J. M. Cobo, J. Fort and N. Isern

The spread of domesticated rice in eastern and southeastern Asia was mainly demic

Supplementary Information

Index

S1. Database, geostatistical analyses and front speed

In the main paper we analyze a database of 201 sites or phases (for sites with several phases, we use the oldest phase). For each phase, if direct radiocarbon dates are available, the median of the 95% distribution is used. Otherwise, i.e. if the date has been estimated by cultural association, we cannot determine the exact moment (within the phase) to which the archaeological evidence for rice corresponds, and the mean date of the phase has been used (see also Silva et al., 2015). Of these 201 sites, 23 have dates obtained by direct accelerator mass spectroscopy, 58 associated C14 dates from the phase/site, 115 were dated by cultural association, and 1 by optically stimulated luminescence. The original database (Silva et al., 2015) does not indicate the method used in the remaining 4 sites. Thus 115 sites (or 57%) were dated by inference from the depositional context and associated artifacts. Admittedly, this makes their dates less precise. Nevertheless, we have decided not to exclude them. The reasons are that if we did, the statistical validity (correlation coefficients) of the results would not increase, and the level of detail in the interpolation maps would decrease substantially. However, we have checked that excluding all phases dated by cultural association would not change the conclusion in the main paper, namely that demic diffusion was more important than cultural diffusion (Sec. S1f). Another limitation of the database is that, as admitted by the authors of a previous version (Fuller et al., 2011), there are very few sites where explicit reporting of morphological details allows an assessment of morphological domestication status. We would like to stress that, in the future, additional sites and more precise archaeological dates will lead to improved databases and, therefore, to more exact results than those reported in the present paper. It is reasonable to expect, however, that the first step presented here may remain useful, similarly to the fact that the first estimation of the spread rate of the Neolithic in Europe (Ammerman and Cavalli-Sforza, 1971), which was based on 53 sites, remains consistent with the results of improved databases, which became available only

several decades later and contain more precise dates and about 1,000 sites (Pinhasi et al., 2005; Fort et al., 2012).

As in our previous work on the spread of the Neolithic in Europe (Fort, 2015), spatial interpolations tend to display small areas that disagree with the overall pattern, i.e. that are anomalously old or young (see e.g. Fig S2a in Fort, 2015). Thus, we perform a smoothing technique that simply averages the interpolated dates over nearby values, as in our previous work (Fig S3 in Fort, 2015). This attempts to remove artificial variations due, e.g., to dating and calibration errors (for an example of how such artificial variations affect the speed vectors, see Figs S2a-c in Fort, 2015). This effect can be especially important if, as in the present paper, the dates of some sites have been estimated by cultural association. A previous, very detailed analysis has shown that, without smoothing, local speed vector and magnitude maps cannot be trusted (compare the non-smoothed Figs S1b-c to the smoothed Figs S4b-c in Fort, 2015). Anyway, smoothing does not have any effect on our results and conclusions, because in the present paper geostatistical maps (Figs 1, S1a, S2-5) have only visualization purposes, whereas the front speed (and, hence, the percentages of demic and cultural diffusion) are estimated using linear regressions (for a detailed justification on the advantages of linear regression over geostatistical methods for the specific database used here, see Sec. S1d-B below). Smoothing is performed 20 times, because repeating the process several additional times introduces negligible changes. The same technique was applied in previous work (see Fig 4a in Fort, 2015).

As it is well-known, when a straight line is fitted to a set of data, the maximum magnitude of the Pearson correlation coefficient is $r = 1$ and corresponds to a perfectly linear relationship. The lower the value of r, the less likely that there is a linear relationship between the variables regressed (time and distance of archaeological sites in our case), and the less reliable the value of the slope of the fitted regression (and the average spread rate estimated from it). The square of the correlation coefficient, r^2, gives the fraction of the variance in the data explained by the linear regression fit. In Neolithic spread rate studies, values $r \geq 0.7$ are usually required to estimate spread rates (Ammerman and Cavalli-Sforza, 1971,

1984; Pinhasi et al., 2005; Fort et al., 2012). If $r < 0.7$, less than 49% of the variance is explained by the linear fit. Then it is usually considered that the confidence in linearity is weak and, therefore, that a statistically sound value for an average spread rate cannot be estimated from the linear fit. In this section S1, we use this approach to interpret results including some sites that have been excluded in the main paper.

Using simulated dispersals, it has been shown that the spatial origin of distances yielding the highest value for the correlation coefficient r may not always be the true dispersal source if $r < 0.3$ (see Fig 3 in Silva et al., 2015). However, in all analyses below we will find that $r > 0.3$, so we can confidently use the highest correlation coefficient to locate the most likely source, as in previous work (Ammerman and Cavalli-Sforza, 1971; Pinhasi et al., 2005).

S1a. Analysis including the Indian subcontinent

In the main paper we have excluded the Indian subcontinent. We discuss this issue in detail here, beginning with the complete Rice Archaeological Database (Silva et al., 2015). It contains 331 sites for which both cultivation and domestication of rice were reported (Supp. Info, S1 Table to the present paper). Figure S1a is an interpolation map of the dates of these sites. As mentioned above, only the oldest phase per site has been used. Note the presence of old (yellow) regions in India and Pakistan. As explained in the main paper, we have excluded this region (the Indian subcontinent) because rice genomic data (Wang et al. 2018) indicate an independent domestication process (*O. sativa* Indica) rather than just introgression from Chinese rice (*O. sativa* Japonica). Here we mention, for the sake of completeness, that the spread rate obtained by regression of dates versus distances using all those 331 sites, with distance origin at Chengtoushan as in the main paper (Fig S1b), is 1.81±0.30 km/yr (95% CL interval) and the correlation coefficient is $r = 0.55$. This value of r is very poor (for comparison, that in the main paper is $r = 0.78$). Such a low value ($r = 0.55$) indicates that only 30% of the variance in the data is explained by the linear fit, so assuming a linear dependence is not justified. In other words, it is not possible to perform a statistically

sound estimation of a constant, average spread rate. In our opinion, the reason may be simply that, as mentioned above, genomic data (Wang et al. 2018) support that in the Indian subcontinent we are dealing with processes independent of the domestication and spread of rice from China. This implies that the dates when *O. sativa* Indica appeared in the Indian subcontinent (or even at several places in it) are independent from the spread of *O. sativa* Japonica from China. Thus, it is not reasonable to expect that a single wave of advance can describe the complete region in Fig S1a. We are probably dealing with several waves of advance of domesticated rice, one from China (*O. sativa* Japonica) and another one (or several) from the yellow regions in the Indian subcontinent in Fig. S1a (*O. sativa* Indica). The former wave of advance is analyzed in the main paper. We discuss the latter in the next subsection.

Finally we mention that the spread rate above (i.e., that implied by the linear regression of the 331 sites in the complete database) is not consistent with the average spread rate obtained using geostatistical techniques (this will be explained in detail in Sec. S1d below). In contrast, using the same reduced database as in the main paper (201 sites, as justified below) leads to agreement between both approaches (see Sec. S1d for details). This again suggests that the database including the Indian subcontinent (331 sites) does not correspond to the spread of a single wave with constant speed, but to several waves with different sources (or to more complicated processes), which obviously cannot be analyzed with a linear regression (including all 331 sites) with a common distance origin.
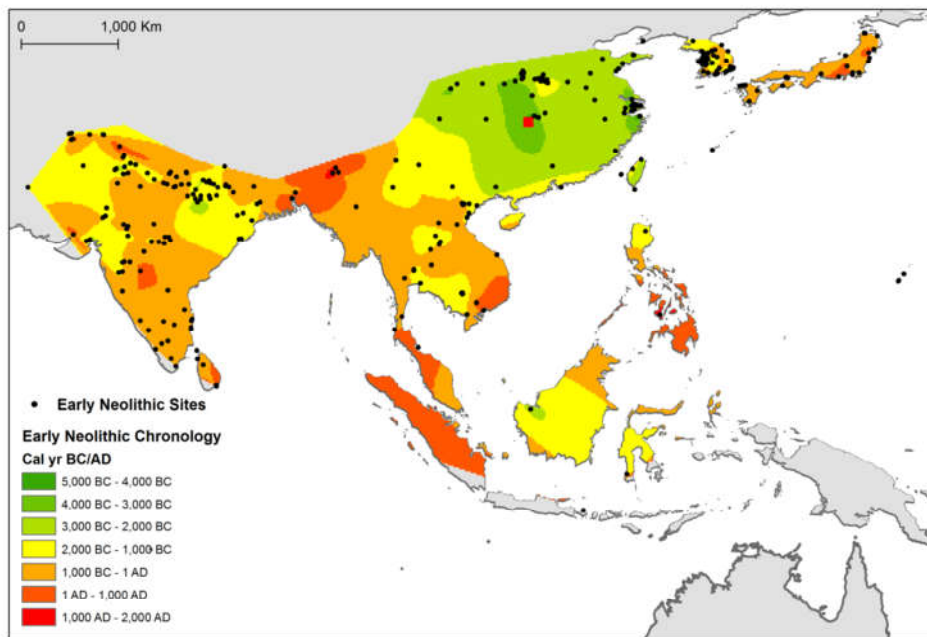
Fig S1a. Map of arrival times of the Neolithic transition in southern and southeastern Asia. Map obtained from a natural neighbor interpolation of 331 early Neolithic sites (symbols) in southern and southeastern Asia (see S1 Table), smoothed 20 times. Color regions correspond to the area advanced by the Neolithic wave every 1,000 years. The red square is the very old site of Chengtoushan (China, dated 4250 BC. The darker green color (5000 BC-4000 BC) in the legend is not seen in the map because only a few sites, namely Hemudu (4800 BC), Chengtoushan (4250 BC), Caoxieshan (4200 BC) and Maojiashan (4000 BC), have dates within this range. Map prepared with ESRI ArcGIS version 10 (http://www.esri.com/) using ArcMap and the ArcGIS Spatial Analyst extension.
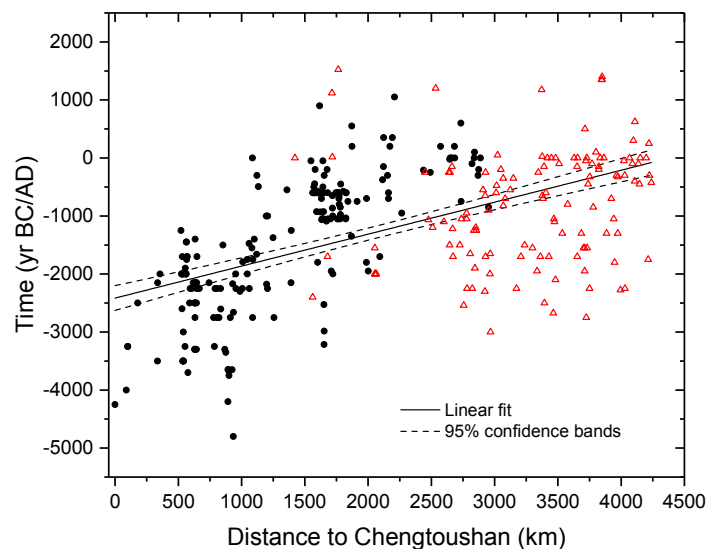

Fig S1b. Linear regression fit to the 331 sites in the map in Fig S1a and S1 Table. Distances are computed from Chengtoushan (China, dated 4250 BC) using great circles. Triangles are the 130 Indian, Pakistani, Sri Lankan, Laotian, Cambodian, Indonesian, Marianan Islands, Malasyan and Philippine sites (they have been excluded in Fig S2 and the main paper). The solid line corresponds to the linear regression fit to all 331 sites, and the dashed lines to the 95% confidence bands. The front speed (inverse of the slope) is 1.81 ± 0.30 km/yr (95% CL interval) and $r = 0.55$.

S1b. Analysis of the Indian subcontinent

The considerations in the previous subsection suggest that it could perhaps be possible to estimate the spread rate of one or several waves of advance of *O. sativa* Indica in the Indian subcontinent, by considering only the 115 sites (S5 Table) in the Rice Archaeological Database (Silva et al., 2015) that are located in India (101 sites), Pakistan (7), Sri Lanka (5) and Bangladesh (2). However, we tried several possibilities and found that the values of the correlation coefficients are too poor. For example, given that there are two old regions in the Indian subcontinent, one in the northeast of India and another one in Pakistan (Fig S1a), we considered a dual-origin model in the Indian subcontinent as follows. We separated the 115 sites in two groups. The first group contains the 96 sites that have lower distance to Kunjhum in Northeast India (the oldest site in India, 3,000 yr BC) than to Ghaleghay (the oldest site in Pakistan, -2,750 yr BC). The second group contains the 19 sites with lower distance to Ghaleghay than to Kunjhum. Note that this is analogous to the dual-origin model for China and southeast Asia proposed in the main paper (Sec. 3.1.3). The correlation coefficient of the regression of the 96 dates versus distances from Kunjhum is only $r = 0.49$, and that of the 19 dates versus distances from Ghaleghay is still worse, $r = 0.28$. With such poor values, unfortunately it is not possible to estimate statistically sound spread rates. Hopefully this situation will improve in the future when the data are more accurate and numerous, but we have to conclude that, at present, it is not possible to estimate spread rates in the Indian subcontinent. Thus, in our main paper we have focused in eastern and southeastern Asia.

S1c. Analysis including Laos and Cambodia

As explained in the main paper, besides the sites in India (101 sites), Pakistan (7) and Sri Lanka (5), we have also excluded those in areas separated by huge distances of Ocean (because they cannot be described by the inland dispersal model), namely the Mariana Islands (3 sites), the Philippines (2), Indonesia (2) and Malaysia (1).

We have also excluded Laos (2 sites) and Cambodia (7 sites with almost the same coordinates). This yields the database of 201 sites used in our main paper. The reason to exclude Laos and Cambodia is that we observe in Fig S1a the presence of two anomalously old (yellow) areas in these regions, the existence of which is supported by only 2 sites for the region in Laos, and by 7 very close sites for the region of Cambodia. Moreover, none of these 9 dates were obtained by direct AMS dating on rice itself. Associated radiocarbon dates from the site considered were used to date 2 sites. The other 7 dates were obtained by cultural association. Thus it seems perfectly possible that the yellow areas in Laos and Cambodia in Fig S1a will disappear with improved databases in the future. For this reason, we have preferred to exclude these sites from the analyses in the main paper. But we would like to stress that, if we included those 9 sites, the conclusions would be the same as those in the main paper. We can see this as follows. In the main paper, linear regression of dates versus distances from Chengtoushan yields (after excluding also 16 sites near Hemudu) an spread rate of 0.72-0.92 km/yr (185 sites, 95% CL). If we included the 9 sites in Laos amb Cambodia (S6 Table), this range would change into 0.76-0.98 km/yr (194 sites, 95% CL). Both ranges are rather similar. The correlation coefficient including Laos and Cambodia is $r = 0.74$ (194 sites), somewhat lower than the value $r = 0.78$ obtained the main paper (185 sites), and this suggests that the rate reported in the main paper is statistically sounder. How would the analysis in the main paper change if we used the rate including the 9 sites in Laos and Cambodia (0.76-0.98 km/yr) instead of the rate excluding them (0.72-0.92 km/yr)? The horizontal rectangle (observed range) in Fig 4 in the main paper would have the upper bound 0.98 km/yr (instead of 0.92 km/yr), so that the upper side of this rectangle would cross the dashed line (minimum predicted speed) at about $C = 0.6$ (instead of $C = 0.5$). Thus, in Fig 5 in the main paper, the maximum cultural effect would be 28% (instead of 24%). Both values are very similar, and well below 50%, so the main conclusion of our main paper (that the spread was mainly demic) would remain exactly the same. Therefore, including the 9 sites in Laos and Cambodia would not change the conclusions of our main paper.

S1d. Two methods to estimate spread rates: linear regressions versus geostatistical techniques

Besides using linear regressions to estimate an average spread rate at the continental level (main paper and previous subsections here), we also attempted to estimate local spread rates by applying the same geostatistical techniques as in previous work on the spread of the Neolithic in Europe (Fort 2015). This procedure is based on the following steps: (i) fitting a surface of arrival time on the map by interpolation and smoothing; (ii) estimating the slope at each point of a grid; and (iii) computing its inverse, which gives a map of local spread rates in km/yr (for a more detailed description of this method, see Fort (2015), especially its Supp. Info. Sec S1). In this subsection, we apply this approach to discuss two different issues.

A. Comparison between the databases of 201 and 331 sites

Applying the 3 steps summarized in the previous paragraph, we have obtained Fig S2a (201 sites). In this map we observe that almost all areas have a speed in the range 0.30-1.36 km/yr, which includes all of the range 0.65-0.83 obtained by linear regression (Fig 2 in the main paper, also with the same 201 sites). Moreover, the average from the speed map (Fig S2a) is 0.78 km/yr, very close to the linear regression speed with the same 201 sites, namely 0.74 km/yr (Fig 2 in the main paper). In contrast, in the speed map for all 331 sites (i.e., including India, etc.), shown in Fig S2b, almost all areas have speeds in the range 0-1.36 km/yr and this is not consistent with the corresponding linear regression speed, namely 1.51-2.11 km/yr (Fig. S1b). Those two ranges do not overlap. As mentioned in the last paragraph in Sec. S1a, this is a third argument (in addition to the independent domestication implied by genomic studies, and the correlation coefficients) to exclude the 130 sites in India, etc. from the regressions (as done in the main paper).
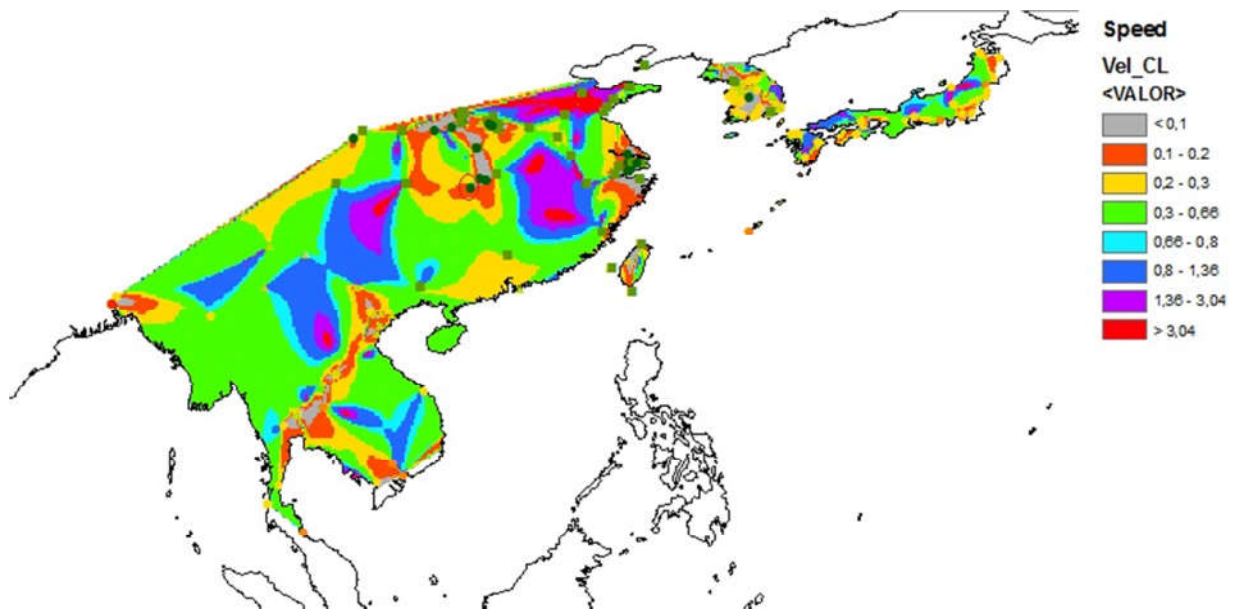
Fig S2a. Local front speeds obtained from the dates of the 201 sites in Fig 2 in the main paper. The large circle is the site of Chengtoushan. Map prepared with ESRI ArcGIS version 10 (http://www.esri.com/) using ArcMap and the ArcGIS Spatial Analyst extension.
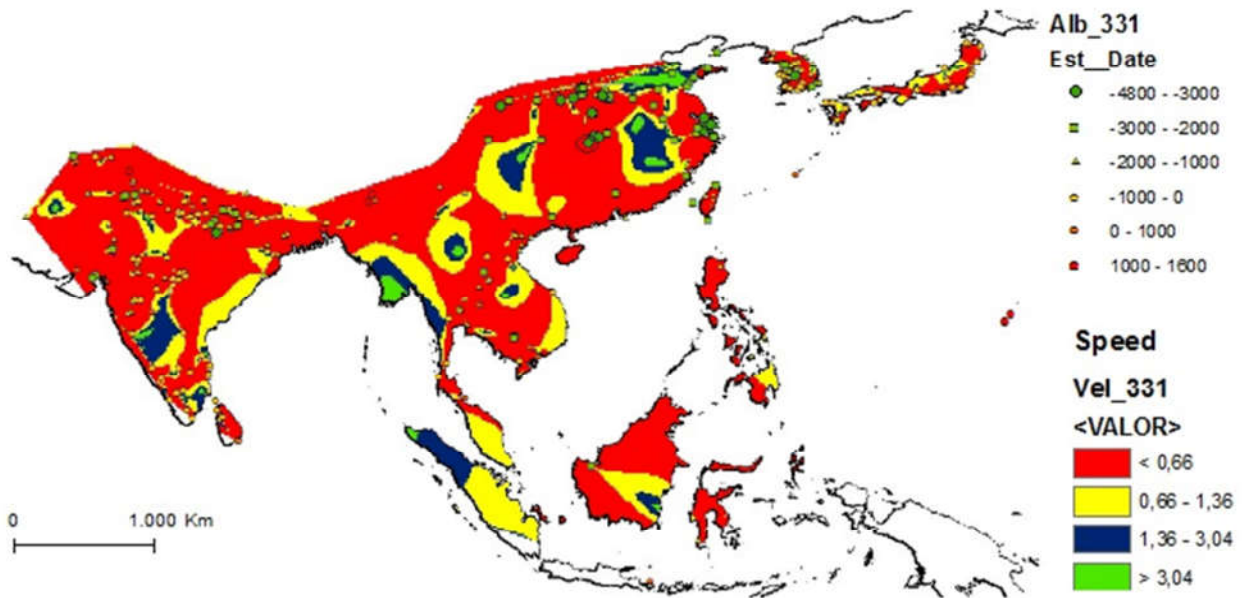

Fig S2b. Local front speeds obtained from the dates of the 331 sites included in Figs S1a-b. The large circle is the site of Chengtoushan. Map prepared with ESRI ArcGIS version 10 (http://www.esri.com/) using ArcMap and the ArcGIS Spatial Analyst extension.

B. Comparison with Europe

In our database there are few sites per unit area. This is seen in, e.g., Fig S2 (201 sites), and much more clearly in the map of isochrones in Fig 1 in the main paper (also for 201 sites). Indeed, in Fig 1 in the main paper we see that there are, e.g., only about 10 sites in the 1,000-yr interval between 1,000BC and 1AD, whereas for Europe (Fig S3a) there are more than 100 sites for each period of 1,000 yr (note that each color in Fig S3a corresponds to a 250 yr interval, whereas in Figs 1 and S1a it corresponds to 1,000 yr). Obviously, with so few sites in Asia (Fig S2) the local speeds are clearly more uncertain than in Europe (Fig S3b). This suggests that local speed maps for Asia (Figs S2-S3) could perhaps change appreciably in the future, when more sites are discovered and dated. Indeed, in Figs S2-S3 individual sites have a clear influence on the shape of colored areas, and thus on the speed values, whereas in Europe they do not (see Fig S3b). This is why for this case study (main paper and Secs S1a-c above) we have preferred to estimate a global spread rate using linear regression (and not local speeds using a speed map such as Fig S2). The same procedure (linear regression) was used to estimate the spread rate of herding across southern Africa, because in that case study there are again rather few sites per unit area available at present (Jerardino et al., 2014).
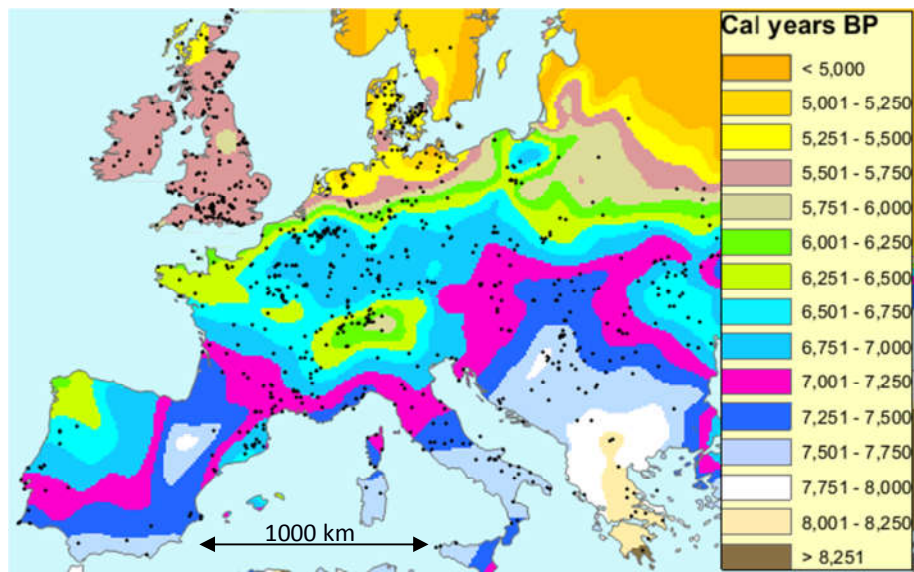
11

Fig S3a. Isochrone map for the spread of the Neolithic in Europe, obtained by interpolation of 918 sites. This figure is the same as Fig S3c in Fort (2015), and is included to compare to Fig 1 in our main paper and Fig S1a. All of these 3 maps have been smoothed 20 times, see the Supp. Info. Sec. S1 in Fort (2015) for further details on the methods used. Map prepared with ESRI ArcGIS version 10 (http://www.esri.com/) using ArcMap and the ArcGIS Spatial Analyst extension.

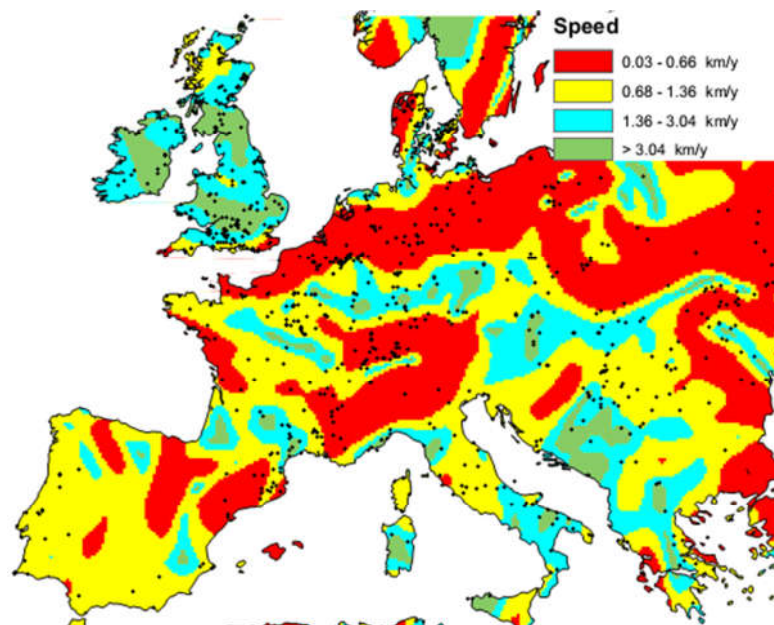**Fig S3b. Local speed magnitude map for the spread of the Neolithic in Europe.** This map is the same as Fig 3 in Fort (2015), and is included to compare to Figs S2-S3 here. It was obtained from Fig S3a above, see the Supp. Info. Sec. S1 in Fort (2015) for further details on the methods used. Map prepared with ESRI ArcGIS version 10 (http://www.esri.com/) using ArcMap and the ArcGIS Spatial Analyst extension.

12

S1e. Analysis excluding Hemudu and nearby sites (N=185 sites)

In the main paper (Sec. 3.1.2) we have explained that the very old dates in the area of the lower Yangtze (including Hemudu, the oldest site in the database) suggest that this was a region of independent domestication. For this reason, in our final analysis we have removed Hemudu and 15 sites nearby (i.e., the 16 blue squares in Fig 2 in the main paper). Figure S4 shows the corresponding linear fit (N=185 sites). Thus, Fig S4 is the same as Fig 2 in the main paper (201 sites), but removing the blue squares and performing the linear fit with the remaining 185 sites. Figure S4 implies a spread rate of 0.72-0.92 km/yr, which is referred to as the *'observed spread rate'* in the main paper.
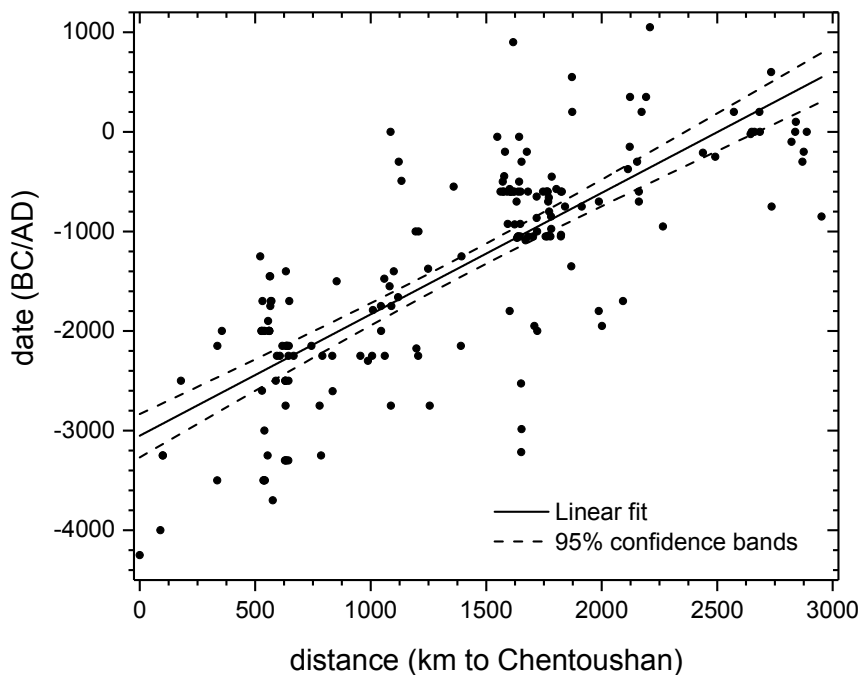
Fig S4. Linear regression fit to 185 sites. This is the same as Fig 2 in the main paper, but excluding Hemudu and 15 nearby sites (blue squares in Fig 2). Distances are computed from Chengtoushan (China, dated 4250 BC) using great circles. The front speed (inverse of the slope) is 0.72-0.92 km/yr (95% CL interval) and the correlation coefficient is $r = 0.78$, slightly larger than for Fig 2 in the main paper (201 sites, $r = 0.76$).

13

S1f. Effect of sites dated by cultural association

As mentioned at the beginning of Sec. S1, 57% of the sites considered were dated by inference from the depositional context and deposits (i.e., using cultural affiliation). In this subsection we check that excluding those sites does not change our main conclusion. Thus, from the list of 185 sites used to obtain the observed spread range in the main paper (Sec. S1e above and Sec. 3.1.2 in the main paper), we excluded all sites dated by cultural association, as well as the 4 sites for which the dating evidence was not reported in the database by Silva et al. (2015). This yielded 23 direct AMS dates on rice itself, 50 associated C14 dates from the site/phase, and 1 date obtained by optically stimulated luminescence (these 74 sites are listed in S3 Table, column W). Figure S5 shows the linear regression obtained using these 74 sites. The speed range (inverse of the slope) is 0.70-1.18 km/yr and $r = 0.67$. The speed error is larger and the statistical confidence (value of r) is lower than using all 185 sites (0.72-0.92 km/yr, $r = 0.78$, see Sec. 3.1.2 in the main paper). But the important point is that both speed ranges are rather similar and, for this reason, if we use the speed range obtained from Fig S5, i.e. 0.70-1.18 km/yr, in Figs 4-5 in the main paper, we easily see that the maximum cultural effect is about 38%. This is less than 50%, thus the main conclusion of our work (i.e., that demic diffusion was more important than cultural diffusion) remains the same. Therefore, this additional analysis strengthens the results in the main paper.
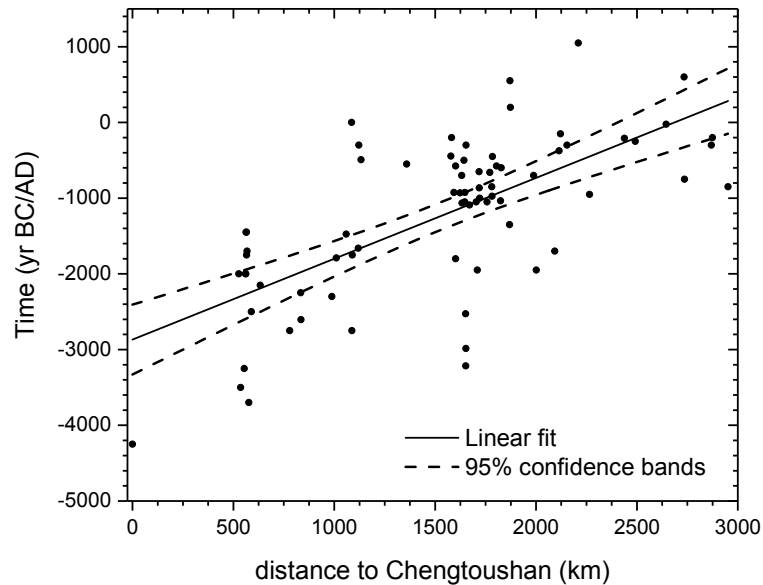
Fig S5. Linear regression excluding dates obtained by cultural affiliation (N=74 sites).

S1g. Cost-distance approach

In the main paper (Sec. 2.3) we have used the concept of great-circle distance, defined as the shortest distance between two points on the Earth surface (considered a sphere). This is a valid approximation in many cases (Pihasi et al. 2005), but here we would like to discuss an alternative approach. The presence of landmasses such as oceans and mountains, and possibly other factors (vegetation, roughness, etc.), can make it impossible (or cause difficulties) for the wave of advance to spread in some regions. Thus in general, the direction of propagation of the wave of advance may not follow the shortest path between two points. Many authors have considered the concept of cost distance to take such effects into account.

It is important to stress that many different techniques are grouped under the name 'cost-distance modelling'. Some such techniques are based on computing a cost distance, for each possible path, that is a

measure of the energy expenditure by a human individual to follow the path considered (Field et al. 2007). Other authors estimate cost distances on the assumption of several hypothetical spread rates in different places and directions (Silva and Steele, 2014). Still others estimate cost distances as travel times (Park and Rilett, 1999). All of these approaches are useful for some objectives, e.g.: (i) to find the least-cost path (between the dispersal origin and an archaeological site); (ii) to compare different assumptions on the values of local spread rates; (iii) to compare alternative dependencies of costs on elevation, slope, vegetation, etc. However, these techniques are not useful for our purposes. The reason is that their cost distances are not geographical distances, i.e., they are not measured in km but in cost units (e.g., Joules or calories, if cost distances measure energy expenditure). Therefore, in such analyses, a spread rate (obtained by regression of dates versus cost distances of archaeological sites) will not be measured in kilometers per year (km/yr) but in cost units per year, e.g., Joules per year (J/yr). But the spread rate of the demic-cultural model, which is given by Eq. (5) in our main paper, has units km/yr. This implies that it is impossible to show the observed spread range in the plot of the spread rate predicted by the demic-cultural model, simply because the former is measured in cost units/yr and the latter in km/yr. Recall that this has been done in the main paper, where we have used great-circle distances so their units are km (not cost units), thus the observed range has units km/yr (not cost units/yr, e.g., J/yr) and can thus be shown as the horizontal hatched rectangle (0.72-0.92 km/yr) in Fig 4 in the main paper. Thus, clearly such cost distances are not useful for our purposes, namely to compare observed and predicted rates (Fig 4 in the main paper) and determine the percentages of demic and cultural diffusion (Fig 5 in the main paper, which follows directly from Fig 4).

Some kinds of cost distances have units km, and are therefore useful for our purposes. To this class belong the so-called shortest-path distances, which we introduced previously (Pinhasi et al., 2005) and simply take into account seas and high mountains as obstacles (so each distance is computed as the sum of a few great-circle distances, which avoid the obstacles). At present, it is easy to use a still more precise

approach by making use of the free internet application google maps (https://www.google.com/maps). This gives the shortest route and distance on foot between two points, following present roads. Obviously, present roads have some modifications (bridges, tunnels, highways, etc.) relative to the foot routes used thousands of years ago. This will have the effect to shorten some distances, but the differences will be presumably small because most of the length of many routes is usually not affected by such important modern modifications. We admit that more precise measures of cost distance can be envisaged, but we would like to stress that present roads have the following advantages over other methods used so far, such as great-circle and shortest-path distances: (i) present roads do take into account the topography, and any other factor that may influence on human motion; (ii) they yield distances in km, as necessary for our purposes (see the previous paragraph); (iii) roads are reasonable measures of the shortest and most feasible routes, given the local topography and environment (Whitley and Dorn, 1993), and are thus very stable over long times; (iv) google maps distinguishes between distances by foot, which are used by us, from distances by car, which are more affected by modern modifications; (v) in principle, using real roads to compute cost distances can be more precise than the shortest-path approximation summarized above (Pinhasi et al., 2005), as the latter only adds a few great-circle distances between two points, whereas google maps follow the tortuous paths of real roads (however, this point is analyzed further the last paragraph in this section).

We considered all 185 sites in our database (Sec. S1e above and Sec. 3.1.2 in the main paper) and used google maps to compute the distance from each site by foot (in km) to Chengtoushan (i.e., the origin of dispersal considered in Sec. S1e and the main paper). For sites affected by sea travel, cost distances have been estimated by following the procedure explained in the caption to Fig S6. The resulting cost distances are included in S3 Table, column AC.
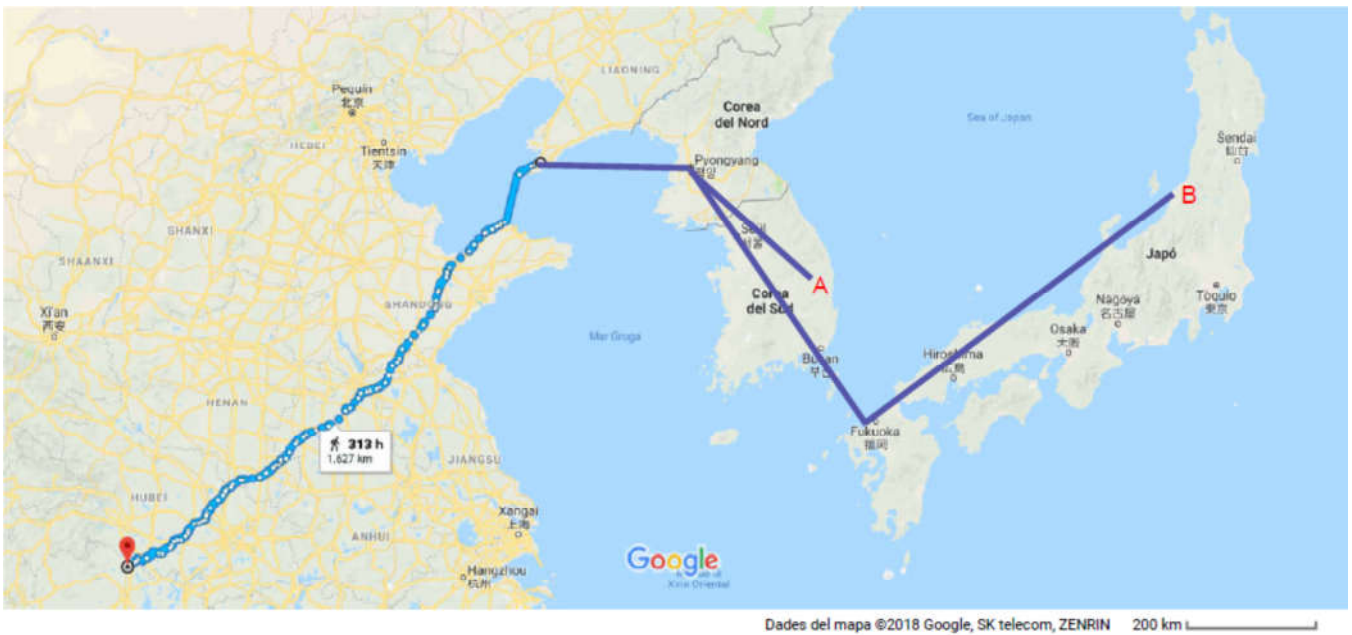
Fig S6. Calculation of cost distances using the free internet application google maps (https://www.google.com/maps). The point at the lower left is the site of Chengtoushan. The path at the left is the cost-distance route by foot estimated by google maps from Chengtoushan to the site of Dazuizi (other sites on the mainland are similarly computed using google maps, but here we display the case of Dazuizi because it is useful to discuss the computation of cost distances to sites in Korea and Japan). Note that google maps allows for sea travels of less than about 200 km, which is reasonable because such distances were easily travelled by sea in prehistory (for example, the island of Cyprus is more than 100 km from mainland Europe and was reach by early Neolithic farmers about 10,000 years ago). Google maps does not find distances crossing North Korea (because it is outside its range), so in the cases for which this was necessary (e.g., for sites A and B on the map) we estimated the total cost-distance from Chengtoushan as follows. If the site is in Korea (e.g., site A on the map), we add up the cost distance (i.e., the google maps distance) from Chengtoushan to Dazuizi (left path) plus the great-circle distance from Dazuizi to the site of Namgyong (horizontal path on the map) plus the great-circle distance from Namgyong (Pyongyang on the map) to the site considered (e.g., site A on the map). Similarly, if the site is in Japan (e.g., site B in the map) we added up the cost distance (i.e., the google maps distance) from Chengtoushan to Dazuizi (left path) plus the great-circle distance from Dazuizi to Namgyong (horizontal path) plus the great-circle distance from Namgyong (Pyongyang) to Nabatake (the western site in Japan, near Fukuoka on the map) plus the great-circle distance from Nabatake to the site considered (e.g., site B on the map). Admittedly, more refined approaches are possible, but we think that they would not lead to substantially different distances, so the conclusions would be the same.

Figure S7 shows the linear regression using cost distances, obtained with the application google maps (see Fig S6 and the text above it). The linear fit in Fig S7 yields $r = 0.82$ (which is higher than the value $r = 0.78$ in Fig S4 and the main paper) and a spread rate of 1.02-1.28 km/yr with 95% CL (which is faster than 0.72-0.92 km/yr in Fig 4 and the main paper, because it follows from Fig S6 that cost distances are always longer than great-circle distances). Using the range obtained from cost distances (1.02-1.28

km/yr) in Figs 4-5 in the main paper, it is easy to see that the maximum cultural effect is about 42%. This is less than 50%, thus the conclusion in the main paper (i.e., that demic diffusion was more important than cultural diffusion) does not change if using cost distances (as done in this section) instead of great-circle distances (as done in the main paper). We stress that alternative cost distance estimations, which do not yield any speed in km/yr, cannot be applied to estimate the percentages of demic and cultural diffusion (see the text above Fig S6). Moreover, we would like to explain that, in our opinion, cost distances are not necessarily more accurate than great-circle distances, and that their respective spread rates should be more properly regarded as upper and lower bounds, respectively (similarly, their respective percentages of cultural diffusion should de regarded as lower and upper bounds, respectively). The reason is the following. First, note that cost distances will be necessarily larger than great-circle distances, because the latter are by definition the shortest possible distances between the points considered. Next, note that if an obstacle is large compared to the displacements per generation of the individuals, then it is reasonable to expect a deflection in the direction of spread of the wave of advance and, therefore, cost distances (computed using google maps) will surely be more accurate than great-circles distances. But imagine a situation in which there are many small obstacles (compared to the displacements per generation of the individuals). In this instance, the obstacles will not have any effect of the spread rate of the wave of advance, because the latter depends only on the initial and final points of the individual displacements, i.e. on the distances in the dispersal kernel, and not on the path joining both points (see, e.g., Fort, 2012). Thus for small obstacles, cost distances will be surely less accurate than great-circle distances. This is why, in our opinion, the maximum cultural effect estimated above using cost distances (42%) and that estimated in the main paper using great circles (24%) should be regarded as an upper and lower bound, respectively. This seems a conceptually important point to us (and it will be applied in Sec. S2e), although it is true that (as shown above) for our case study any of both approaches (cost distances and great circles) yields the same conclusion, namely that cultural diffusion was less important than demic diffusion.
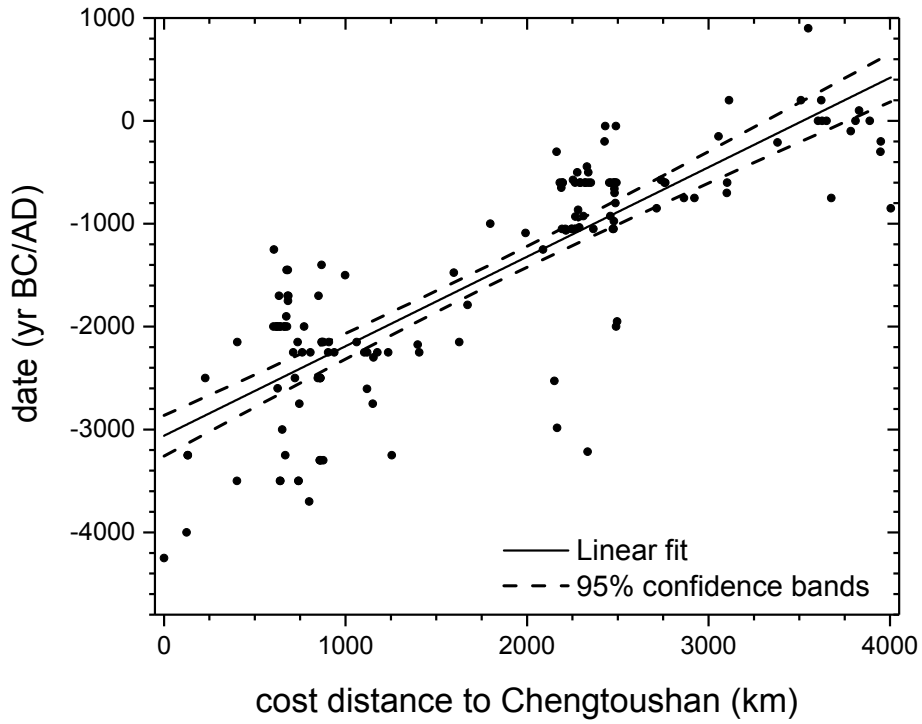
Fig S7. Linear regression using cost distances. The linear fit implies a spread rate of 1.02-1.28 km/yr (95% CL) and $r = 0.82$. The circles correspond to N=150 sites (we ignored some sites for which google maps could not compute the distance to Chengtoushan by foot).

S2. Dispersal kernels of rice cultivators

We have performed a detailed search of the dispersal kernels of rice cultivators that have been reported in the literature. In this section we discuss their implications.

S2a. The kernel reported by Mehrai (1984)

This is the dispersal kernel used in the main paper (Figs 4 and 5). It is extremely interesting that Mehrai (1984) is apparently the only author who has reported distances between birthplaces of parents and children in rural Asian populations of rice cultivators. In contrast, all other kernels for rice cultivators in the literature (see below) apparently give only marriage distances (i.e., distances between the places of birth or residence of spouses). The sample of one of the 3 populations studied by Mehrai (1984) is very small (see his table 5.27), so Mehrai (1984) suggests (p. 199) that it is not representative. Another of these 3 populations does not refer to rice cultivators but to pastoralists (table 5.28 in Mehrai, 1984). Thus we use the data on the individuals of the remaining population, who are rice cultivators in Bakhsh Markazi (table 5.26 in Mehrai, 1984). Those data refer to grandparents-parents as well as to parents-children distances. For the parents-children data, improvements in roads increased the distances moved by the population (see p. 150 in Mehrai, 1984). Thus we consider grandparents-parents distances. The corresponding raw data (*N*=714) are not given by Mehrai (1984), but they are divided into four subgroups labeled FF-F (father's father-father), FM-F (father's mother-mother), MF-M (mother's father-mother) and MM-M (mother's mother-mother). For the distance classes below 40 km we pooled together the individuals from the four subgroups, whereas for the upper-distance class (>40 km) we estimated a characteristic distance for each subgroup from its reported mean, as done previously for populations that are not rice cultivators (Isern et al., 2008). The dispersal kernel is then given by the probabilities $\{p_j\} = \{0.803, 0.040, 0.022, 0.025, 0.063, 0.005, 0.009, 0.019, 0.014\}$ and their respective distances $\{r_j\} = \{0.5, 5.5, 15, 25, 35, 50.03, 57.20, 60.51, 97.65\}$ km. Using this kernel, we have obtained the results shown in Figs 4 and 5 in the main paper.

S2b. The kernel reported by Singh and Singh (2015)

As mentioned in the main paper, the most direct way to estimate the dispersal kernel is to consider distances between the birthplaces of parents and their children. It is unfortunate that such distances seem to have been reported for rice cultivators in Asia only by Mehrai (1984). Interestingly, however, it has been shown that parent-offspring distances can be estimated by means of marriage distances, defined as the distances between the places of birth or residence of spouses (Cavalli-Sforza 1962; Malhotra 1980). This is why most work on modelling the spread of the Neolithic across Europe has used (in addition to the very few parent-children data available) marriage distances (Ammerman and Cavalli-Sforza, 1984; Isern et al., 2008; Fort 2012, 2015). It has been indeed observed that both kinds of data yield the same conclusions concerning the relative importance of demic and cultural diffusion in the spread of the Neolithic in Europe (Fort, 2012). Analogously, we shall here use data on marriage distances and compare the implied relative importance of demic and cultural diffusion with that obtained from parent-child data in the main paper. This will allow us to see if dispersal data from different populations lead to the same conclusions or not.

Singh and Singh (2015) reported a histogram of marriage distances (N=930) for a present agricultural population in North-Eastern Bihar (India), where the staple food is rice (Choudhary, 2009; Sethy et al., 2016). For this population, the dispersal kernel is given by the following set of probabilities $\{p_j\} = \{0.018, 0.081, 0.105, 0.129, 0.140, 0.125, 0.107, 0.079, 0.068, 0.057, 0.036, 0.025, 0.030\}$ to move respectively distances $\{r_j\} = \{2.5, 7.5, 12.5, 17.5, 22.5, 27.5, 32.5, 37.5, 42.5, 47.5, 52.5, 57.5, 62.5\}$ km. Using this kernel, Fig S8a displays the predicted speed by the demic-cultural model, i.e. by Eq (5) in the main paper (similarly to Fig 4 in the main paper, which has been obtained using the kernel by Mehrai, 1984). The range of observed speeds (0.72 – 0.92 km/yr) obtained from the archaeological data (Fig 2 in the main paper) is shown in Fig S8a as a horizontal hatched rectangle. The region of consistency between observed and predicted speeds (black region in Fig S8a) indicates that the value of the cultural transmission

parameter is C is below 0.3. Thus we can safely state that $C < 0.5$, as concluded from Fig 4 in the main paper. On the other hand, the kernel reported by Singh and Singh (2015) implies that cultural effect on the Neolithic wave of advance was lower than 13% (Fig S8b). Similarly, using the kernel by Mehrai (1984), Fig 5 in the main paper implies that it was lower than 24%. Thus, according to both kernels, the cultural effect was clearly below 50%. Therefore, both this dispersal kernel and that used in the main paper lead to the same conclusion, namely that demic diffusion explains more than 50% of the spread rate and that, in this sense, the spread of rice across southeastern Asia was mainly demic.
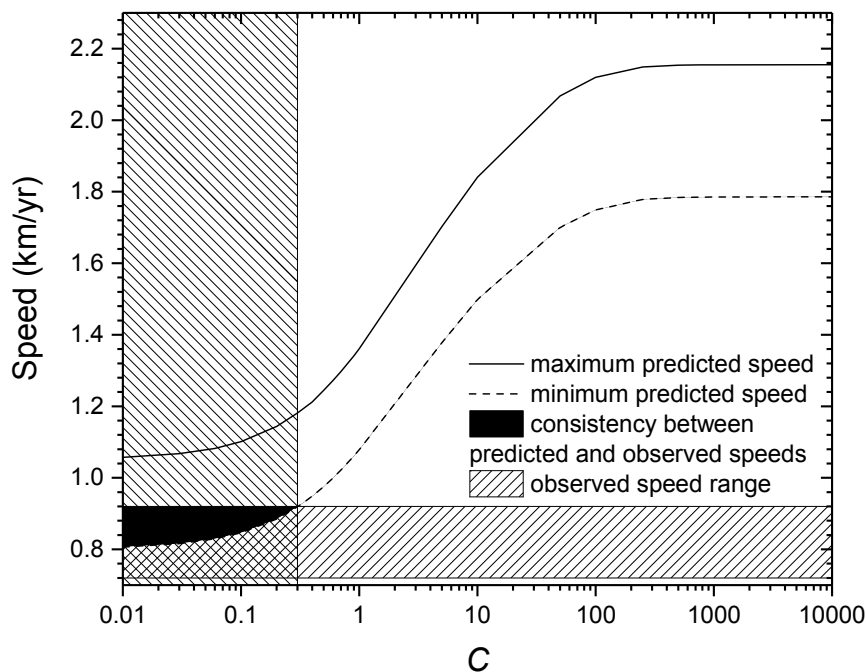
Fig S8a. Neolithic front speeds using the dispersal kernel reported by Singh and Singh (2015). The full curve is the maximum speed predicted by the demic-cultrual model, Eq (5) in the main paper, and the dashed curve is the minimum speed predicted by the same model. The horizontal rectangle is the observed speed range, and the vertical rectangle is the implied range of the cultural transmission intensity C. The purely demic model corresponds to C=0 (i.e., no incorporation of hunter-gatherers into the pioneering populations of rice cultivators). This figure is equivalent to Fig 4 in the main paper, but using the dispersion kernel by Singh and Singh (2015) instead of that by Mehrai (1984).

23

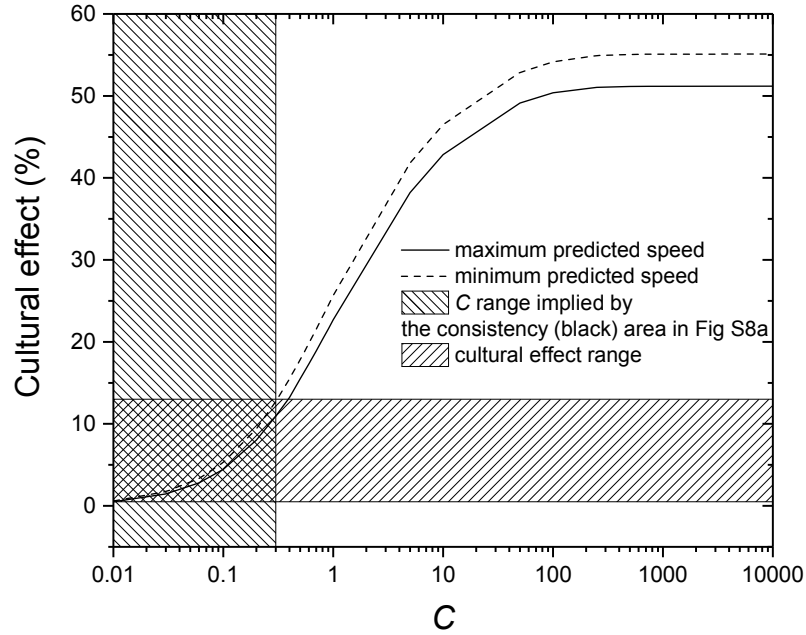**Fig S8b. Effect of cultural transmission on the front speed using the dispersal kernel reported by Singh and Singh (2015).** This is defined as the speed with cultural transmission (i.e., that plotted in Fig S8a for the value of C considered) minus the speed without cultural transmission (i.e., that plotted in Fig S8a for $C=0$), divided by the former and multiplied by 100. The full and dashed curves have been obtained using the full and dashed curves in Fig S8a respectively. The vertical rectangle is the range of cultural transmission intensity C implied from Fig S8a. The purely demic model (cultural effect of 0%) corresponds to $C=0$ (i.e., no conversion of hunter-gatherers into rice cultivators). This figure is equivalent to Fig 5 in the main paper, but using the dispersion kernel reported by Singh and Singh (2015) instead of that by Mehrai (1984).

S2c. The kernel reported by Shukla (2015)

Shukla (2015) reports marriage distances (N=345) of a population in Varanasi (Uttar Pradesh, India) whose staple food is rice (Anden and Barker 1974; Singh et al. 2015). For this population, the dispersal kernel is given by the probabilities $\{p_j\} =$ {0.058, 0.122, 0.191, 0.256, 0.168, 0.101, 0.069, 0.023, 0.012} and their respective distances $\{r_j\} =$ {4, 12, 20, 28, 36, 44, 52, 60, 68} km. This kernel is less detailed than that reported by Singh and Singh (2015), because the distances are grouped in a lower number of intervals. But since they have been obtained from different populations, it is interesting to compare the results. Using the kernel from Shukla (2015), the region of consistency between observed and predicted speeds (black

region in Fig S9a) indicates that the value of the cultural transmission parameter C is below 0.3. Thus we can safely conclude that $C < 0.5$, which is the same conclusion as that reached in the main paper and Sec. S2b (using other observed kernels for rice cultivators). The kernel reported by Shukla (2015) implies that cultural effect on the Neolithic wave of advance was lower than 12% (Fig S9b), which is again well below 50%. Therefore, this kernel leads to the same conclusion as those in the main paper and Sec. S2b, namely that demic diffusion explains most of the spread rate (i.e., that the spread of rice across southeastern Asia was mainly demic).
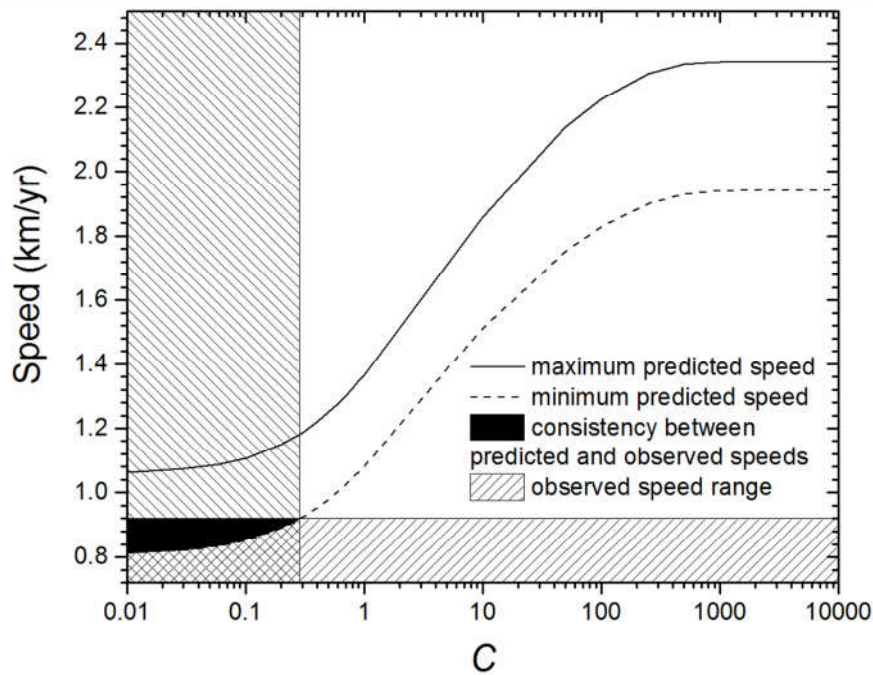
Fig S9a. Neolithic front speeds using the dispersal kernel reported by Shukla (2015). The full curve is the maximum speed predicted by the demic-cultrual model, Eq (5) in the main paper, and the dashed curve is the minimum speed predicted by the same model. The horizontal rectangle is the observed speed range, and the vertical rectangle is the implied range of the cultural transmission intensity C. The purely demic model corresponds to C=0 (i.e., no incorporation of hunter-gatherers into the pioneering populations of rice cultivators). This figure is equivalent to Fig 4 in the main paper, but using the dispersion kernel by Shukla (2015) instead of that by Mehrai (1984).

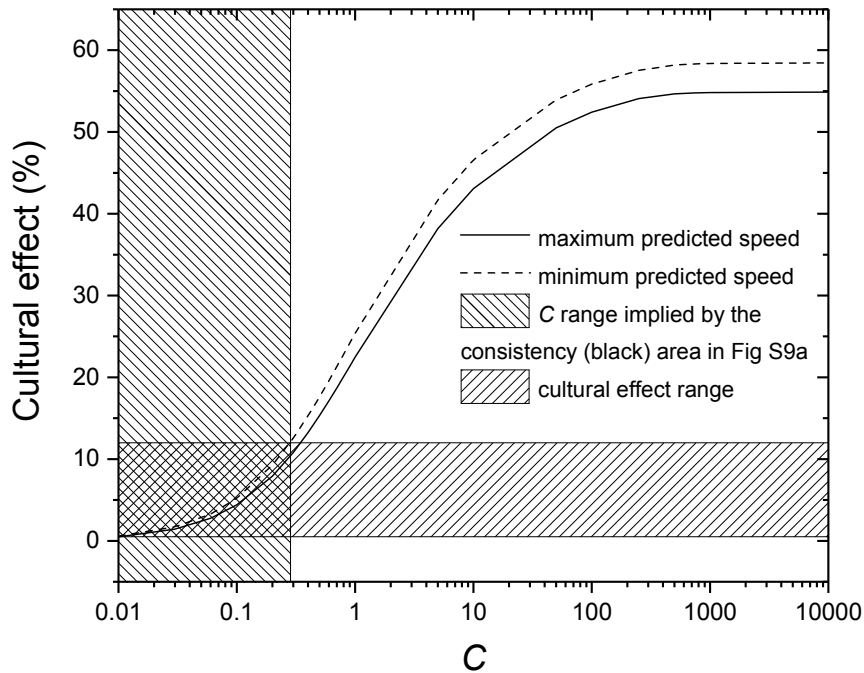**Fig S9b. Effect of cultural transmission on the front speed using the dispersal kernel reported by Shukla (2015)**. This is defined as the speed with cultural transmission (i.e., that plotted in Fig S9a for the value of C considered) minus the speed without cultural transmission (i.e., that plotted in Fig S9a for C=0), divided by the former and multiplied by 100. The full and dashed curves have been obtained using the full and dashed curves in Fig S9a respectively. The vertical rectangle is the range of cultural transmission intensity C implied from Fig S9a. The purely demic model (cultural effect of 0%) corresponds to C=0 (i.e., no conversion of hunter-gatherers into rice cultivators). This figure is equivalent to Fig 5 in the main paper, but using the dispersion kernel reported by Shukla (2015) instead of that by Mehrai (1984).

S2d. Other kernels of rice cultivators

The kernel by Deshorjit and Nabakumar (2010) gives probabilities for several intervals of marriage distances (N =481) recorded for the inhabitants of Phayeng, a village in Manipur (India), who are also rice cultivators (Ghosh and Ghosh 1997; Devi and Suresh Kumar 2012). Unfortunately, for the upper-distance class they report only a lower bound, namely >45 km. Therefore, this kernel is clearly less precise than those by Singh and Singh (Sec. S2b) and Shukla (Sec. S2c), who instead report the upper-distance classes with lower and upper limits (60-65 km and 64-72 km, respectively). This is one limitation of the kernel by Deshorjit and Nabakumar (2010). Another limitation is due to the fact that (as in Secs. S2b-c) we are dealing with marriage

distances which, in contrast to the parent-child distances (main paper and Sec. S2a), are a less direct measure of the mobility used in the mathematical model. Both limitations imply that there are several sources of uncertainty in this kernel. Hence we will not use it, in view of the fact that we have 3 more precise kernels (one used in the main paper, and two used in Secs. S2b and S2c above). The same happens for a marriage kernel reported by Yadav (2016), who again gives only a lower bound for the upper-distance class. There are a few additional marriage kernels for rice cultivators, but they cannot be used because, besides giving only a lower bound for the upper-distance class, apparently they do not report the mean or any other value from which we may estimate a characteristic distance for the upper-distance class (Yasuda 1975; Malhotra 1980; Gautman and Kshatriya 2011).

Basu (1973) gives raw data for marriage distances (*N*=140) for a population in the neighborhood of Giridih (Bihar, India), where the staple food of this population is also rice (Fuwa et al. 2005; Banik et al. 2014). Unfortunately, we cannot use these data either, because the maximum distances included are very short (28 km) in comparison with those reported by Singh and Singh (2015) (65 km) or by Shukla (2015) (72 km). One possible explanation to this is that the author explains that he could not find on the map the locations corresponding to 41 of the 181 distances, probably because the location of at least one of the spouses was outside the study area considered. This is likely the reason why the longest distances were not included.

Finally, Majumder and Malhotra (1979) present a database on distances between the birthplaces of spouses collected from 2260 married individuals of the Dhangar, a caste of people located in Maharashtra (India). Unfortunately, we cannot use those data either, because distances were measured along existing roads rather than as great-circle distances, and the latter corresponds to our estimation of the front speed from archaeological data (see Eq (2) in the main paper).

S2e. Sensitivity analysis of the dispersal kernel

The dispersal kernel reported by Mehrai (Fig 4 in the main paper) predicts spread rates that are a bit slower than those predicted by the kernels reported by Singh and Singh (Fig S8a) and by Shukla (Fig S9a). This is because of the shorter distances in the kernel by Mehrai (for which 80% of individuals move 0-1 km) compared to those by Singh and Singh and by Shukla (for which about 60% of individuals move 22.5±12.5 km and 28±12 km, respectively). These three kernels have been obtained from ethnographic data of rice farmers, and they yield spread rates consistent with the observed one (Figs 4, S8a and S9a). However, using these dispersal kernels implies assuming that they are similar to those of rice farmers that lived up to 6,200 years ago and in regions with different environments and separated up to 3,000 km. Perhaps in the future this assumption will become unnecessary, because hopefully it will be possible to measure dispersal kernels directly from archaeological (rather than ethnographic) data, by genetic identification of parent-child pairs buried in different locations. But unfortunately, up to now all prehistoric parent-child pairs that have been genetically identified were buried in the same site (Monroy et al. 2018). Therefore, at present we can only use ethnographic kernels. Then the question arises of which extreme, hypothetical kernels (beyond those attested ethnographically) would cause our model not to hold. We tackled this question by taking into account that, as explained at the end of Sec. S1g, great circles and cost distances give lower and upper bounds to the spread rate, respectively. Therefore, in order to determine the complete range of the dispersal kernel (for which our model holds), it is more reasonable to consider the complete range of the spread rate (implied by both great circles and cost distances), i.e. 0.72-1.28 km/yr. Obviously, the model can break down in two cases: (A) if it predicts spread rates that are faster than the observed one (0.72-1.28km/yr), and this can happen only for kernels with longer distances than the three ethnographic kernels used by us (Figs 4, S8a and S9a); (B) the model will also break down if it predicts spread rates that are slower than the observed one (0.72-1.28 km/yr), and this can happen only for kernels

with shorter distances than the three ethnographic kernels used by us (Figs 4, S8a and S9a). We discuss both cases separately.

A. Kernels with long distances

Firstly we look for hypothetical kernels with distances long enough so that the model predicts that all spread rates are faster than the observed one, i.e. above 1.28 km/yr. Note that in this case it is sufficient to consider the minimum speed case, i.e. a_N =0.023 yr^{-1}, T =35 yr and $C = 0$ (see the main paper, Sec. 3.2), because other values in the ranges justified in the main paper (0.023 yr^{-1} $\leq a_N \leq$ 0.033 yr^{-1}, 29 yr $\leq T \leq$ 35 and $0 \leq C \leq \infty$) would lead to still faster speeds.

As explained above, the ethnographic kernel with faster speeds (due to its longer distances) is that reported by Shukla (Figs S9a-b), and is given by the probabilities $\{p_j\}$ ={0.058, 0.122, 0.191, 0.256, 0.168, 0.101, 0.069, 0.023, 0.012} for the distances $\{r_j\}$ ={4, 12, 20, 28, 36, 44, 52, 60, 68} km (Sec. S2c). In a first approach, we modify this kernel by increasing all distances by the same amount (for the other two kernels, larger increases in distance would be necessary, and the conclusion would be the same). Using Eq. (5) in the main paper, it turns out that, for the spread rate to be faster than the observed one (1.28 km/yr), we have to increase all of these distances by at least 23 km. This implies that about 6% of the population moves less than 31 km (because for the kernel by Shukla $p_1 = 0.058 \approx 6\%$ and $r_1 = 4$ km corresponds to 0-8 km, to which we have added 23 km), and the other 94% moves longer distances than 31 km. However, we think that such a modified kernel is not acceptable because we have not found any human population for which *only 6%* of individuals move less than 31 km. In all cases known to us, *much more than 6%* of the population moves up to 31 km (or even a shorter distance). For the three kernels of rice cultivators introduced above, we have that 80% of the population moves less than 1 km (Mehrai), 10% moves less than 10 km (Singh and Singh) and 18% moves less than 12 km (Shukla). Similarly, for pre-industrial farmers that are not rice cultivators (Isern et al. 2008), 54% moves less than 4.8 km (Gilishi 15), 40%

29

moves less than 4.8 km (Gilishi 25), 19% moves less than 4.8 km (Shiri 15), 19% moves less than 20 km (Yanomano), and 42% moves less than 4.6 km (Issongos). And for a modern population (Isern et al. 2008) 77% moves less than 2.5 km (Parma). Therefore, as far as we know, there is no human population for which *only 6%* of individuals moves less than 31 km per generation. In other words, the increase in distances necessary for the model to break down is not defensible, and the reason is that too few children would have to stay with or near their parents (as compared to real human populations).

As an alternative approach, we also tried to increase only the longest distances. Thus, we again increased distances in the kernel by Shukla (Sec. S2c), but only the 10 % longest distances. Therefore, we increased by the same amount the distances of 52 km (which corresponds to the interval 48-56 km, and has probability 0.069=6.9%,), 60 km (56-64 km, probability 0.023=2.3%) and 68 km (>64 km, probability 0.012=1.2%). Then the increase necessary for the model not to hold (i.e., to yield all speeds above 1.28 km/yr) is 54 km. This would imply that about 10% of the population moves distances above 102 km. Again such a high intergenerational mobility is much higher than all data known for human populations. Indeed, it is always observed that *much less than 10%* of the population moves more than 102 km (or even a shorter distance). For the three kernels of rice cultivators introduced above, we have that only 1.4% of the population moves more than about 97.65 km (Mehrai), only 3% moves more than 60 km (Singh and Singh) and only 1.2% moves more than 64 km (Shukla). But the same happens also for pre-industrial farmers that are not rice cultivators (Isern et al. 2008), for which only 2% moves more than 100 km (Yanomano), and only 1% moves more than 60 km (Issongos) [we mention that the data for the populations Gilishi 15, Gilishi 25 and Shiri 15 are not detailed enough for this analysis, because the longest-distance interval is >48 km and has probabilities >25% (Isern et al. 2008)]. Also for a modern population (Isern et al. 2008), only 5% moves more than 72.5 km (Parma). Therefore, as far as we know, there is not any human population for which 10% or more of the individuals move distances above 102 km per generation. Thus,

using this second approach, we conclude again that our model does not break down unless we assume a dispersal kernel that is unreasonable for human populations.

B. Kernels with short distances

Here we look for hypothetical kernels with distances short enough so that the model predicts all spread rates slower than the observed one, i.e. below 0.72 km/yr. Note that in this case it is sufficient to consider the maximum speed case, i.e. $a_N = 0.033$ yr^{-1}, $T = 29$ yr and $C \to \infty$ (see the main paper, Sec. 3.2), because other values in the ranges justified in the main paper (0.023 yr$^{-1} \leq a_N \leq 0.033$ yr^{-1}, 29 yr $\leq T \leq 35$ and $0 \leq C \leq \infty$) would lead to still slower speeds.

As mentioned above, the ethnographic kernel with slower speeds (due to its shorter distances) is that reported by Mehrai (Figs 4-5 in the main paper), and has probabilities $\{p_j\} = \{0.803, 0.040, 0.022,$ $0.025, 0.063, 0.005, 0.009, 0.019, 0.014\}$ for distances $\{r_j\} = \{0.5, 5.5, 15, 25, 35, 50.03, 57.20, 60.51,$ $97.65\}$ km (Sec. S2a). In a first approach, we modify this kernel by decreasing all distances by the same amount (for the other two kernels, larger decreases in distance would be necessary, and the conclusion would be the same). Using this first approach and Eq. (5) in the main paper, it turns out that it is not possible for the spread rate to be slower than the observed one (0.79 km/yr) for all values of C. The reason is that for a decrease of 0.5 km some spread rates are still substantially faster than 0.72 km/yr, and we cannot decrease all distances by more than 0.5 km due to the fact that the shortest distance would be negative (because in the kernel by Mehari $r_1 = 0.5$ km).

In a second approach, we modify again the kernel reported by Mehrai (Sec. S2a), but now by dividing all distances by the same number. Note that in this way, we cannot get any negative distance. We find that for the model not to hold (i.e., to get all spread rates slower than 0.72 km/yr using Eq. (5) in the main paper), we need to divide all of these distances by 4.7. Is such a kernel defensible from what is known about the intergenerational mobility of humans? Its mean distance is only 1.5 km, which is extremely small

compared to the mean distance of the three kernels of rice cultivators introduced above, namely 7 km (Mehrai), 28 km (Singh and Singh) and 29 km (Shukla). Similarly, for pre-industrial farmers that are not rice cultivators (Isern et al. 2008), the mean distance is 21 km (Gilishi 15), 25 km (Gilishi 25), 41 km (Shiri 15), 36 km (Yanomano), and 12 km (Issongos). And for a modern population (Isern et al. 2008) the mean distance is 9 km (Parma). These values are all between 5 and 27 times larger than the value (1.5 km) that is need for the model to break down. Thus, it does not seem possible to defend the validity of such a kernel for human populations.

Obviously, many other hypothetical kernels can be considered, but these results strongly suggest that they will have similar problems. Therefore, it is reasonable to conclude that our model holds for reasonable dispersal kernels.

S3. Derivation of the cultural transmission model

The cultural transmission model used in the main paper was derived (Fort, 2012) by generalizing a previous model due to Cavalli-Sforza and Feldman (1981). In order to make the present work as self-contained as possible, in this section we summarize the derivation of the model and the interpretation of the corresponding parameters. Here we deal with cultural transmission, so we consider homogeneous systems for notational simplicity (the non-homogeneous extension in the main paper is straightforward). We could consider any cultural trait, but for our purposes it will be clearer to consider rice farming. At time t, let $N(t)$ stand for the number of rice farmers and $P(t)$ for the number of hunter-gatherers. As mentioned in the main paper (Sec. 2.4), after reproduction during a generation time T the number of rice farmers is $R_T[N(t)]$ and the number of hunter-gatherers is $R_T[P(t)]$, where $R_T[...]$ is the logistic function (its corresponding equation is given below Eqs. (4b) in the main paper). In order to find how many of these

hunter-gatherers become farmers, Cavalli-Sforza and Feldman (1981) considered that during her/his lifetime, each hunter-gatherer contacts (on average) n teachers, and assumed that a proportion $u = \frac{R_T[N(t)]}{R_T[N(t)]+R_T[P(t)]}$ of them is of type N (this model will be generalized in the next paragraph). Then, introducing the probability g that a P individual becomes a N individual due to contact with a single N individual, the probability that she/he will become N after n contacts is $1 - (1 - g)^{nu}$, which is approximately equal to ngu if g is small (Cavalli-Sforza and Feldman 1981). This implies that the number of hunter-gatherers becoming rice farmers per generation is $fuR_T[P(t)] = f\frac{R_T[N(t)]R_T[P(t)]}{R_T[N(t)]+R_T[P(t)]}$, where $f = ng$. Thus, in this model the numbers of rice farmers and hunter-gatherers in the next generation (i.e., at time $t + T$) are

$$N(t + T) = R_T[N(t)] + f\frac{R_T[N(t)]R_T[P(t)]}{R_T[N(t)] + R_T[P(t)]}$$

$$P(t + T) = R_T[P(t)] - f\frac{R_T[N(t)]R_T[P(t)]}{R_T[N(t)] + R_T[P(t)]}$$

If there are many more rice cultivators than hunter-gatherers ($R_T[N(t)] >> R_T[P(t)]$), the last equation becomes $P(t + T) = (1 - f)R_T[P(t)]$, which implies that $f < 1$ because population numbers cannot be negative. This model has the problem that in the opposite limit, $R_T[N(t)] << R_T[P(t)]$, the equation for the number of rice farmers becomes $N(t + T) = R_T[N(t)] + fR_T[N(t)]$, where $R_T[N(t)]$ is the 'initial' number of farmers (i.e., before cultural transmission), so the condition $f < 1$ implies that each rice farmer cannot convert more than a hunter-gatherer into a rice farmers during their lifetime. To avoid this limitation, it was assumed (Fort 2012) that, for learning purposes, each hunter-gatherer contacts only a fraction α of her/his N neighbors and a fraction β of her/his P neighbors. Then, the proportion of teachers of type N that she/he contacts is $\frac{\alpha R_T[N(t)]}{\alpha R_T[N(t)]+\beta R_T[P(t)]}$ and the two equations above are generalized into

$$N(t + T) = R_T[N(t)] + f\frac{R_T[N(t)]R_T[P(t)]}{R_T[N(t)] + \gamma R_T[P(t)]}$$

$$P(t + T) = R_T[P(t)] - f \frac{R_T[N(t)]R_T[P(t)]}{R_T[N(t)] + \gamma R_T[P(t)]}$$

where $\gamma = \frac{\beta}{\alpha}$ (the simpler model discussed above is the special case $\alpha = \beta$ or $\gamma = 1$).

This completes the derivation of the cultural transmission model (Eqs. (4b) in the main paper).

The derivation above leads immediately to the following interpretation of the parameters in the model (γ and f). If $\gamma > 1$, then hunter-gatherers are more likely to copy other hunter-gatherers than rice farmers and γ is a measure of this preference. Similarly, if $\gamma < 1$, then hunter-gatherers are more likely to copy rice farmers than other hunter-gatherers and γ measures this preference. If $\gamma = 1$, then hunter-gatherers are equally likely to copy the behavior of rice farmers or that of other hunter-gatherers, and this corresponds to the model by Cavalli-Sforza and Feldman (1981). Concerning parameter f, in the limit $R_T[N(t)] >> R_T[P(t)]$ the last equation becomes $P(t + T) = R_T[P(t)] - f R_T[P(t)]$. The first term on the right-hand side is the 'initial' number of hunter-gatherers (i.e., before cultural transmission), and the last term is the number of hunter-gatherers that adopt rice cultivation per generation (in this limit). Thus, we can interpret f as the fraction of hunter-gatherers converted into rice farming per generation in situations such that there are many more rice farmers than hunter-gatherers. In the opposite limit, $R_T[N(t)] << R_T[P(t)]$, there are very few rice farmers compared to hunter-gatherers (this corresponds, e.g., to the pioneering rice farmers in the arrival of a wave of advance of rice farmers) and we have $N(t + T) = R_T[N(t)] + \frac{f}{\gamma}R_T[N(t)]$. The first term in the initial number of rice farmers and the last one is its increase due to cultural transmission, so we can interpret $C = \frac{f}{\gamma}$ as the number of hunter-gatherers converted into rice cultivation per generation and pioneering farmer. $C = \frac{f}{\gamma}$ was called the intensity of cultural transmission (Fort 2012) and it can be arbitrarily large, which means that the number of hunter-gatherers converted by each rice farmers is not limited (in contrast to what happens in the simpler model discussed above, where $\gamma = 1$ and therefore $C = f < 1$). Note that in Figs 4-5 in the main paper (as well as in those in Sec. S2) the parameter that appears in the horizontal axis is C, because the spread rate of the waves of

34

advance of rice farmers depends only on the quotient $C = \frac{f}{\gamma}$ and not on parameters f and γ separately (see Eq. (5) in the main paper).

For completeness we mention that this model can be further generalized in two ways: (i) including frequency-dependent cultural transmission (conformism) (Fort 2012, Eqs. [S3]); and (ii) including interbreeding (vertical cultural transmission) (Fort, Pareta and Sørensen 2018). However, the results and conclusions of our work would be the same, so we do not repeat the details of those more complicated models here.

S4. Comparison to previous work

A previous study on the spread of rice considered models with several sources (Silva et al. 2015), although unfortunately speeds in km/yr were not reported, so we cannot perform a separate comparison for each speed. However, Silva et al. (2105) report a speed ratio of 0.78 (for the speed of the front from the lower Yangtze relative to that from the middle Yangtze), and our dual-source model (Sec. 3.1.3) yields a very close value, namely $\frac{0.60 \text{ km/yr}}{0.74 \text{ km/yr}} = 0.81$. The fact that both ratios are not exactly the same is more than acceptable, given that there are substantial differences between both approaches: (i) Silva et al. (2015) used least-cost distances (or more precisely, a kind of shortest-paths distances, see below), whereas here we have used great-circle distances in the main paper; (ii) Silva et al. (2015) estimated the speeds using fast-marching simulations, whereas here we have used only the archaeological dates; (iii) Silva et al. (2105) included sites with cultivated but not domesticated rice (so they included very old sites and could not obtain linear space-time dependences); and (iv) Silva et al. (2015) apparently do not clarify if the distances of all sites in their time-distance plots were computed relative to a single or several origins.

The most important point is that unless speeds are estimated in km/yr (and this was not done by Silva et al., 2015, 2018), it does not seem possible to determine the relative importance of demic and

cultural diffusion, which is our purpose. This is done in our main paper (Sec. 3.2), and was not attempted by Silva et al. (2015, 2018).

The 'best-fitting' model by Silva et al. (2015) clearly arrives too early to the great majority of sites (Fig 9 in Silva et al., 2015 or Fig 1 in Silva et al., 2018) because they arbitrarily 'give greatest weight to the earliest 10% of dates' (using quartile regression), but there is no justification to require this for 10% rather than, e.g., 5% or 20% of dates. In contrast, in the present paper we have not applied quantile regression but ordinary regression. The sum of the squared residuals is minimized in ordinary regression. In the simplest quantile regression (i.e., without weighting the residuals) the number of positive and negative residuals is the same, thus assuring that there are the same number of observations above and below the fitted function. In this sense, the fitted function gives the median or 50th percentile. More generally, quantile regression weights positive and negative residuals differently, so that the fitted function satisfies that there is a prescribed percentage of observations below it (Koenker and Hallock, 2001). For example, Silva et al. (2015) and Fuller et al. (2011) write that their quantile regression 'gives greatest weight to the earliest 10% of dates', without explaining the precise meaning of this statement. Perhaps it means that 10% of dates are older than the fitted function (evaluated at the same distance as that of each date). In any case, as mentioned above, using the 10% (rather than e.g. 5% or 20%) is an arbitrary, unjustified choice. Moreover, such an approach would clearly lead to erroneous results in our case study (the spread of domesticated rice), because it would rely heavily on many of the orange triangles in Fig S1b and blue squares in Fig 2 in the main paper (due to their old ages). But these are precisely the sites that should be excluded from the analysis (as justified in Sec. S1).

Finally we stress that some authors (see the citations in Silva et al., 2015) have used least-cost distances instead of great-circle distances (as in our main paper) or shortest-path distances (as in Pinhasi et al., 2005). As already explained in Sec. S1g, in many cases such cost distances assume that each kind of terrain implies some cost or modification in the distance, but the cost of each kind of terrain is not rigorously derived but chosen *ad hoc* (e.g., Silva and Steele, 2014). Moreover, usually they do not make it

possible to estimate the physical front speed in km/yr. This is a severe limitation indeed for our purposes, and is due to the fact that usually cost distances are not true physical (i.e., geographic) distances. The latter are the distances used in demic-cultural models (Fort, 2012, 2015) based on reaction-diffusion front propagation theory (Fort and Méndez 2002; Fort and Pujol 2008). An approach that makes possible to use cost distances for our purposes has been presented in Sec. S1g, and it leads to the same conclusions as the great-circle approach in our main paper.

Silva et al. (2015; 2018) do not strictly use least-cost distances but a kind of shortest-path distances, because their approach is to mask out deserts and cold regions, and extend the mainland 40 km offshore. However, unfortunately Silva et al. (2015, 2018) do not estimate any speed in km/yr, so comparison to demic-cultural models (Fort 2012) and estimation of the percentages of demic and cultural diffusion is not possible using their results.

References

Ammerman AJ, Cavalli-Sforza LL. Measuring the rate of spread of early farming in Europe. *Man*. 1971; 6: 674-688.

Ammerman AJ, Cavalli-Sforza LL. The Neolithic transition and the genetics of populations in Europe. Princeton: Princeton University Press; 1984.

Anden T, Barker R. Changes in rice farming in selected areas of Asia. Los Banos, Philippines: International Rice Research Institute; 1974.

Banik P, Edmonds C, Fuwa N. Sustainability implications of the evolution of rice farming amid rural poverty: the case of the Chhotanagpur Plateau in Eastern India. *J Sustain Dev*. 2014; 7: 282.

Basu A. A note of the distribution of marriage distance among the santals in the neighbourhood of Giridih, Bihar. J Biosoc Sci. 1973; 5: 367-376.

Cavalli-Sforza LL. The distribution of migration distances: models and applications to genetics. In: Sutter J, editor. Les Deplacements Humains. Monaco: Hachette; 1962, pp. 139-166.

Cavalli-Sforza LL, Feldman M. W. Cultural transmission and evolution: a quantitative approach. Princeton: Princeton University Press; 1981, chapter 3.

Choudhary KK. Occurrence of Chroococcaceae during rice cultivation in North Bihar, India. *Bangladesh J Plant Taxon*. 2009; 16: 57-63.

Deshorjit T, Nabakumar W. Endogamy and marital distance among the Phayeng of Manipur. *Studies on Tribes and Tribals*. 2010; 8: 61-65.

Devi P, Suresh Kumar P. Traditional, ethnic and fermented foods of different tribes of Manipur. *Indian Journal of Traditional Knowledge*. 2012; 11: 70-77.

Field JS, Petraglia MD, Lahr MM. The southern dispersal hypothesis and the South Asian archaeological record: examination of the dispersal routes through GIS analysis. *J. Anthropol. Archaeol.* 2017; 26: 88-108.

Fort J. Synthesis between demic and cultural diffusion in the Neolithic transition in Europe. *Proc Natl Acad Sci. USA*. 2012; 109: 18669-18673.

Fort J. Demic and cultural diffusion propagated the Neolithic transition across different regions of Europe. *J R Soc Interface.* 2015; 12: 20150166.

Fort F, Méndez V. Wavefronts in time-delayed reaction-diffusion systems. Theory and comparison to experiments. *Rep Progr Phys.* 2002; 65: 895-954.

Fort J, Pareta MM, Søresnsen L. Estimating the relative importance of demic and cultural diffusion in the spread of the Neolithic in Scandinavia. *J R S Interface.* 2018; 15: 20180597.

Fort J, Pujol T. Progress in front propagation research. *Rep Progr Phys*. 2008; 71: 086001.

Fort J, Pujol T, vander Linden M. Modelling the Neolithic transition in the Near East and Europe. *Amer Antiq.* 2012; 77: 203-220.

Fuller DQ, van Etten J, Manning K, Castillo C, Kingwell-Banham E, Weisskopf A, Qin L, Sato Y-I, Hijmans RJ. The contribution of rice agriculture and livestock pastoralism to prehistoric methane levels: an archaeological assessment*. The Holocene*. 2011; 21: 743-759.

Fuwa N, Edmonds C, Banik P. How inefficient are small-scale rice farmers in eastern India really? Examining the effects of microtopography on technical efficiency estimates (Working paper 79). Honolulu: East-West Center; 2005.

Gautman RK, Kshatriya GK. Polyandry: a case study of Kinnauras. *Ind. J. Phys. Anthrop. & Hum. Gen.* 2011; 30: 145-161.

Ghosh GK, Ghosh S. Women of Manipur. New Delhi: APH Publishing; 1997.

Glass C, Steele J, Wheatley D. Modelling spatial range expansion across a heterogeneous cost surface. In: Procs. CAA 97, Birmingham. *BAR Int. Series* 750: 67-72. ArcheoPress, Oxford; 1999.

Isern N, Fort J, Pérez-Losada J. Realistic dispersion kernels and cohabitation reaction-dispersion equations. *J Stat Mechs: Theory and Experiment.* 2008; 2008: P10012.

Jerardino A, Fort J, Isern N, Rondelli B. Cultural diffusion was the main driving mechanism of the Neolithic transition in southern Africa. *PLoS One*. 2014; 9: e113672.

Koenker R and Hallock KF. Quantile regression. *J Econ Pers.* 2001; 15: 143-156.

Majumder PP, Malhotra KC. Matrimonial distance, inbreeding coefficient and population size: Dhangar data. *Ann Hum Biol.* 1979; 6: 17-27.

Malhotra KC. Gene dispersion in man: the Indian case. *Curr Anthropol*. 1980; 21: 135-136.

Mehrai H. A demographic study of some populations in Shahrestan Nowshahr, Mazandaran, Iran, with reference to the genetic structure. PhD Thesis, Durham University. 1984. Available from: http://etheses.dur.ac.uk/7169/

Monroy JM, Jakobsson M, Günther T. Estimating genetic kin relationships in prehistoric populations. *PLoS One.* 2018; 13: e0195491.

Park D, Rilett LR. Forecasting freeway link travel times with a multilayer feedforward neural network. *Computer-aided civil and infrastructure engineering*. 1999; 14: 357-367.

Pinhasi R, Fort J, Ammerman AJ. Tracing the origin and spread of agriculture in Europe. *PLoS Biol*. 2005; 3: e410.

Sethy JP, Singh DR, Ankhila RH, Sharma V. Adoption and sustainability of eco-friendly SRI method of rice cultivation in Bihar. Research Journal of Agricultural Sciences. 2016; 7: 15-18.

Shukla KK. Models to describe the distribution of distance associated with marriage migration. Journal of Institute of Science and Technology. 2015; 20: 145-147.

Silva F, Stevens CJ, Weisskopf A, Castillo C, Qin L, Bevan A, Fuller D Q. Modelling the geographical origin of rice cultivation in Asia using the Rice Archaeological Database. *PLoS One.* 2015; 10: e0137024.

Silva F, Weisskopf A, Castillo C, Murphy C, Kingwell-Banham E, Qin L, Fuller D Q. A tale of two rice varieties: modelling the prehistoric dispersals of japonica and proto-indica rices. *The Holocene*. 2018; 28: 1745-1758.

Singh NK, Singh BP. Study of distance associated with marriage migration. *International Journal of Mathematics and Computer Applications Research.* 2015; 5: 111-116.

Singh AK, Sahu RK, Singh AK. Morpho-taxonomic distinction between *Oryza rufipogon* and *Oryza sativa* (Poaceae) at seedling stage. *Annals of Plant Sciences.* 2015; 4: 1068-1071.

Wang W, et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature.* 2018; 557: 43-49.

Whitley D S, Dorn R I. New perspectives of the Clois vs. Pre-Clovis controversy. *Amer. Antiq.* 1993; 58, 626-647.

Yadav RK, Kumar U. Study of marriage migration and distance. *Res J Math & Stat Sci.* 2016; 4: 11-17, table 3.

Yasuda N. The distribution of distance between birthplaces of mates. *Human Biology*. 1975; 47: 81-100.