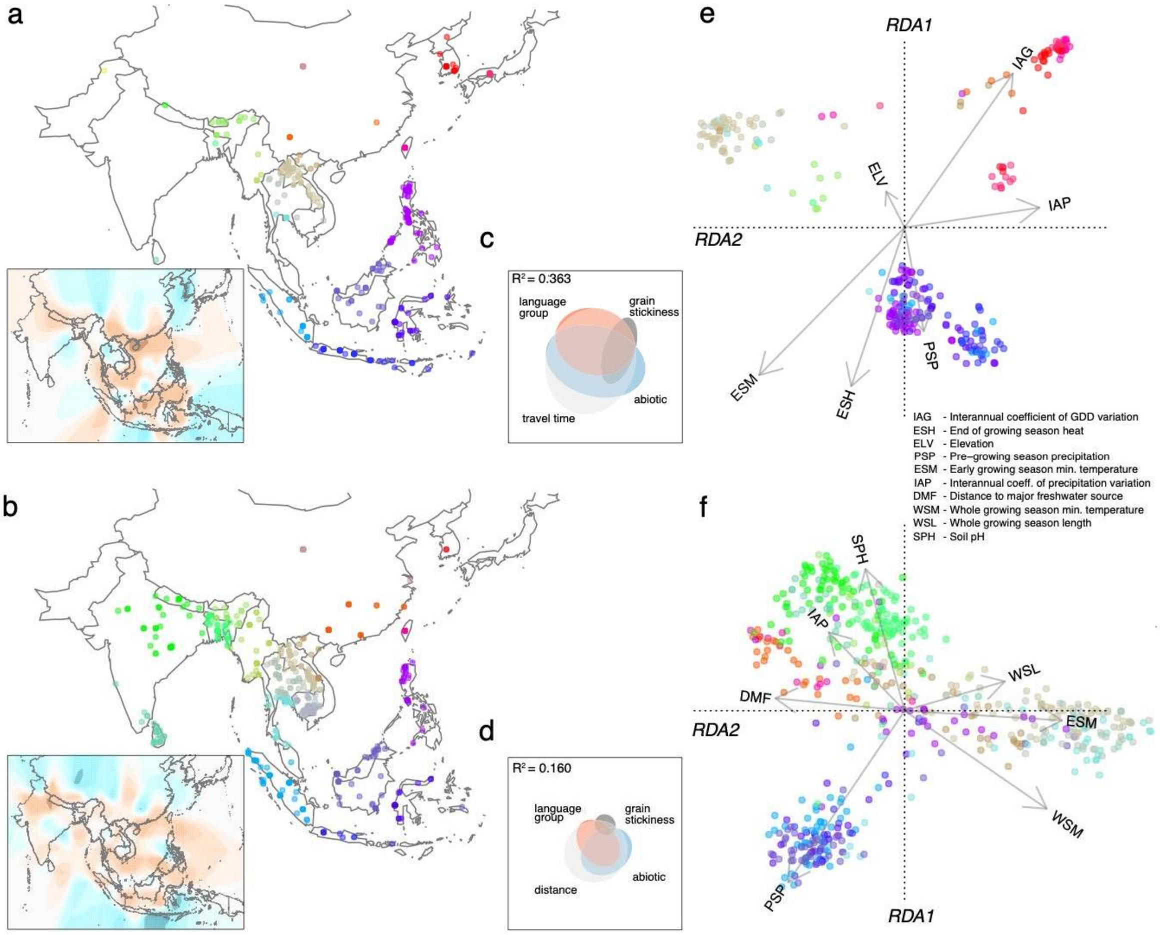

Figure 1: Factors underlying geographic distribution of genomic diversity in japonica and indica. Maps of collection sites for (a) japonica and (b) indica landraces used in this study. Colors represent regions of origin. In insets are effective migration surfaces representing migration barriers (orange) and channels (cyan). (c) Japonica and (d) indica genomic diversity is best explained by a combination of four factors represented in Euler plots: 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 particular group of variables calculated using variance partitioning with redundancy analysis ordination. (e) Japonica and (f) indica genotypes projected on the first two canonical axes of redundancy analysis. Arrows represent environmental predictors (acronyms explained in the legend) that strongly correlate with a maximal proportion of linear combinations of SNPs.

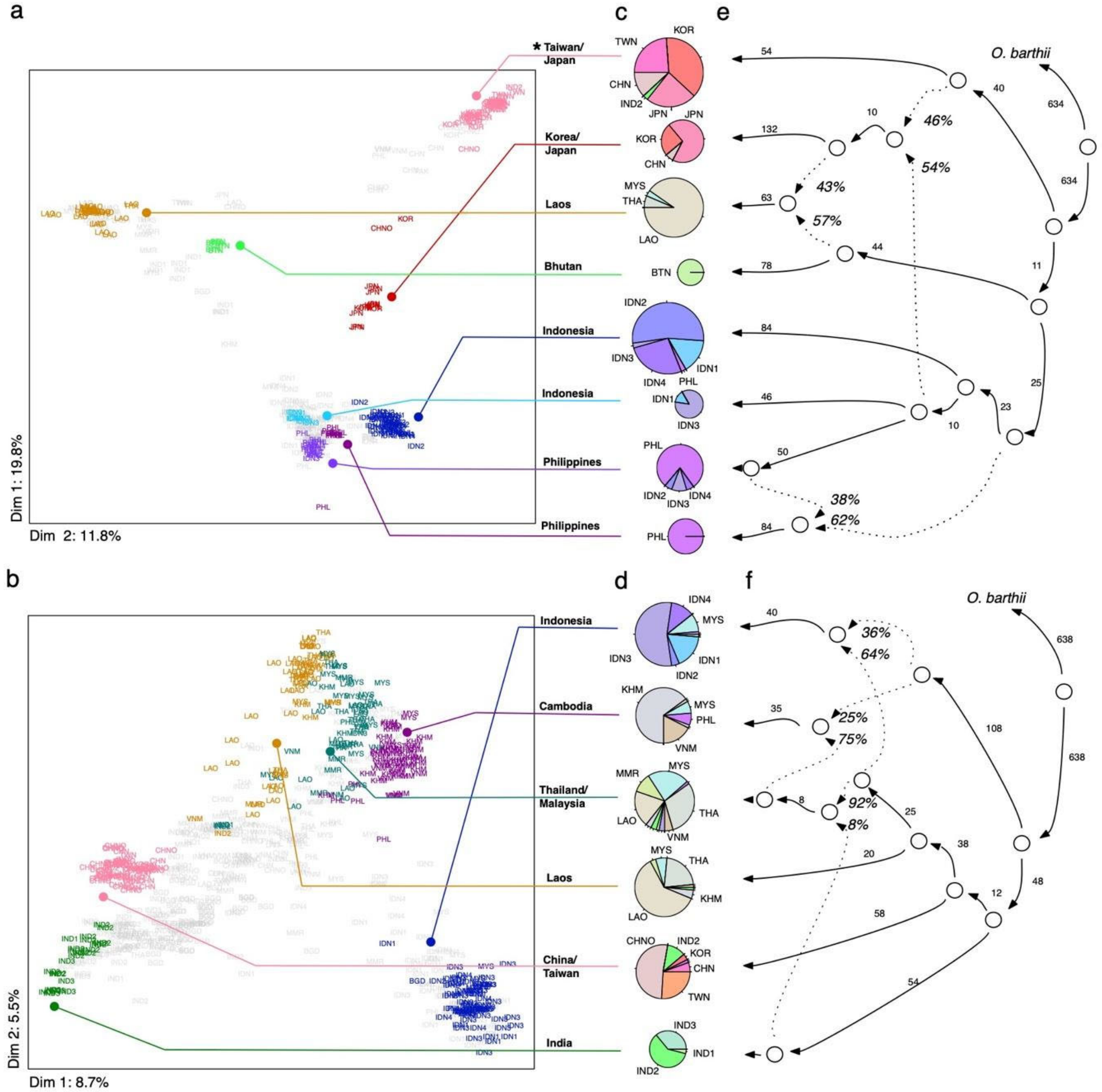

Figure 2: Japonica and indica rice subpopulations. (a) All japonica and (b) indica landraces projected onto first two dimensions after multidimensional scaling of genomic distances. (a) japonica genotypes were clustered using k-medoids (k = 9 subpopulations) and filtered using silhouette parameters, which resulted in k_d = 8 discrete subpopulations (colored labels). Asterisk denotes 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 (colored labels). Pie charts representing the geographical composition of each discrete subpopulation of (c) japonica and (d) indica 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 consistent between models for all lower k's. (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 k's, likely due to complex history of indica. (e and f) Solid lines with arrowheads represent uniform ancestries (attached numbers show scaled drift parameter f_2), while dashed lines represent mixed ancestries (% values indicate estimated proportion of ancestry).

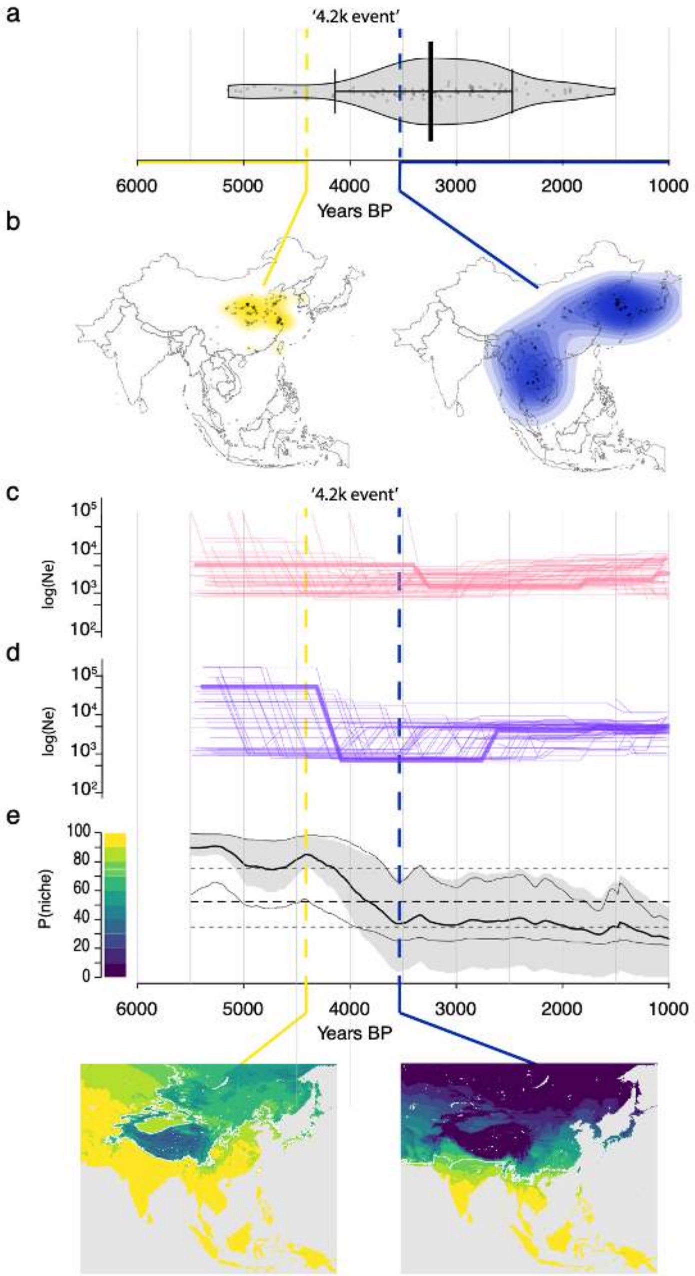

Figure 3: Demographic, paleoenvironmental and archaeological context of temperate japonica rice emergence. (a) The distribution of temperate-tropical split times estimated from cross-coalescence analysis carried out for 50 pairs of temperate and tropical individuals; bar represents mean, bands represent 90% interquartile range. (b) Maps indicating geographic locations and densities of archaeological sites with rice macro-remains. To the left: cumulative archaeobotanical evidence from 9,000-4,400 years BP, to the right: cumulative archaeobotanical evidence from 3,500-1,000 years BP. Effective population sizes over time in (c) tropical and (d) temperate japonica subpopulations. Thin lines represent demographic histories for 50 randomly sampled individuals, while bold lines represent joint models. (e) Probability of tropical rice being in the thermal niche (assuming requirement of 2900 growing degree days, at 10°C base) over time. The mean (thick black line) and the interquartile range, 25% to 75% (gray shaded area) of probability of being in the thermal niche. The thin black lines are the mean probabilities of being in the thermal niche across the study area when modeled using the 1σ uncertainty intervals as provided by the northern

hemisphere temperature reconstruction. The two inset maps show the geographic distribution of niche probabilities; to the left: before climate cooling (4,400 years BP), to the right: after climate cooling (3,500 years BP).

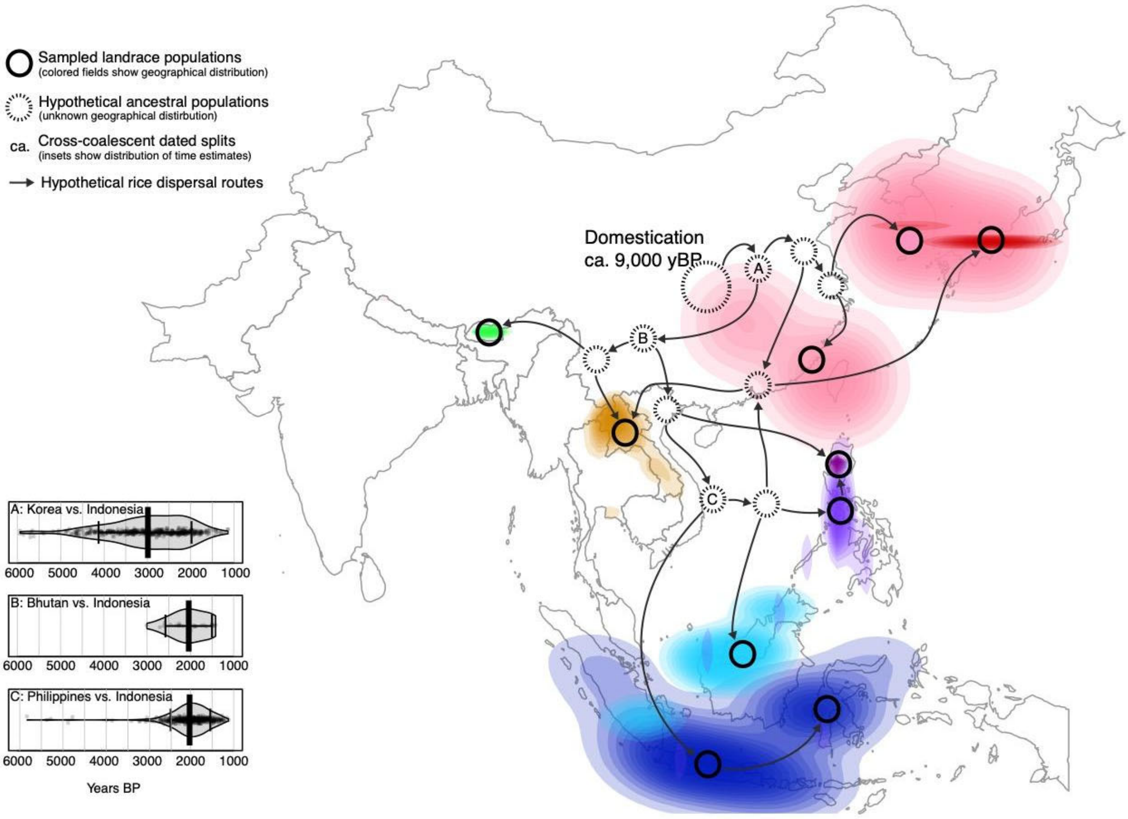
Sampled landrace populations
(colored fields show geographical distribution)
Hypothetical ancestral populations
(unknown geographical distribution)
ca. Cross-coalescent dated splits
(insets show distribution of time estimates)
Hypothetical rice dispersal routes
Domestication
ca. 9,000 yBP
A
B
C
A: Korea vs. Indonesia
B: Bhutan vs. Indonesia
C: Philippines vs. Indonesia
6000 5000 4000 3000 2000 1000
Years BP

Figure 4: Proposed dispersal map of japonica rice in Asia. Map generated for japonica, k_d = 8 discrete subpopulations. The geographic distributions of subpopulations were represented as colored, two-dimensional Kernel density fields. Bold circles represent leaves in the admixture graphs and are mapped close to the centers of subpopulation distributions. Dashed circles represent hypothetical ancestral subpopulations inferred from splits in best-matching admixture graphs; their precise geographic placement is uncertain. The distribution of split times between non-admixed subpopulations was created from cross-coalescence estimates summarized over all k_d levels and presented as violin plots; bar represents mean, bands represent 90% interquartile range. Arrows indicate hypothetical routes of dispersal.

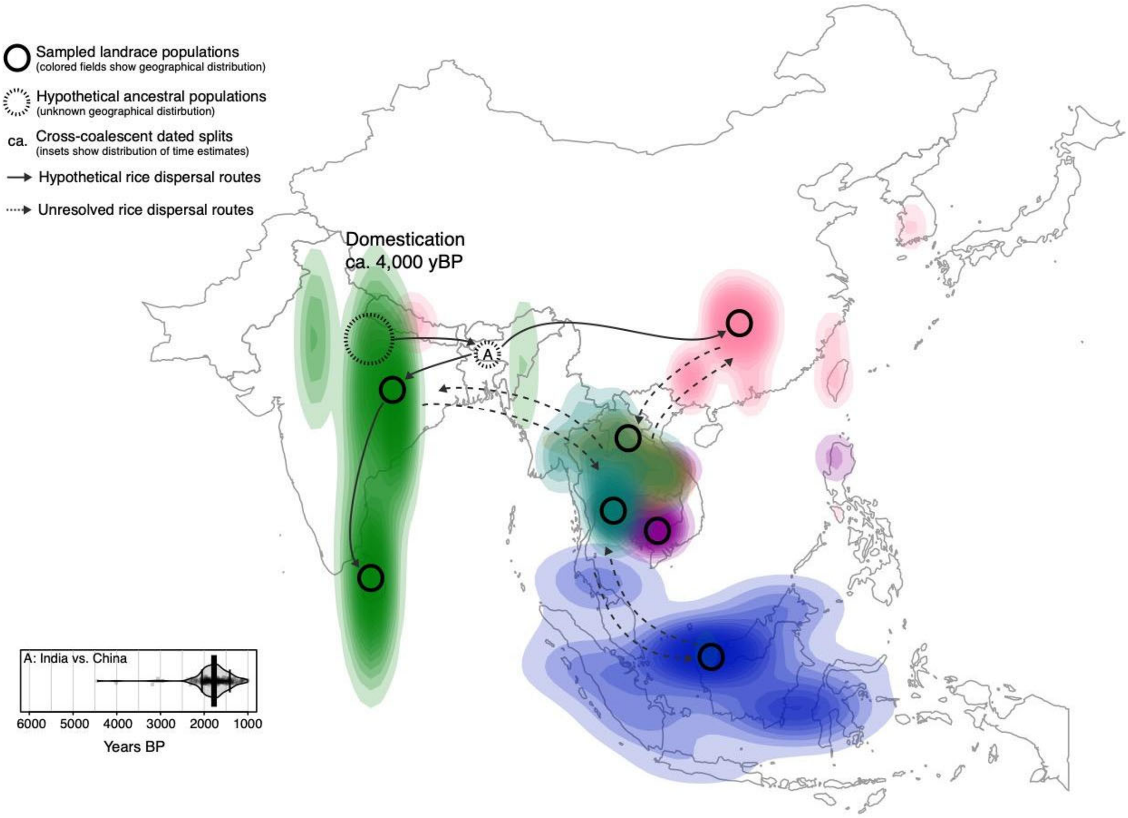
Sampled landrace populations
(colored fields show geographical distribution)
Hypothetical ancestral populations
(unknown geographical distribution)
ca. Cross-coalescent dated splits
(insets show distribution of time estimates)
Hypothetical rice dispersal routes
Unresolved rice dispersal routes
Domestication
ca. 4,000 yBP
A
A: India vs. China
6000
5000
4000
3000
2000
1000
Years BP

Figure 5: Proposed dispersal map of indica rice in Asia. Map generated for indica, k_d = 6 discrete subpopulations. The geographic distributions of subpopulations were represented as colored, two-dimensional Kernel density fields. Bold circles represent leaves in the admixture graphs and are mapped close to the centers of subpopulation distributions. Dashed circle represents consistent split; its geographic position is uncertain. The distribution of split times between non-admixed subpopulations was created from cross-coalescence estimates summarized over all k_d levels and presented as violin plots; bar represents mean, bands represent 90% interquartile range. Solid arrows indicate hypothetical routes of dispersal, while dotted arrows indicate possible routes that remain unresolved from admixture graphs.

Online methods

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

The definitions of landrace are very complex[57,58] and hard 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. To that end we used the following criteria: 1) Pre-selected 'candidate' landraces from available annotation, and 2) filtered them based on their joint genetic and geographic clustering.

Accession passport data were obtained from the International Rice Information System (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 that 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 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]. Based on their agro-ecosystem of origin, accessions are 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 through 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 ‘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 was 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 that 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 $k = 6$[63]. Resulting ancestry probabilities do not represent any biological individuals, but rather a most likely hypothetical wild relative in the area of rice cultivation.

We also considered rice-human relationships not covered by geographic distance/resistance dispersal. One important rice grain property 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 modeling the general ability of people to talk to each other and trade seeds. To this end 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), growing degree days [GDD] in growing season(s), and inter-annual 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 inter-annual coefficient of variation of precipitation) were derived from CHELSA climatological data (v1.2), which provides monthly and mean annual precipitation and temperature data at 30 arc-second resolution for the time period 1979 to 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-CSI 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 v1.2 (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 section 1) from the 3K-RG project were downloaded in fastq format from the Short Read Archive (SRA) using FASTQ-DUMP tool with option to split reads into forward, reverse and trimmed.

We generated sequencing data for 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, Hilden, Germany). Yields ranged between 3 ng/ul and 102 ng/ul. 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 – 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] 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 eight 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 GATK's GENOTYPEGVCFS function 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 only kept 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 GATK's VARIANTRECALIBRATOR function referencing a true-positive set of SNPs that were discovered independently in the 3K-RG project[3], and in the rice diversity panel (RDP) that was genotyped with a high-density SNP array[80]. We applied the dynamic filter to our raw set of SNPs using GATK's APPLYRECALIBRATION function 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 equals 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-square 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 mis-mapped reads in chromosomal regions with structural variants that are present in the re-sequencing 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 lower than 80% genotyping rate 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 section 1) to annotate each candidate landrace with country of origin and subgroup designation (indica, japonica, *circum*-aus, *circum*-basmati) provided by the 3K-RG and RDP[1,75] 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 2). 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 mislabeling. 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 5 million Monte-Carlo Markov Chain iterations of which the first 2 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]. Traveling speed over land given the slope between adjacent grid cells was calculated according to Tobler's Hiking Function[87], based on elevation data at 30 arc second resolution from WorldClim v1.4. 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 degree to reduce 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 linear mixed model with maximum likelihood estimation of Clarke *et al.*[90] and used Akaike Information Criterion (AIC)[91] to select between geographic distance versus 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 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' and a group of 'archipelago' landraces. Mainland landraces were defined to include those north of 9.7ºN latitude and west of 110ºE longitude, thus excluding the relatively small number of isolated mainland landraces to the east (e.g. eastern China). Archipelago landraces included those from the Malay Archipelago and the Malay Peninsula, but not from the islands to the north (*i.e.*, Taiwan or Japan).

Redundancy analyses. Redundancy analyses (RDA) are eigenanalyses 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 (a)biotic 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 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 neighboring 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 (*i.e.*, 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/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 travel time for japonica, because model selection favored these distance measures. For indica and geographic distance, optimization selected 25 MEM eigenvectors from the connectivity matrix based on 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 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 in the main text indicate those loading most strongly in each direction along each RDA canonical axis as well as those loading most strongly in each diagonal (identified by multiplying the loadings on the first two canonical axes). All RDA (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 based on pairwise genetic matrices with the partitioning around medoids (PAM) method[98] implemented as the PAM function in 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 based on 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 co-localized geographic distribution within each discrete cluster. We investigated composition of clusters with regard to region of origin to see if each fulfilled our latter criterion for landrace status. One indica cluster exhibited poor geographic co-localization and was therefore removed from all analyses (see section 1). 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 *Oryza 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 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 Fig. 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 Fig. 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 Sequential Markov Coalescent method (SMC++)[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 through 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 subpopulations. In subpopulations with fewer than 50 individuals assigned, we sampled all of them 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 year 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-year 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].

In order to predict how changing temperatures might have impacted 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 growing degree-days (GDDs) 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 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 GDDs, 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 arc min (c. 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/gutaker2019.

Data availability: Raw FASTQ reads for 178 accessions whose genomes were re-sequenced for this study have been deposited in the Sequence Read Archive under SRA Bioproject accession numbers PRJNA422249 and PRJNA557122. Sources for all downloaded data are referred to in the supplementary materials.

Code availability: Code repositories are referred to in the supplementary materials.

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 Biol. Sci*. **372**, (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**, 11 (2017).

14. Fuller, D. Q., Castillo, C. C. & Murphy, C. How rice failed to unify Asia: globalization and regionalism of early farming traditions in the Monsoon World. 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**, 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*. (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*. (2019) doi:10.1093/molbev/msz280.

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. Clustering by means of medoids. *Proc. Statistical Data Analysis Based on the L1 Norm* (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. *Journal of Archaeological Science: Reports* **20**, 711–721 (2018).

35. Fuller, D. Q., Weisskopf, A. R. & Castillo, C. Pathways of rice diversification across Asia. *Archaeology International* **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 University 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. U. S. A*. **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. The spread and adaptation of Buddhism in early medieval China. *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.* (IRRI, 1997).

61. Maclean, J., Hardy, B. & Hettel, G. *Rice Almanac, 4th edition: source book for one of the most important economic activities on Earth.* (IRRI, 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. Glottolog4.0. (2019) doi:10.5281/zenodo.3260726.

66. Hijmans, R. J. & van Etten, J. raster: Geographic data analysis and modeling. *R package version 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) doi: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 (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. *arXiv [q-bio.GN]* (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**, 11.10.1–33 (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. Team, R. C. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013. (2014).

84. Oksanen, J. Vegan: an introduction to ordination. *URL http://cran. r-project. org/web/packages/vegan/vignettes/introvegan. pdf* 8, 19 (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. *Journal of Statistical Software* **76** (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* (93-1). (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*. (University of 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 version 0.3-7*. (2019).

97. Mardia, K. V. Some properties of classical multi-dimesional scaling. *Communications in Statistics - Theory and Methods* **7**, 1233–1241 (1978).

98. Schubert, E. & Rousseeuw, P. J. Faster k-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms. *arXiv [cs.LG]* (2018).

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. U. S. A.* **93**, 10274–10279 (1996).

104. Menne, M. J. *et al.* Global historical climatology network-daily (GHCN-Daily), Version 3. *NOAA National Climatic Data Center* **10**, V5D21VHZ (2012).

105. Edwards, M. Data announcement 88-MGG-02: digital relief of the surface of the earth. (National Oceanic and Atmospheric Administration, National Geophysical Data Center, Boulder, CO, USA, 1988).

Acknowledgments:

We would like to 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 MDP, PortuguesePortugal Fundação para a Ciência e a Tecnologia grant #EXPL/BIA-BIC/0947/2012 to SN, SFRH/BD/68835/2010 to ISP and UID/Multi/04551/2013 to MMO, Gordon and Betty Moore Foundation/Life Sciences Research Foundation grant GBMF2550.06 to SCG, US National Science Foundation grant PRFB #1711950 to ESB, Natural Environment Research Council, UK grant NE/N010957/1 to CCC and DQF, US National Science Foundation grant BCS-1632207 to JdAG and United States Department of Agriculture, National Institute of Food and Agriculture grant 2019-67009-29006 to JRL.

Author contributions:

RMG and MDP conceived and designed the study with input from JRL and SCG. JYC, ISP and OW generated sequencing data. MDP, SN and MMO supervised laboratory work. RMG assembled and processed the sequencing data. SCG and ESB assembled and processed the environmental data with input from JRL. JRL lead the spatial analyses with input from RMG. ESB and ERS carried out travel time analyses with input from JRL. JRL carried out RDA analyses. RMG carried out population structure, admixture graph and coalescence analyses. RKB and JAdG conducted thermal niche modelling. DQF, CCC and JAdG provided archaeological context. MDP, RMG and JRL wrote the manuscript with input from all authors.

Competing interests: Authors declare no competing interests.

Materials and correspondence:

All requests should be addressed to MDP (sequencing data analyses) and JRL (environmental data analyses).