J. Fort & J. Pérez-Losada, *Interbreeding between farmers and hunter-gatherers along the inland and Mediterranean routes of Neolithic spread in Europe*

Supplementary Information

INDEX

S1. Comparison to previous results

S1-A. Review of previous simulations of genetic clines

In this subsection we provide a list of previous simulations of genetic clines due to the spread of the Neolithic under the combined effects of demic and cultural diffusion. Here we aim to compare the assumptions of those models to ours (Sec. S2 below), so only the main issues relevant for this purpose are summarized.

Recall that the wave-of-advance model of Neolithic spread was introduced in 1971, together with the suggestion that the combination of demic and cultural diffusion could have led to genetic clines [1].

(1) In year 1973 Sgaramella-Zonta and Cavalli-Sforza [2] reported the first simulation of such clines on a grid. They assumed that initially there are farmers on the central cell and hunter-gatherers elsewhere. For each cell and generation, some individuals migrate to neighboring cells, each population reproduces constrained by its carrying capacity (which is higher for farmers), and some hunter-gatherers convert into farmers. They described cultural transmission using a Lotka-Volterra interaction, i.e., they assumed that each generation $\gamma P_F P_{HG}$ hunter-gatherers per unit area are converted into farmers (with P_F and P_{HG} the numbers of farmers and hunter-gatherers in a cell). Rounding the results of calculations to integers is not described in this work, so they used real rather than integer numbers, which avoids problems due to low population sizes (see Sec. S2-B below). This approximation (i.e., using real rather than integer numbers) has been also applied in recent simulations [3].

(2) In year 1986 Rendine, Piazza and Cavalli-Sforza [4] performed a spatial simulation that is similar to (1) in two ways. First, the Lotka-Volterra interaction (i.e., $\gamma P_F P_{HG}$) was again applied. Second, they apparently did not discuss any rounding of real numbers to integer ones. Thus, they did not replace the population size resulting from each calculation (migration, reproduction and cultural transmission) by an integer. In other words, real numbers were again used (rather than integer ones), avoiding the problems that would otherwise arise due to the low population sizes that are always present at the population front of a wave of advance (Sec. S2-B below).

(3) In year 1995 Barbujani, Sokal and Oden [5] considered a grid of cells in which the number of hunter-gatherers that became farmers per cell and generation was given by $2\gamma P_F^2 P_{HG}/(P_F+P_{HG})^2$. They also compared to the Lotka-Volterra interaction (i.e., $\gamma P_F P_{HG}$). In contrast to models (1)-(2) above, they used integer numbers for population sizes by replacing the result of each calculated number by its integer part (p. 114). Using integers for population sizes is surely more realistic biologically than using real numbers, although it leads to problems that can be avoided if a dispersal threshold is introduced (as explained in detail in Sec. S2-B below). This was indeed done in Ref. [5], by assuming that individuals do not jump to another cell until a minimum population size is reached (pp. 112 and 115). The authors consider this reasonable because farming people are attached to their land, and tend to move only when they really have to (G. Barbujani, personal communication). Additional justifications of a dispersal threshold (based on ethnography and archaeology) can be found in Sec. S2-C below.

(4) In year 2005 Currat and Excoffier [6] applied the same interaction as in Ref. [5], i.e., they assumed an increase in P_F and a decrease in P_{HG} per cell and generation given by $2\gamma P_F^2 P_{HG}/(P_F+P_{HG})^2$ [7]. They used non-integer numbers for the computation of population densities due, e.g., to migration (Currat, personal communication). As mentioned above, this has been also applied in Refs. [2, 3, 4] and avoids the problems that would otherwise arise due to low population sizes at the population front and, therefore, the need to apply a dispersal threshold (although the latter approach is more realistic biologically because population sizes are necessarily integer, see also Sec. S2-B below).

(5) In year 2012 Rasteiro et al. [8] applied the same interaction as in Refs. [5, 6, 7], i.e., they assumed an increase in P_F and a decrease in P_{HG} per cell (or deme) and generation given by $2\gamma P_F^2 P_{HG}/(P_F+P_{HG})^2$. They considered female and male subpopulations (because they were interested in comparing the consequences of matrilocality, patrilocality and bilocality). They used integer population numbers and did not allow the foundation of a population unless there is at least one male and one female. In their Supp. Info., they analyzed the results as a function of time and therefore simulated both ancient and modern DNA.

(6) Simulations (1)-(5) were performed before it became possible to analyze ancient DNA (aDNA) for a large number of individuals, so the authors could not compare to an ancient genetic cline and estimate the corresponding cultural transmission parameter γ. This was done in year 2017 [3] using a model [9] based on cultural transmission theory [10], according to which interbreeding (vertical cultural transmission) leads to the result that each generation the increase in the number of farmers (and the decrease in the number of HGs) is $\eta p_F P_{HG}/(p_F+P_{HG})$, with $0 \leq \eta \leq 1$. The value of η was estimated from the observed cline of aDNA mitochondrial haplogroup K [3]. Note that in this model, if $p_F \ll P_{HG}$ the increase in the number of farmers becomes independent of the total number of HGs P_{HG}, which is reasonable because it means that interbreeding is limited by the number of farmers p_F. This does not happen with the models used in simulations (1)-(4) above. The detailed derivation of the expression $\eta p_F P_{HG}/(p_F+P_{HG})$ [9] is based on cultural transmission theory [10]. In Ref. [3], Text S9, it was also argued that acculturation (horizontal cultural transmission) would give similar results, and in this case the corresponding equation is $f p_F P_{HG}/(p_F+\gamma P_{HG})$, where f and γ are cultural transmission parameters [11].

S1-B. Alternative models

A related point of interest is that clines can also arise in models that are not based on the combined effects of demic and cultural diffusion. For completeness, we next summarize those alternative models.

(a) One of such models, called isolation by distance, shows that clines can result from random fluctuations (genetic drift) *after* the spread of the Neolithic wave of advance, i.e., in a scenario in which farming populations in all of Europe are initially at their carrying capacity [5]. However, such simulated clines did not encompass the whole continent [5, 12], in contrast to the observed cline of haplogroup K [3]. Moreover, ancient genetics has now shown that the cline of mitochondrial (mt) haplogroup K was formed *simultaneously (not after)* the spread of the Neolithic [3]. Thus a main assumption of this model (farmers at carrying capacity before the formation of the cline) is not satisfied for the Neolithic cline of mt haplogroup K. Therefore, this model is not realistic for the cline of haplogroup K. An additional reason to reject this model (and models (b)-(c) below) is given at the end of this subsection.

(b) Another model, called surfing, assumes that a genetic marker increases its frequency in the low-density region (leading edge) of a wave of advance (again due to random sampling, i.e., drift). In this case the frequency would increase westwards for the Neolithic spread. In fact the frequency of haplogroup K decreases westwards (Figs. 3a-b in our main paper), so in this scenario increases in the frequencies of other haplogroups due to surfing would have caused the decrease of haplogroup K. However, Neolithic clines resulting from simulations with surfing cover distances of only about 500 km (Fig. 6D in Ref. [13]), i.e., apparently they are not consistent with a continent-wide cline such as that of haplogroup K. An additional reason to reject this model (as well as models (a) and (c)) is given at the end of this subsection.

(c) A third mechanism is natural selection (see pp. 86-98 in Ref. [14]). A classical hypothesis is that selection might have been important during Neolithic expansions due to diseases associated with domesticated animals [15]. However selection might have taken place also due to other reasons, and some epidemiological studies in modern populations have indeed suggested that mitochondrial haplogroup K might hypothetically experience positive selection [16, 17]. But positive selection would lead to an increase of the percentage of haplogroup K in the direction of the Neolithic wave of advance (i.e., westwards). Instead, a decrease is observed (Figs. 3a-b in our main paper). Moreover, there is quantitative evidence against both positive and negative selection on haplogroup K because Tajima's D and Fu's F_s tests of early Neolithic K haplotypes indicate neutrality of the corresponding mutations (see Text S1-1 in Ref. [3]). For these reasons, this model is not realistic either. An additional reason to reject this model (and models (a) and (b) above) is given in the next paragraph.

Although we think that the reasons above are enough to dismiss models (a), (b) and (c), in this paragraph we give an additional reason to reject them. There is a substantial increase in the frequency of haplogroups initially present only in HGs (U, U2, U4, U5, U8, ..., see Supplementary Table 6b) along the sea route (from 16% to 60%, Supplementary Table 7b). Along the inland route the increase is not so large but also shows up clearly (Supplementary Table 7c). This is expected in the framework of our model, because the incorporation of HGs into the populations of farmers via interbreeding (and/or acculturation) should obviously increase the frequencies of HG haplogroups in the populations of farmers as their wave of advance travels westwards. Neither isolation by distance, nor surfing, nor natural selection models predict such substantial increases in the frequency *of HG haplogroups* in early farmers along the direction of propagation of the wave of advance. This is an additional reason why we believe that models (a), (b) and (c) above should de dismissed and that our model based on interbreeding (and/or acculturation) seems reasonable.

S1-C. Haplogroup K is useful to analyze the interbreeding behavior

As explained in the introduction of the main paper, a specific genetic marker that meets the following conditions would be ideal for a quantitative study aiming to estimate the percentage of farmers who interbred with hunter-gatherers (HGs): (1) the marker is (nearly) absent in the HGs before the arrival of the firsts farmers; (2) selection and (3) drift (including surfing) effects are not important; and (4) its frequency is high enough in some regions so that a clear gradual decrease, if it exists, can be detected (see Sec. S11). We next discuss these conditions separately.

(1) Condition (1) holds because the subclades of haplogroup K that have been found in European Neolithic individuals were absent in Europe before the spread of farming, except in very few HGs and taking them into account does not change the results (see Sec. S1-D below, specially Suppl. Fig. 1a).

(2) Condition (2) also agrees with the data because several previous analyses have shown that selective effects on haplogroup K are unlikely and that its diversity can be explained by a demographic expansion of farmers during the Neolithic spread. Those analyses include Tajima’s D and Fu’s F_S tests, the dependence of haplotype diversity versus distance from the origin of the Neolithic spread, the shape of mismatch distributions, Mantel tests for genetic and geographic distances, a principal component analysis of the K haplotypes, the star-like shape of their phylogenetic network, and a Bayesian skyline plot (Ref. [3], text S1).

(3) Concerning condition (3), in the last paragraph of the previous subsection we have explained that any model based on drift (not on interbreeding as our model) that attempts to explain the *decrease* in the % of haplogroup K along both routes is not sufficient to explain also the *increase* of the %s of HG-haplogroups along them, whereas our model does explain both spatial trends (precisely because of interbreeding). Thus a model based on drift is clearly useless. In fact our simulations do include drift effects, because we use integer population numbers (not population densities with exactly the same dispersal in all four directions, as in our previous work [3]) and random dispersal, so each simulation run yields slightly different results (Sec S2-D). Thus drift is indeed included in our simulations, but its effect turns out to be negligible. In fact, this is a major difference with our previous model: whereas in Ref. [3] we used real numbers for population sizes (so each simulation run with the same parameter values yielded the same result, i.e., drift and other random effects were not taken into account), here we use integer numbers so that random effects are necessarily

present (each simulation run yields a different result, see Sec. S2-D below), although the results show clearly that it is unimportant (the differences in haplogroup K frequencies between runs are below 0.1%, see Methods in the main paper).

(4) Haplogroup K meets condition (4) because it attains the highest frequency (about 50%) of all mitochondrial haplogroups and, for this reason, its cline can be detected visually (Suppl. Fig. 25) and confirmed computing spatial correlograms (Sec. S8). In Suppl. Fig. 24 below we show that this is possible only for haplogroup K because all other Neolithic mitochondrial haplogroups have too low frequencies (below 20%, see Suppl. Tables 7b-c).

S1-D. Mesolithic samples with haplogroup K

A database of HGs is included as Suppl. Table 5, where individuals with haplogroup K are highlighted with yellow background. We consider the 16 geographical regions listed in the main paper, Figs. 2-3, and classify HGs with haplogroup K into the following groups:

1. Group 1 is composed of 9 HGs[1] with haplogroup K that were contemporaneous with Neolithic farmers in the same region, so it is reasonable to consider the possibility that the presence of haplogroup K in these individuals is due to interbreeding with farmers. Moreover, possibly 4 of these 9 HGs are irrelevant due to their subclades[2].

2. Group 2 consists of 4 HGs[3] with haplogroup-K subclades that have never been found among early farmers to date (i.e., they do not appear in our Suppl. Table 1). For our purposes, these individuals are irrelevant (according to present knowledge) because they do not affect the subclades of haplogroup K considered by us, i.e., those introduced into Europe by the Neolithic population wave of advance.

3. Group 3 is the most interesting one. It is composed by 4 HGs[4] with haplogroup K who satisfy two conditions: (i) they lived clearly before the arrival of the first farmers to their regions, and (ii) all of them display subclades that have been also found among Neolithic farmers. The first observation implies that their subclades are clearly pre-Neolithic (in their respective regions), and the second one that their subclades are part of the cline of haplogroup K analyzed and modelled by us. In other words, these 4 HGs show conclusively that haplogroup K[5] was not entirely absent before the Neolithic in the regions analyzed by us[6]. This implies that we have to check if our simulations (which assume that haplogroup K[7] was entirely absent in HGs before the arrival of farmers) are a reasonable approximation or not. In order to do so, we note that the 4 HGs in group 3 amount to 1.6% of the 254 HGs included in our database (Suppl. Table 5) for the 16 regions considered in our analysis[8]. Therefore, we have repeated our simulations by assuming that 2% of HGs had haplogroup K in all of the cells in our simulation grid, and thus in all 16 regions considered (Suppl.

[1] The 9 HGs in group 1 are, from Suppl. Table 5, a HG with haplogroup K (no subclade was reported) from Ostorf (a Mesolithic enclave surrounded by farmers in Germany) [61], four HGs with subclade K1a1 from Sweden [69, 70] (Pitted Ware culture, which overlapped chronologically with farmers during almost a millennium [62, 63, 64]), a HG with K1e in Denmark [18] slightly more recent than the arrival of the first farmers [23] and three HGs with subclade K1 from Romania [66, 22] whose date ranges overlap with those of the Neolithic site of Ogradena-Icoana in Romania [67]).

[2] Note from the previous footnote that 4 of these 9 HGs (group 1) were reported as K or K1, so a more detailed subclade would be necessary in order to determine if they belong also to group 2 or not.

[3] The 4 HGs in group 2 are two HGs with subclade K1c from Greece [65], a HG with K1c from Romania [22] and a HG with K1b2 also from Romania [33] (Suppl. Table 5).

[4] The 4 HGs in group 3 are, from Suppl. Table 5, three HGs from Denmark [33] with subclade K1e (which is also displayed by 5 farmers from Sweden, see Suppl. Table 1) and a HG from central Anatolia [68] with subclade K2b (which has been also found in 7 early European farmers, of which 2 are from Sweden, 1 from northern France, 1 from Italy and 3 from England, see Suppl. Table 1).

[5] More precisely, the subclades of haplogroup K considered by us, i.e., all of them except those in group 2.

[6] The 16 regions listed in Figs. 2-3 (main paper) and also in Suppl. Table 1, Secs. A, B and C.

[7] More precisely, the subclades of haplogroup K considered by us, i.e., all of them except those in group 2.

[8] The 16 regions listed in Figs. 2-3 (main paper) and also in Suppl. Table 1, Secs. A, B and C.

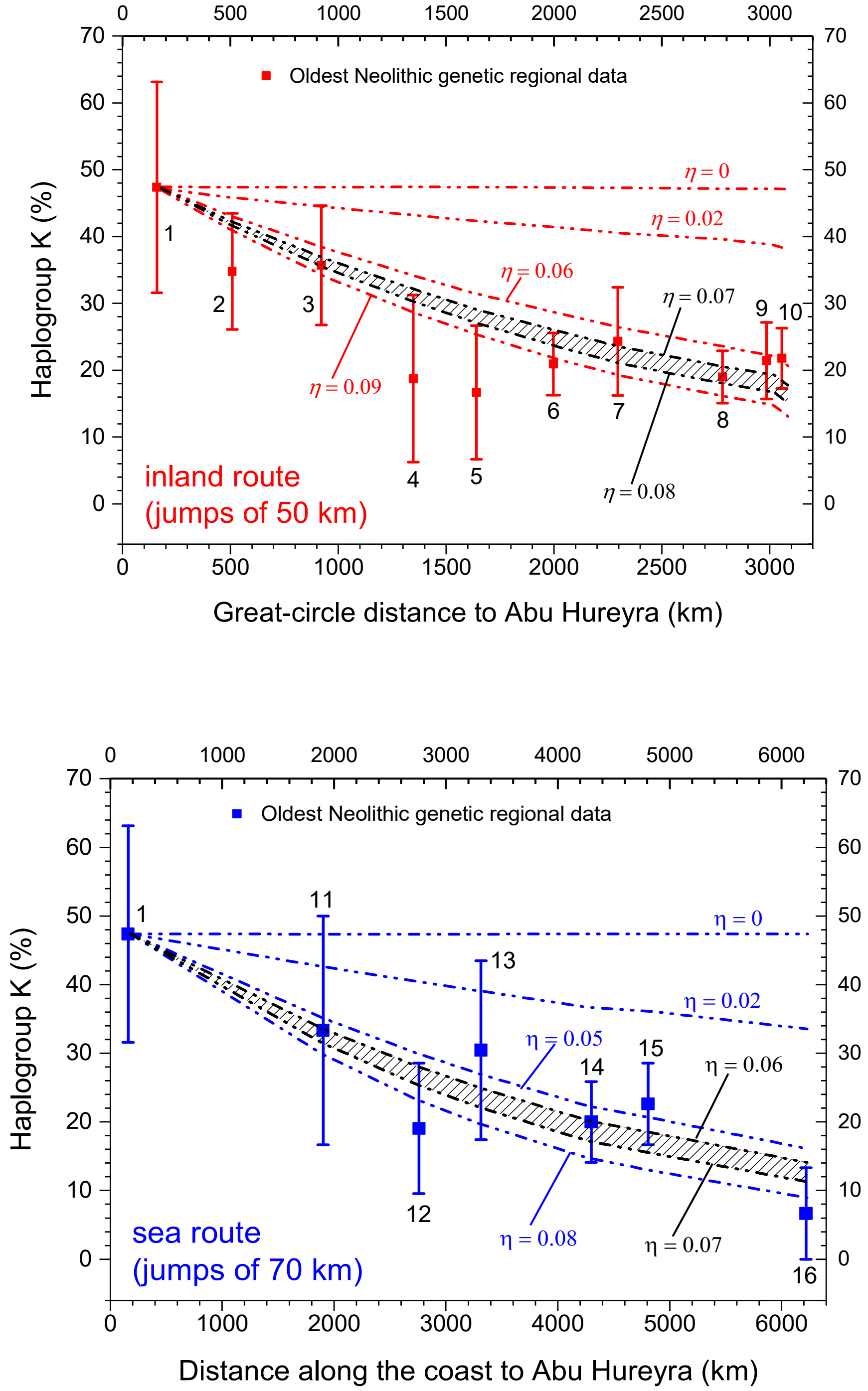
Table 1). It is easy to repeat the mathematical derivation and see that then Eq. (8) in our main paper is the same but Eqs. (11) and (12) are replaced by

$$P_N(x, y, t+1) = R_{0,F}\left[\ P_N(x, y, t) + \frac{2}{100} couples\ HN + \frac{2}{100} couples\ HX\right],$$

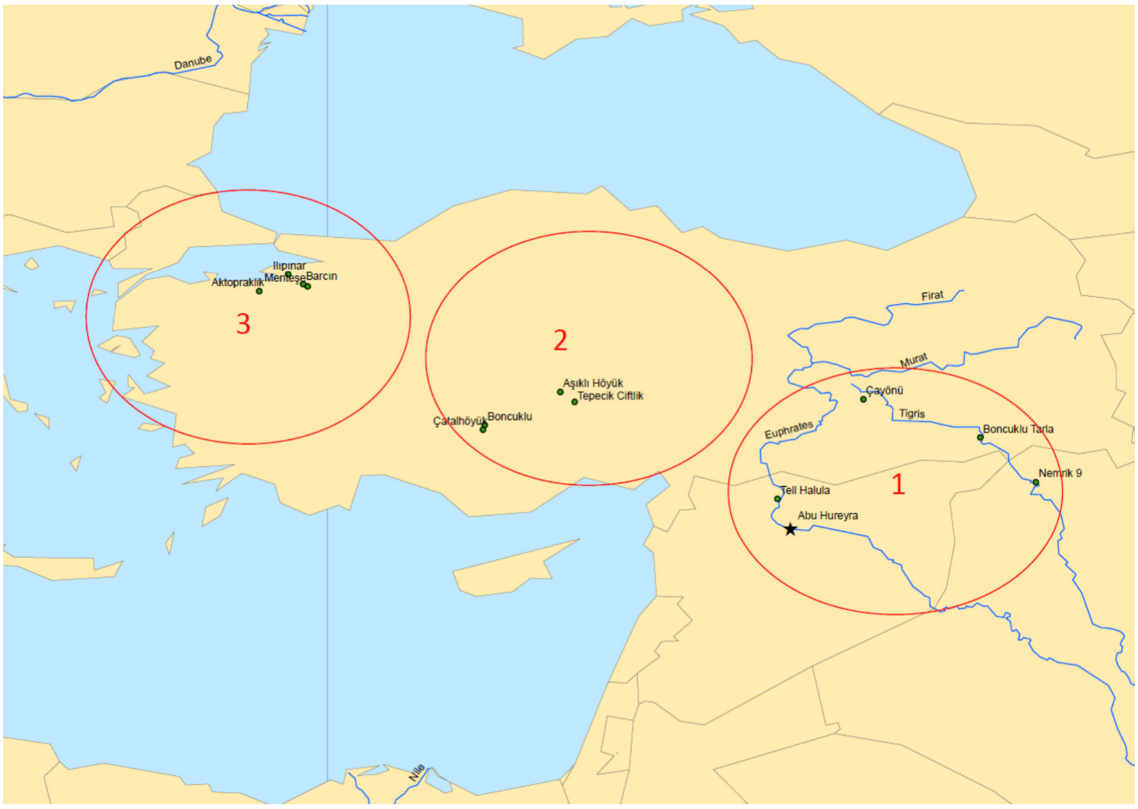
$$P_X(x, y, t+1) = R_{0,F}\left[P_X(x, y, t) + \left(1 - \frac{2}{100}\right) couples\ HN + \left(1 - \frac{2}{100}\right) couples\ HX\right]$$

Our computer programs (publicly available at https://zenodo.org/records/11099220) use these equations and lead to the results in Suppl. Fig. 1a. By comparing to Fig. 3 in our main paper, it is seen that the simulated values of the percentage of haplogroup K (% K) in farmers (curves) are higher in Suppl. Fig. 1a than in Fig. 3 (as expected intuitively). However, the simulation results (curves) in Suppl. Fig. 1a are almost the same as those in Fig. 3 in our main paper. Indeed, the differences between the curves in Suppl. Fig. 1a and Fig. 3 in the main paper are about 1% K or less. We conclude that the results and conclusions in our main paper are not affected by the fact that a small percentage (about 2% or less) of HGs who lived before the arrival of farming had haplogroup K[9]. This also shows that recent criticisms [18, 19] to neglecting that some HGs had haplogroup K are unfounded.

[9] More precisely, the subclades of haplogroup K that have been found in early Neolithic farmers, i.e., all subclades except those in group 2.



Supplementary Figure 1a. This figure is the same as Fig. 3 in the main paper, with only the following difference in the simulations (curves). In Fig. 3 we have assumed that 0% of HGs have haplogroup K. In contrast, here we have assumed that 2% of HGs have haplogroup K in all cells of our simulation grid, and therefore in all 16 regions considered. Regions 1, 2 and 3 are shown in the map in Supplementary Fig. 1b.



Supplementary Figure 1b. Location of Abu Hureyra in northern Syria (star), which is the presumed stating point of Neolithic dispersal in our simulations (see Sec. S2), and the sites in three first regions included in our genetic database (Suppl. Table 1 and Fig. 2a in our main paper). Region 1 is northern Mesopotamia, i.e., south-eastern Anatolia (present-day Turkey), northern Syria and north-western Iraq. Region 2 is central Anatolia. Region 3 is western Anatolia.

S1-E. Neolithic individuals not included in this study

We have computed the error bars in Fig. 3 (main paper) using bootstrap re-sampling (Methods) for regions that have at least 15 early farmers whose mt DNA haplogroup is known. The only regions excluded for this reason are Cyprus (2 individuals), Switzerland (1 individual) and Albania (1 individual), as recorded in Suppl. Table 1.

We have also ignored those regions in the Near East whose populations were not involved in the spread of the Neolithic across Anatolia and Europe. Those excluded regions are, on one hand, the southern Levant (Jordan and Israel, see [20] and Text S3 in [3]) and, on the other hand, the Zagros region[10] [20, 21].

Data from the Danubian sites of the Iron Gates gorge in Serbia (Vlasac and Lepenski Vir) and the site of Malak Preslavets in Bulgaria [22] have been excluded because of the anomalously strong interactions with HGs that have been detected archaeologically [23] and confirmed by an anomalously high HG genomic ancestry due to interbreeding and/or acculturation [22]. It has been suggested that the reason may be the very high densities of HGs near the Danube that have been detected archaeologically, likely due to the considerable fishing sources available to them [23, 24]. Anyway, in Vlasac there is not a single individual whose mt DNA haplogroup is known and that has been dated to the Neolithic (Suppl. Table 1). On the other hand, the site of Lepenski Vir is exceptional because it was not founded exclusively by an incoming population of Neolithic ancestry and culture. Instead, there is a clear co-existence of Mesolithic and Neolithic culture and genes during the Mesolithic-Neolithic transitional phase. Then there are no domestic

[10] The Zagros region includes the sites of Shanidar and Bestansur in northern Iraq [21].

animals neither crops but there is, e.g., pottery characteristic of the Balkan early farmers (see p. 73 in Ref. [23]). New haplogroups appear due to incoming individuals with Neolithic ancestry, but Mesolithic haplogroups are still present during this phase and individuals with Neolithic ancestry are primarily buried in the habitation area (either in houses or between them), whereas those with HG ancestry are preferentially buried in the upslope area at the rear of the village [25]. This is totally different than what we see all over early Neolithic Europe, where there is an abrupt replacement in the economy, culture and genes due to the foundation of new sites by incoming Neolithic populations. It is not reasonable to expect that interbreeding will take place at a similar rate in so different situations, namely a single site with high fractions of individuals with Neolithic and Mesolithic genetic ancestry and culture (Lepenski Vir) versus settlements founded by Neolithic individuals with only occasional incorporation of HGs (in most of Europe). Therefore, it is clearly very reasonable not to include in our analysis the site of Lepenski Vir. Anyway, in Lepenski Vir there are only 2 individuals whose mt haplogroup is known and that have been dated to the Neolithic (Suppl. Table 1). For clarity and consistency with the reasoning in this paragraph, we prefer not to include them. However including these 2 individuals would not change the conclusions, because the %K in Suppl. Table 3 (region 5 Romania and Serbia) would be 15.6% instead of 16.7%, so Fig. 3 would be essentially the same. Finally for Malak Preslavets Mathieson et al. [22] detected that its farmers "have significantly more HG ancestry than other Balkan Neolithic populations", so they are definitely a local exception to the continent-wide genetic cline that we analyze and it is not reasonable to include them. Indeed, the percentage of haplogroup K is 0% in Malak Preslavets ($n = 14$, Suppl. Table 1).

Similarly, data from the UK and Ireland have not been included because they also show huge deviations from the continental trend of the percentage of haplogroup K. This is easily seen from Fig. 3a in the main paper by noting that Suppl. Table 1 gives percentages of haplogroup K of 33.9% in the UK and 29.1% in Ireland, with great-circle distances to Abu Hureyra of 3,738 km and 4,097 km, respectively. These values are far larger than those suggested by the continental cline (Fig. 3a in the main paper). The origin of these insular deviations is unknown, but three possible reasons are the following.

(1) A first possible reason could be genetic bottleneck effects, due to the fact that perhaps small communities crossed the Channel from the continent to England and Ireland. Indeed, archaeologists have noted for long that pottery styles are similar but not exactly the same at both sides of the Channel, and consider that 'it is highly likely that founder effects and drift have been operating in any groups that crossed the Channel. That is to say, specific founder communities would only have produced some fraction of the range of forms present in their ancestors' in the continent (p. 184 in [23]). If this was the case, it seems probable that founder effects could also show up in the genetic composition of the populations.

(2) A second possible reason (perhaps in combination with the first one above) is that genomic work has detected strong affinities between British and Iberian Neolithic populations, indicating that British Neolithic people derived much of their ancestry from the Mediterranean route via Iberia [26] but not all of it, so it is unclear to what extent farmers of the Mediterranean and inland routes mixed during the 1,000-yr delay that it took for Neolithic wave of advance to cross the Channel [23] and how much each stream contributed to the populations crossing of the Channel.

(3) A third possible reason is the following. In sharp contrast to most of early Neolithic continental Europe, in the UK and England specific families used megaliths for burial practices [27, 28], so a substantial part of the sampled individuals are not representative of the whole population and may instead belong to ruling elites [28]. We can quantify this point by means of the following table, that contains all Neolithic pairs of individuals that are relatives and have been detected in the UK and Ireland so far. Some papers in our database (Suppl. Table 1) that report mitochondrial haplogroups from Neolithic individuals in the UK and Britain [29, 30, 26] did not include kinship analyses. Such analyses were performed by Scheib et al. [31] for 2 individuals, Sánchez-Quinto et al. for 16 individuals [27], Fowler et al. for 35 individuals [32], Allentoft et al. for 7 individuals [33], and Cassidy et al. [28] (main paper and Sec. S6.5) for 75 individuals (besides 11

individuals already analyzed in Ref. [27]). This gives a total of 135 individuals among which kinship relationships have been investigated. Out of these 135 individuals, according to the table below there are 63 different individuals (in bold and underlined) for which relatives have been detected. This is 47%, i.e., about half of the population. Thus a substantial part of the sampled individuals are not representative of the whole population but of specific families. This may lead to values of the percentage of haplogroup K that are not representative of the whole population. In contrast, the simulations (lines in Fig. 3 in our main paper) consider the whole population. This is a third reason that suggests not to include data (error bars) from the UK neither Ireland in Fig. 3 in our main paper.

pair of individuals	**site(s)**	**region**	**Ref.**
Sk.4/799 & **Sk.1/880**	Trumptington Meadows	England, UK	[31]
prs002 & **prs017**	Primrose	Ireland	[27]
prs017 & **prs018**	Primrose	Ireland	[27]
car004 & **prs007**	Carrowmore & Primrose	Ireland	[27]
prs006 & prs007	Primrose	Ireland	[27]
prs006 & prs004	Primrose	Ireland	[27]
car004 & **NG10**	Carrowmore & Newgrange	Ireland	[28]
car004 & **CAK533**	Carrowmore & Carrowkeel	Ireland	[28]
car004& **MB6**	Carrowmore & Millin Bay	Ireland	[28]
NG10 & **CAK532**	Newgrange & Carrowkeel	Ireland	[28]
PB675 & **PB357**	Parknabinnia	Ireland	[28]
GNM1007 & **GNM1076**	Glennamong cave	Ireland	[28]
PA Sk 3332 & **PA SK 3324**	Fussles Lodge [30]	England, UK	[28]
bal004 & **mid001**	Balintore & Knowe of Midhowe [27]	Scotland, UK	[28]
bal004 & **mid002**	Balintore & Knowe of Midhowe [27]	Scotland, UK	[28]
prs009 & **prs016**	Primrose	Scotland, UK	[28]
bal004 & **lai001**	Balintore & Knowe of Lairo [27]	Scotland, UK	[28]
prs009 & **prs013**	Primrose	Scotland, UK	[28]
PB672 & **PB754**	Parknabinnia	Ireland	[28]
PN107 & **PN04**	Poulnabrone	Ireland	[28]
ARD2 & **PB443**	Ardcrony & Parknabinnia	Ireland	[28]
PB186 & **PN06**	Parknabinnia & Poulnabrone	Ireland	[28]
PB186 & **PN07**	Parknabinnia & Poulnabrone	Ireland	[28]
NEO624 & **NEO625**	Banks tomb (Ref. [33], Supp. Info. part 1, table S3c1)	Scotland, UK	[33]
27 individuals (Fig. 1 in [32])	Hazleton North	England, UK	[32]

Pairs of Neolithic individuals in the UK and Ireland that are relatives. Individuals appear the first time in bold and underlined.

Ancient DNA from Latvia and the Dnieper Rapids has led to the conclusion that Anatolian farmer-related ancestry did not reach the Baltic neither Ukraine [34]. This suggests that these regions underwent a Neolithisation process totally different from that modelled in the present paper, so they are not included in our analysis.

The new genetic database of Neolithic individuals is included as Supplementary Tables 1-3.

S2. Details on the simulation model

Our model modifies a previous one [3] in the following ways.

(i) Initial archaeological condition. Neolithic traits (i.e., domestication of plants and animals) appeared at different places and times in the Near East and over several thousand years. These traits eventually led to a homogeneous, well-established set of farming and stockbreeding practices (called the Neolithic package). This happened in the so-called pre-pottery Neolithic B and C (PPNB/C) cultures that spread the Neolithic package across Europe [35]. In a previous study [3] we considered the oldest PPNB site in Syria according to the database in Ref. [35], namely Ras Shamra. It is dated 8,233 calibrated (cal.) yr before the common era (BCE). In Ref. [3] we assumed that at this date all of the cells in our simulation grid were empty of farmers except the cell that contained Ras Shamra, and we simulated a Neolithic wave of advance that spread from that Syrian cell to Anatolia and Europe. However, data published after Ref. [3] include a Neolithic site in Anatolia (Boncuklu) that is older (8,260 cal. yr BCE) [36] and located to the West of Ras Shamra (Syria, 8,233 cal. yr BCE). We avoid this inconsistency by using as origin of the expansion, instead of Ras Shamra as in Ref. [3], the oldest PPNB site in Syria, Anatolia and Iraq, namely Abu Hureyra[11]. Supplementary Figure 1b shows the location of this site (as a star), which is located in northern Mesopotamia (northern Syria, northwestern Iraq and southeastern Anatolia), i.e., region 1 in our database, which is the presumed region of origin of the Neolithic wave of advance that spread across Anatolia and Europe [21]. Alternatively, instead of using a single cell as the spatial origin of agriculture, we could use a wider region but it is known from previous simulations that the results would be much the same [35].

In order to attain agreement between the inland simulations and the archaeological (not genetic) data shown in Fig. 1b in our main paper (red lines and error bars, respectively), we take into account that the oldest Neolithic date of Abu Hureyra is 9,557 cal. yr BCE (Suppl. Table 4) but the front could have spread from there later. Thus, to attain this agreement we use the date 8,718 cal. yr BCE for the start of the spread from Abu Hureyra. Such a date for the start of the Neolithic spread is within the PPNB period of Abu Hureyra [37]. Note that we choose the intercept but not the slope (inverse of the speed) of the Neolithic front (red line in Fig. 1b in the main paper) because the slope is fixed by the dispersal distance, net fecundity and generation time, all of them estimated from ethnographic data (their values are given and justified in Methods).

For the reasons explained in the previous two paragraphs, we use Abu Hureyra and 8,718 cal. yr BCE as origin of the Neolithic spread in the main paper. However, **in this section S2 we are interested in comparing the new model to that in Ref** [3]**, so we use Ras Shamra and 8,233 cal. yr BCE as origin (as in Ref.** [3]**).**

As a side note, we mention that a non-genetic simulation model published a decade ago [35] assumed a later date for the initial of the dispersal, namely 7,051 cal. yr BCE, i.e., 9,000 cal. yr Before Present (BP) because it considered all early Neolithic archaeological data in Europe, leading to a dispersion of several millennia for each distance from the dispersal origin (see Fig. 5 in Ref. [35]). In contrast, in more recent work [3, 38, 39] and the present paper only the earliest date per region is considered, which is more appropriate to compare to the corresponding arrival time from the simulations.

(ii) Integer population sizes. The most important change is that our model improves the previous one [3] by using integer numbers for population sizes. The reason is that non-integer values for the number of individuals are obviously unrealistic from a biological perspective. Subsections B-D below introduce and discuss this issue in detail.

(iii) Homogeneous geography. In this work we use a homogeneous geography, in the sense that individuals can live in any cell of the simulation grid (in contrast, in Ref. [3] we used a real geography, i.e., a map with seas and mountains). A homogeneous geography has three clear advantages: (1) it makes simulations substantially faster and simpler; (2) the spread rate in a homogeneous geography can be checked by comparing to analytical results (this is done in subsection A2 below); and (3) in a real geography it is observed in some cases that a node of the simulation grid is on a sea location, which avoids that individuals can cross it in spite of the fact that their dispersal distance (assumed in the simulations) is longer than the sea

[11] This is the oldest PPNB site in the area mentioned, according to the database in Ref. [47].

distance involved. This happens, e.g., from Albania to southern Italy (see Text S6 in Ref. [3]). A more detailed discussion on this point is provided in Sec. S3-B below.

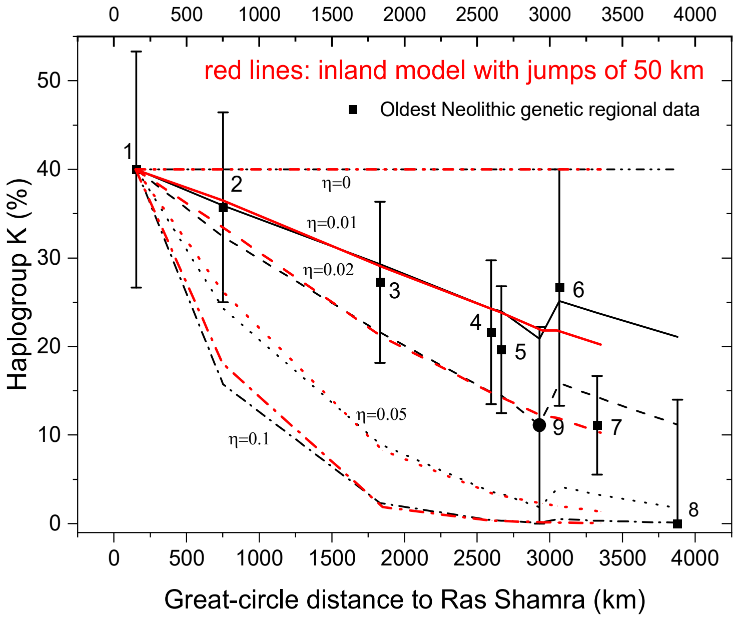
We need to be sure that in our simulations a homogeneous geography is a valid approximation to the actual one (i.e., to a map of Europe with seas and mountains). In other words, we need to check that, using real numbers as in Ref. [3], the homogeneous geography (used in the present paper) yields similar results to the non-homogeneous geography (map of Europe) used in Ref. [3]. Therefore, we begin (subsection A below) by using our new, homogeneous geography with real numbers for the population sizes (as in Ref. [3]) to compare to the results of Ref. [3] (non-homogenous geography).

S2-A. Homogeneous versus non-homogeneous geographies

S2-A1. Genetic cline

Although in the main paper we consider an inland model (with square cells of 50x50 km) and a sea model (with cells of 70x70 km), for clarity we consider here only the inland model (the sea one will be discussed in Sec. S3 below). The homogenous geography used in this work represents the surface of Europe, the Near East and parts of Asia and northern Africa as a square grid of 300 x 300 nodes (corresponding to a grid size of 300 x 50 km = 15,000 km). Any geographical features (sea, mountains, rivers, etc.) are not considered. In this subsection A1 we use real numbers for population sizes (as in Ref. [3]). The only difference between the model in this subsection A1 and that in Ref. [3] is that the latter used a real, non-homogeneous geography (i.e., a real map with seas and mountains [3]). Therefore, both models share the following features. Three processes are computed per generation, namely dispersion, interbreeding and population growth (net reproduction), as explained in the main paper (Methods). As in Ref. [3], the saturation density of the farmer population is 1.28 individuals/km^2 (as estimated from ethnographic data), i.e., 3,200 individuals/cell of (50 km)^2, and the initial population of farmers in the simulations consists of 3,200 farmers located at the cell containing the oldest PPNB site in Syria considered in Ref. [3], Ras Shamra, dated 8,233 cal. yr BCE.

Supplementary Figure 2a shows the fraction of farmers with haplogroup K versus distance from the origin (initial node) of the dispersal, for several values of the interbreeding parameter η. The red lines are the results for the *homogeneous* geography used in our main paper (Methods). Note moreover that in the main paper we use *integer* numbers for the population sizes (as explained in Sec. S2-B below). In contrast, the red lines in Suppl. Fig. 2a have been obtained by using *real* numbers for the population sizes, so that we can compare to Ref. [3], which also used *real* numbers. The black lines in Suppl. Fig. 2a are the same as those in Fig. 3 in Ref. [3], i.e., they give the results from the *non-homogeneous* geography using *real* numbers [3]. The genetic data (squares and error bars) in Suppl. Fig. 2a are the same as in Fig. 3 in Ref. [3]. They give the mean and 80% confidence-level bounds of the observed averages of the percentage of haplogroup K (%K) for early farmers in several regions.



Supplementary Figure 2a. Observed and simulated percentages of haplogroup K among early farmers as a function of distance. Red lines have been obtained using the homogeneous geography (inland model) introduced in Sec. S2. Black lines were obtained in Ref. [3] using a non-homogeneous geography (i.e., with seas and mountains). ***Both models use real numbers** for the population sizes.* Black squares are the observed fractions with their corresponding error bars (80% confidence-level), as reported in Ref. [3], for regions 1 Syria PPNB, 2 Anatolia, 3 Hungary-Croatia, 4 Eastern Germany LBK, 5 Western Germany LBK, 6 North-eastern Spain Cardial, 7 Spain Navarre, 8 Portugal coastal Early Neolithic, and 9 Sweden. To facilitate comparison to Ref. [3], simulations begin in Ras Shamra at 8,233 cal. BCE. Each simulation ends at the average date of the individuals in the considered region whose mtDNA haplogroup has been determined (Data S1 or S3 in Ref. [3]).

It is very interesting that in Suppl. Fig. 2a both the homogeneous (red) and the non-homogeneous (black) geographies produce very similar results. For η =0 (no interbreeding), both geographies predict that the initial fraction remains constant (horizontal straight line). Also for both geographies, for all values of $\eta \neq 0$ the %K decreases with increasing distance from the origin, as expected due to interbreeding with hunter-gatherers (who lack haplogroup K). The quantitative agreement between both geographies (black and red lines in Suppl. Fig. 2a) is remarkable. It implies that a homogeneous geography is a valid approximation. For the case $\eta = 0.02$, both geographies yield predictions consistent with the observed data (error bars). A larger value of η, i.e., more interbreeding between farmers and hunter-gatherers, results in a more rapid decline of fraction of haplogroup K with increasing distance, also as expected. For $\eta > 0.02$, both geographies yield fractions that are lower than the measured ones. Thus we see that the *homogeneous* geography (introduced in this section) leads to the same conclusion as that obtained from the *non-homogeneous* geography (map of Europe) used in Ref. [3], namely that $\eta \approx 0.02$ [3]. Therefore, there is no need to consider non-homogeneous geographies (which lead to substantially more complicated and slower simulations than homogeneous ones). This is one reason why we have used a homogeneous geography to obtain the results reported in our main paper (other reasons have been explained in point (iii) above).

We stress that in this subsection we have used *real* numbers to compare our new model (homogeneous geography) to Ref. [3] (non-homogeneous geography). However, in the main paper and the sections below we will use *integer* numbers and find that this leads to substantial corrections.

Although this is not strictly necessary for the purposes of this section, for the sake of clarity we mention two related points.

(1) In the homogeneous geography (red lines in Suppl. Fig. 2a) the simulated wave of advance has not arrived yet to Portugal (error bar 8) at the average date of the early farmers in Portugal whose mtDNA has

been determined. This is simply due to the fact that this first model ('inland model') does not include long travels along the coast for simplicity, but in reality Portugal was reached following a sea route [38] by means of longer jumps than those of inland travel (50 km, from ethnographic data [40]), and this effect naturally leads to a faster spread (this is solved in the main paper and Secs. S3-S5 by introducing a sea model, i.e., simulations with longer jumps along the coast).

(2) Another difference between the homogeneous geography (this section) and the non-homogeneous one (Ref. [3]) is that the homogeneous geography does not yield a minimum in Sweden (error bar 9 in Suppl. Fig. 2a). This is also due to the existence of a sea dispersal route along the Mediterranean Sea (see Sec. S8 in Ref. [3]).

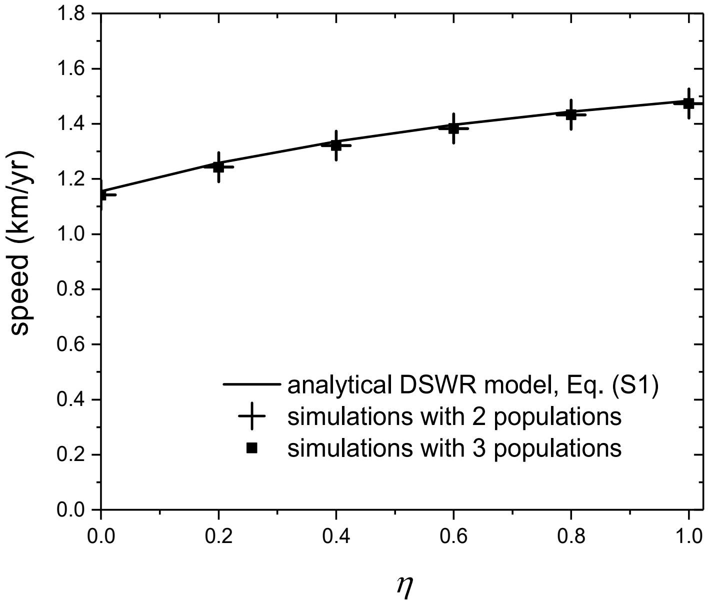
In the main paper and Secs. S3-S5 we consider not only an inland but also a sea homogeneous geography, and this will improve both points (1) and (2).

We stress that the difference with the models in our main paper is that in this subsection A (as in Ref. [3]) we use *real* population numbers. In contrast, the model used in the main paper computes *integer* population numbers, by rounding each real number to its nearest integer number. Given that in reality all population numbers are obviously integer, the model in our main paper is more reasonable than the model in Ref. [3] and this subsection A. However, using integer number causes additional complications, which we discuss and solve in subsections B-D below.

S2-A2. Front speed

As mentioned at the beginning of Sec. S2, point (iii), an important advantage of performing simulations on homogeneous geographies is that it is possible to check the front speed (i.e., the spread rate of the wave of advance) by comparing results obtained from simulations to those from analytical equations. This is done in this subsection.

First we consider a two-population model, namely famers and hunter-gatherers (HGs) interacting via vertical cultural transmission (interbreeding), as described previously [9]. The speed of the front on a homogeneous geography is estimated by using a simulation program (written in FORTRAN) to determine, at each node of the grid, the time (number of iterations or generations) when the population number of farmers is equal to its saturation value divided by 10 (other values instead of 10 give the same results, because the shape of the front is constant according to well-established Physics theory [9, 11]). The results of the speed along the horizontal (or vertical) direction of the grid are shown as crosses in Suppl. Fig. 2b. As in subsection A1, all simulations in this subsection A2 use real numbers.



Supplementary Figure 2b. Front speed (i.e., spread rate) as a function of the intensity of interbreeding η according to Eq. (S1) and to simulations using two and three populations. Simulations have been performed by using real numbers for the population sizes (as in Suppl. Fig. 2a and Ref. [3]).

The full line in Suppl. Fig. 2b is the front speed according to the corresponding analytical model, namely

$$speed = min_{\lambda > 0} \frac{ln\left\{R_{0,F}(1+\eta)\left[\frac{p_e+1}{2}+\frac{1-p_e}{2}cosh(\lambda r)\right]\right\}}{\lambda T}. \tag{S1}$$

This result is easily obtained by using Eq. (17) in Ref. [40] and replacing $R_{0,F}$ by $R_{0,F}(1+\eta)$, as follows from the method derived in detail in Ref. [9] [12]. As in our main paper and Ref. [3], we have used the values $r = 50$ km, $R_{0,F} = 2.45$, $p_e = 0.38$ and $T = 32$ yr, estimated from ethnographic data.

We also run another FORTRAN program to simulate the three-population model in the main paper and the rest of these Supplementary Methods (namely HGs, farmers with haplogroup K and farmers without haplogroup K). In this case we estimate the front speed by considering the sum of the population numbers of farmers with and without haplogroup K. The results are shown as squares in Suppl. Fig. 2b. Obviously the front speed of the total population of farmers should be the same as for the farmers in the two-population model (because the haplogroup of farmers does not have any effect on their dispersal neither on their reproduction), and this is indeed seen in Suppl. Fig. 2b (the crosses agree exactly with the squares). More importantly, the differences between the simulations and Eq. (S1) are about 1% or less. Therefore, we can be sure that our programs implement correctly the processes of reproduction, dispersal and cultural transmission as described in the Methods section of the main paper. This possibility to check the simulations by comparing to analytical results is a crucial advantage of using a homogeneous model.[13]

[12] For the purposes of checking simulation results, Eq. (S1) is more accurate than Eq. (S60) in Ref. [3], because that one was derived under the assumption a continuous space, i.e. using a continuous-space random walk (CSRW) model, but simulations are necessarily performed on a grid, so they correspond to a discrete-space random walk (DSRW) model [40].

[13] For the sake of clarity, we mention that the speed in Fig. 1b in the main paper (red line) is slower than the speed in Suppl. Fig. 2b for $\eta = 0$ due to two reasons. First, the speed in Suppl. Fig. 2b corresponds to the horizontal/vertical direction of the simulation grid. In contrast, the regions considered in Fig. 1b in the main paper are not located on a such a direction. This is explicitly seen from their rectangular coordinates, which are given in Supp. Table 4, columns 'Number of nodes X (d=50km)' and 'Number of nodes Y (d=50 km)'. This difference leads to slower speeds in Fig. 1b in the main paper relative to Suppl. Fig. 2b (see Figs. 1-2 in Ref. [40] for an explicit proof and its intuitive explanation). Second, the simulations in

S2-B. Difficulties in modelling dispersal using integer numbers

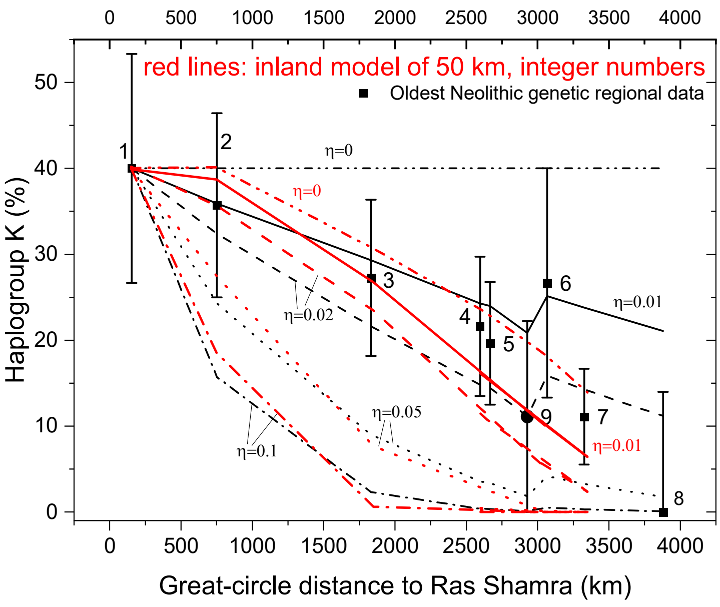
The simulations in subsection A above use real numbers for the population sizes (this has been useful to compare to Ref. [3] and, in this way, check that the homogeneous geography used here is accurate). However, as explained in Methods (and also at the beginning of Sec. S2, point (ii)), it is much more reasonable biologically to use integers. This has been done in previous work, both on genetic clines [5] and front speeds [41]. The simplest approach is to replace, in all calculations in the simulations, each computed number of farmers or HGs by its nearest integer (which yields the red lines in Suppl. Fig. 3). But this causes some problems that we next discuss and solve. For clarity we consider the case without interbreeding ($\eta = 0$). The problems are most easily seen by considering the following very simple examples.

(i) As a first example, suppose that in a cell of our simulation grid we have 9 K-farmers (i.e., farmers with haplogroup K) and 15 non-K farmers (i.e., farmers without haplogroup K). Assume that a fraction $p_e = 0.38$ of these individuals stays at the original cell (main paper, Methods) and the rest are distributed equally among the 4 nearest cells. Then, if we approximate the result of each computation by its nearest integer ($NINT$), the number of K-farmers that jump in each of the 4 directions is $NINT\big((1 - 0.38) \cdot 9/4\big) = NINT(1.395) = 1$, and the number of non-K farmers that jump in each of the 4 directions is $NINT\big((1 - 0.38) \cdot 15/4\big) = NINT(2.325) = 2$. Thus, in the nodes where farmers arrive, their percentage of haplogroup K is $\%K = 1/(1 + 2) \cdot 100 = 33\%$. But this is less than the initial value, namely $\%K = 9/(9 + 15) \cdot 100 = 38\%$. This is why, if using integer numbers (red line in Suppl. Fig. 3), the line for $\eta = 0$ is not horizontal but decreases (in contrast, the lines in Suppl. Fig. 2a for $\eta = 0$ are horizontal and have been obtained by using real numbers, i.e., without using the nearest-integer approximation above). Clearly this is just a drift effect, due to the low population numbers considered (as shown explicitly in example (ii) below). It could be argued that such an effect might have happened in reality. However, in Secs. S2-C-D below we show that it is more reasonable to take into account that both Ethnography [42] and Archaeology [43, 23] indicate that humans do not live in populations of arbitrarily low size (moreover, taking this into account leads to a horizontal cline for $\eta = 0$).

(ii) We stress that this problem is important only if the population number is sufficiently small. We show this explicitly with a second example. Consider a node in which, instead of 9 K-farmers and 15 non-K farmers (as in example (i) above), there are 30 K-farmers and 50 non-K farmers. Therefore, initially $\%K = 30/(30 + 50) \cdot 100 = 38\%$ is the same as in example (i) above. The number of K-farmers that jump in each of the 4 directions is $NINT\big((1 - 0.38) \cdot 30/4\big) = 5$, and the number of non-K farmers that jump in each of the 4 directions is $NINT\big((1 - 0.38) \cdot 50/4\big) = 8$. Thus, in the nodes where farmers arrive, their percentage of haplogroup K is $\%K = 5/(5 + 8) \cdot 100 = 38\%$. This is the same as the initial value, namely 38%. Therefore, we see that the %K is constant if the population numbers are large enough. This does not happen in example (i) above. Similarly, at the leading edge of the front (pioneering populations of farmers) the population is small, and this is why the red line in Suppl. Fig. 3 for $\eta = 0$ is not horizontal. In Secs. C-D below we show that, in fact, this line is horizontal if proper care is taken of the fact that humans do not live in arbitrarily small groups (as mentioned above).[14]

Suppl. Fig. 2b use real numbers but those in Fig. 1b in the main paper use integer numbers and a dispersal threshold, which also slows down the speed (for an example, see the caption to Suppl. Fig. 4 below).

[14] It is interesting to note that the cline for $\eta = 0.02$ in Suppl. Fig. 3 using integer numbers (red dashed line) is not so different after all that the cline for the same value ($\eta = 0.02$) in the same figure without using integer numbers, and that this cline crosses all error bars (black dashed line [3]). So we might be tempted to apply the model with integer numbers, without any of the additional improvements introduced below. However, this would imply assuming that a single individual can disperse and reproduce, which is not realistic biologically. Moreover, both ethnographic and archaeological data strongly indicate that assuming that only one or a few families disperse together is not realistic either (see below). This dispersal-threshold effect leads



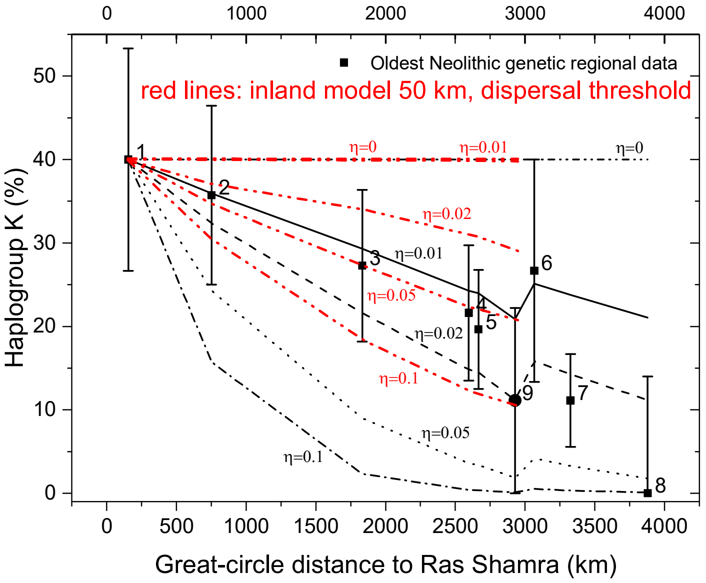
Supplementary Figure 3. Observed and simulated percentages of haplogroup K among early farmers as a function of distance. Red lines have been obtained using the homogeneous geography (inland model) *with* ***integer numbers without dispersal threshold.*** As in Supplementary Figure 2a, the simulated wave of advance has not arrived yet to Portugal (error bar 8) due to the fact that Portugal was reached by sea (this is solved in the main paper and Secs. S3-S5 by using a sea model with longer jumps than the inland model).

S2-C. First model. Integer numbers and a dispersal threshold

We solve the difficulty explained in the previous subsection as follows. We impose the condition that dispersal takes place only after the number of farmers at a given node reaches a minimal number (dispersal threshold). This is very reasonable because there is a well-known minimum population size for human groups, which is thought to be related to the benefits of food-sharing, division of labor and other forms of cooperation [42]. Moreover, archaeologists have remarked that first Neolithic settlements in local regions generally consisted of several houses, not a single one (see Ref. [23], p. 97). For the Early Neolithic Bandkeramik, the population density has been estimated as a function of time using archaeological data and the minimum density is about $p_{F\,min}$ =0.06 individuals/km^2 (see Fig. 6 in Ref. [43]). Therefore, since each cell in our simulations has 50x50=2,500 km^2, the minimum population number acceptable per cell is 0.06 individuals/km^2 · 2,500 km^2 = 150 individuals per cell. Interestingly, this value can be also justified without resorting to any archaeological data, because it is close to the minimum size for a human reproductive network to be viable [42]. Thus, for the simulations to be ethnographically realistic, if a cell is empty of farmers, the first group of arriving farmers must include more than about 150 individuals. Therefore, we require that farmers do not jump from any cell unless it has at least 970 farmers, because the number of farmers that jump in each direction is then at least $NINT\big((1-0.38)\cdot 970/4\big) = 150$ individuals. Thus, in Suppl. Fig. 4 we present the results using integer population numbers as in Suppl. Fig. 3 but now requiring that dispersal takes place only from cells with at least 970 individuals (red lines). In Suppl. Fig. 4 we see that the red cline for $\eta = 0$ is horizontal (in contrast to Suppl. Fig. 3). **Note that the cline that agrees best with the data is that for $\eta = 0.02$ if using the model with real numbers (black lines in Suppl. Fig. 4)** [3] **but this changes to between $\eta = 0.05$ and $\eta = 0.1$ if using the model with integer numbers and a dispersal threshold (red lines in Suppl. Fig. 4).** We will comment on this after we present our final model (subsection S2-D below).

to a much higher value of η, namely between $\eta = 0.05$ and $\eta = 0.1$ (Secs. S2-C and S2-D). For all these reasons, it is necessary to develop more realistic models, as done in Secs. S2-C and S2-D.

The implementation of a dispersal threshold in coastal cells is discussed in Sec. S3-B below.



Supplementary Figure 4. Observed and simulated percentages of haplogroup K among early farmers. Red lines have been obtained using the homogeneous geography (inland model) *with* ***integer numbers and a dispersal threshold*** of 970 individuals (in this model, the population number is not conserved, and this is solved in Sec. S2-D and Supplementary Figure 5). The simulated wave of advance has not arrived yet to regions 6-8 (not only to region 8 as in Supplementary Figures 2a and 3) because the dispersal threshold leads to a slower front. This is solved in the main paper and Secs. S3-S5 (and the main paper) by using a model with longer jumps for the sea route implied by Archaeology [38] for those regions.

S2-D. Second model. Integer numbers, dispersal threshold and random dispersal

This second model improves the first one (previous subsection) and has been used to obtain all of the results reported in our main paper.

Consider a simple example in which the initial number of farmers in a given cell is 132. Then the number of farmers who jump in each direction is $NINT\big((1-0.38)\cdot 132/4\big) = NINT(20.46) = 20$ and the number who stay at the original cell is $NINT(0.38 \cdot 132) = NINT(50.16) = 50$. Hence the total final number of individuals is 20·4+50=130, but this is different from the original number (132). In other words, 2 individuals have disappeared without reason. Thus the number of individuals is not conserved in the dispersal. In the simulations leading to Suppl. Fig. 5 we have solved this difficulty in the following way. For each generation, node and direction of dispersal (North, South, East and West), a random number between 0 and 1 is generated. If the random number is smaller than the decimal part of the number of farmers who jump in that direction (computed without using *NINT*), an extra farmer is added. For example, if the initial number of farmers in a given node is 132, then the number of farmers who jump in each direction (computed without using *NINT*) is $(1-0.38)\cdot 132/4 = 20.46$. If for a given direction, the random number is, e.g., 0.2, then since 0.2 < 0.46, applying the rule above we have that 21 farmers move in that direction. The number of individuals who stay at the node considered is 132 minus the sum of those who disperse in each of the 4 directions (which may be different). In this way, the final number (individuals dispersed plus individuals who stay) and the initial number (individuals before dispersal) are the same. Each model run produces a slightly different result (so we call this approach 'random dispersal'). However, for any given values of the position and time, the difference in the percentage of haplogroup K between different simulation runs are only about 0.1%, so the figures do not change (i.e., it is not necessary to average over different runs).

The red curves in Suppl. Fig. 5 (especially those crossing more error bars, i.e., $\eta = 0.05$ and $\eta = 0.1$) **are rather similar to those in Suppl. Fig. 4, so the conservation of the population number** (i.e., the random dispersal algorithm introduced in this Sec. S2-D) **does not imply important corrections quantitatively**. Note, however, that this model (Suppl. Fig. 5) solves all problems, namely: (1) the fact that population numbers are integer (which was not taken into account in Ref. [3]); (2) the existence of a dispersal threshold, which is implied by both ethnographic and archaeological data and leads to a uniform and constant percentage of haplogroup K for $\eta = 0$); and (3) and the conservation of the population number in the dispersal step.

Importantly, in Suppl. Fig. 5 the red curve that agrees best with the data (error bars) is not that for the model in Ref. [3]**, namely $\eta = 0.02$ (dashed black curve in Suppl. Fig. 5), but for substantially higher values of η, between $\eta = 0.05$ and $\eta = 0.1$ (red curves). This is why, also in our main paper (Fig. 3) for the two routes, the value $\eta = 0.07$ is substantially higher than the value estimated previously ($\eta = 0.02$)** [3]**.** This shows that using integer for population sizes with a dispersal threshold (rather than real numbers) not only makes more biological sense but also has substantial quantitative implications. Here we have plotted only the curves for the values $\eta = 0.02$, $\eta = 0.05$, etc. because our main aim here is to compare with the curves in Ref. [3]. Figure 3 in the main paper shows curves for the inland and Mediterranean routes, with additional values of η and updated error bars (including additional regions), and using Abu Hureyra as origin (not Ras Shamra as in Ref. [3] and this section). Figure 3 in the main paper shows that the best agreement between the new model (lines) and the data (error bars) is attained for $\eta = 0.07$. Therefore, using a threshold for the dispersing population size is not only reasonable ethnographically [42] and archaeologically [43, 23], but also leads to substantially different predictions (the value of η that agrees best with the data increases by more than a factor of 3).

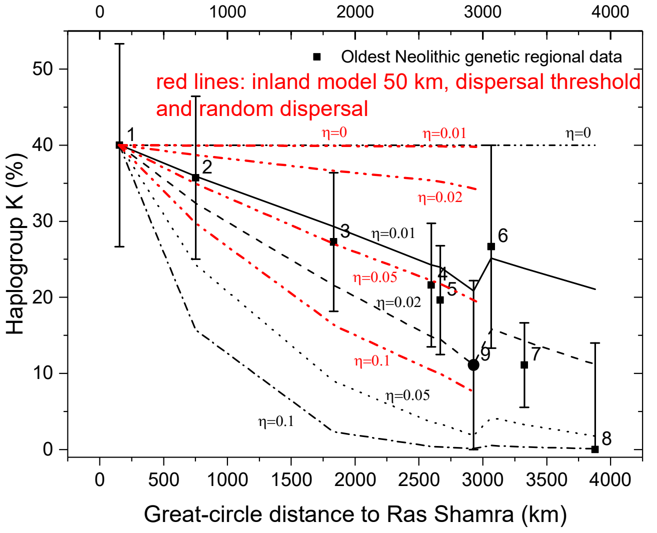


Figure 5. Observed and simulated Neolithic percentages of haplogroup K. Red lines are simulation results using the homogeneous geography (inland model) *with* ***integer numbers, dispersal threshold and random dispersal*, i.e. the model used in the main paper**. Black lines were obtained in Ref. [3] using a non-homogeneous geography with real numbers. The simulated wave of advance has not arrived yet to regions 6-8 (not only to region 8 as in Supplementary Figures 2a and 3) because the dispersal threshold leads to a slower front (this is solved in the main paper and Secs. S3-S5 by using a model with longer jumps for the sea route implied by Archaeology [38] for those regions).

S3. Inland routes and sea routes. Delayed and advanced regions

As mentioned in the caption to Suppl. Fig. 2a, when comparing to *genetic* data (Figs. 3-4 in the main paper and Supppl. Figs. 3-5) the results of the simulations are computed at the average date of the individuals in the considered region whose mtDNA haplogroup has been determined. In contrast, when comparing to *archaeological* data (e.g., Fig. 1b in the main paper) we need a simulated date that can be compared to that of the oldest Neolithic site in the region considered. For this reason, the arrival time of the Neolithic to each region (main paper, lines in Fig. 1b) has been estimated from the simulations as that when the population density of farmers reaches 10% of their saturation density $p_{F\,max}$. This value 10% has been applied previously to compare to archaeological data [3, 38] because it is unlikely that the archaeological record corresponds to the earliest farmers per region. However, any other reasonable value of this percentage (instead of 10%) would lead to essentially the same results.[15]

It is well known from Archaeology that the Neolithic spread across Europe following two main routes: one propagated inland (Balkans and Central Europe) and the other one along the sea (Northern Mediterranean coast) [23], the latter being substantially faster [38, 39] (see Fig. 1b in our main paper). Accordingly, we run our simulation program in a homogeneous geography (Sec. S2-A) by taking the existence of these two routes into account, as follows.

S3-A. Inland route

In order to identify each geographical location with a point in our simulation grid, we need to transform latitude and longitude into rectangular (X and Y) coordinates. Some previous studies have used specific projections for this purpose. For example, an Albers conic equal-area projection was used in Ref. [35], which has the property that the area is conserved, i.e., equal areas on the rectangular grid correspond quite closely to equal areas on the surface of the Earth. Note, however, that for regions reached following the inland route, we use great-circle distances (not areas) to compare the results of the simulation to archaeological and genetic data (Figs. 1a and 3a in the main paper, respectively). For this reason, here we prefer to look for a transformation of latitude and longitude into rectangular (X and Y) coordinates such that the distance is conserved, i.e. that the distance between two points on the rectangular grid is the same as the great-circle distance between their corresponding latitudes and longitudes. But there is not any projection satisfying this condition exactly for all possible pairs of points [44]. For example, in the sinusoidal projection, only distances along the equator and other parallels are conserved [44]. Therefore, the following approach is more appropriate for our purposes.

As explained at the beginning of Sec. S2, point (i), the origin of our simulations (central node in a grid of squared cells) corresponds to the site of Abu Hureyra (Syria). Let $\psi_o = 38.40^{\text{o}}\text{E}$ and $\varphi_o = 35.87\text{N}$ stand for its longitude and latitude, respectively. Let ψ and φ stand for the longitude and latitude of an arbitrary location. We determine the number of nodes separating this point (ψ, φ) on the simulation grid from the origin (ψ_o, φ_o) along the horizontal (X) and vertical (Y) directions as follows. The *curved* distance along a meridian (i.e., a circle on the Earth surface with constant longitude) between the two latitudes is

$$\Delta Y = R(\varphi - \varphi_o), \qquad \text{(S2)}$$

where the angles are measured in radians. On our *flat* simulation grid, this is the height of a rectangle such that the points point (ψ, φ) and (ψ_o, φ_o) correspond to two corners on its diagonal. Therefore the number of nodes of a vertical side of this rectangle (on our flat simulation grid) is

$$N_Y = \Delta Y / 50\ \text{km}, \qquad \text{(S3)}$$

[15] For the saturation density of farmers used in our simulations ($p_{F\,max} = 1.28$ farmers/km^2), this value 10% corresponds to $1.28\frac{\text{farmers}}{\text{km}^2} \cdot 50^2 \cdot \frac{1}{10}$ =320 farmers per cell in our inland simulation grid, which is higher (as it should) than the threshold or minum of $0.06\frac{\text{farmers}}{\text{km}^2} \cdot 50^2$ =150 farmers that can arrive to a cell (Sec. S2-C). Similarly for our coastal simulation grid (Sec. S3-B) the 10% of $p_{F\,max}$ gives $1.28\frac{\text{farmers}}{\text{km}^2} \cdot 70^2 \cdot \frac{1}{10}$ =627 farmers per cell, which is again consistently above the corresponding threshold, i.e., $0.06\frac{\text{farmers}}{\text{km}^2} \cdot 70^2 = 294$ farmers per cell.

because the cell size (distance between two neighboring nodes) corresponds to the mean dispersal distance per generation of pre-industrial farmers, which is 50 km according to ethnographic data [40].

For the X direction the calculation is less straightforward. We begin by noting that all points on the surface of the Earth with the same latitude φ are located on a plane, such that its intersection with the surface of the Earth is a circle (called a parallel) with radius $R\cos\varphi$, where R is the radius of the Earth. Therefore, the curved distance along the parallel at latitude φ between the two longitudes is

$$\Delta X_{\varphi} = R\cos\varphi \;(\psi - \psi_o), \tag{S4}$$

and similarly the curved distance along the parallel at latitude φ_o between the two longitudes is

$$\Delta X_0 = R\cos\varphi_o\,(\psi - \psi_o), \tag{S5}$$

where ψ and ψ_o are measured in radians. Thus on the Earth surface, the *curved* area defined by the parallels and meridians crossing locations (ψ, φ) and (ψ_o, φ_o), which corresponds to the rectangle on our *flat* simulation grid, has constant height (Eq. (S2)) but variable width (Eqs. (S4)-(S5)). A reasonable solution seems to consider the average width on the Earth surface for the width of the rectangle on the *flat* simulation grid, i.e.

$$\Delta X = \frac{\Delta X_{\varphi} + \Delta X_0}{2}, \tag{S6}$$

and therefore the number of nodes of the horizontal side of the rectangle on the *flat* simulation grid is

$$N_X = \Delta X / 50 \text{ km}. \tag{S7}$$

We have found that using this simple approximation, the great-circle distance is reasonably conserved. For example, consider the point in Sweden with longitude $\psi = 14.21$ºE and latitude $\varphi = 57.60$ºN. Using Eqs. (S2)-(S7) and the radius of the Earth ($R = 6{,}371$ km) we find that ΔY=2,416 km, $N_Y = 48$ nodes, $\Delta X_{\varphi} = 1{,}453$ km, ΔX_0 =2,179 km, $\Delta X = 1{,}816$ km and $N_X = 36$ nodes. This implies that in the simulation rectangular grid, the distance between this point (located in Sweden) and the origin (Abu Hureyra) is $\sqrt{2{,}416^2 + 1{,}816^{\,2}} = 3{,}022$ km (note that this equation would not hold for a curved grid). On the other hand, the great-circle distance between both points on the Earth surface can be computed by using the Haversine equation [45],

$$d = 2\,R\,\sin^{-1}\left(\sqrt{\sin^2\left(\frac{\varphi - \varphi_0}{2}\right) + \sin^2\left(\frac{\psi - \psi_0}{2}\right)\,\cos\varphi\,\cos\varphi_0}\right) \tag{S8}$$

and yields $d = 3{,}005$ km, very close to the value of 3,022 km on our *flat* simulation grid. The absolute error is only 17 km, which is less than the cell side, i.e. 50 km, and the relative error is only $\frac{17\text{ km}}{3{,}005\text{ km}} \cdot 100 = 0.6\%$, i.e., below 1%. We have also checked that there is reasonable agreement for all other locations at which we have compared simulations to genetic and archaeological observations (Suppl. Tables 2 and 4, respectively). We think that this is a powerful reason for using the method above to determine the number of X and Y nodes in the simulation grid between the origin and an arbitrary location.[16]

[16] For completeness we mention that alternative approaches based on *curved* simulation grids have substantial problems. The reason is that, as seen above, if we consider an area *A* on the Earth surface with sides on the parallels and meridians defined by two locations (ψ_o, φ_o) and (ψ, φ), then the horizontal size (i.e., the width) of *A* decreases with increasing latitude. This implies two options if considering a curved grid: (i) the number of nodes of *A* depends on latitude (in the example above, from 44 nodes at latitude φ_o to only 29 nodes at latitude φ), but this is complicated to implement because the X-coordinate along a meridian would not be constant; (ii) alternatively, we could require a uniform number of nodes and the distance between two neighbouring nodes would be variable. In the example above, assuming the value 50 km at latitude φ_o it would be only 33 km at latitude φ, but this would deviate substantially from the average value of 50 km implied by the ethnographic data [40]. For these reasons, we think that it is simpler and very reasonable to follow the approach explained above (which is based on using a *flat* simulation grid).

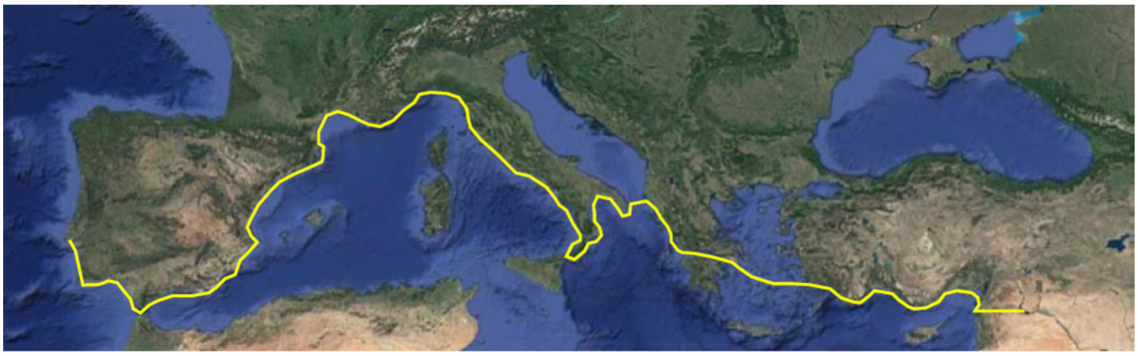
S3-B. Sea route

In Fig. 1b (main paper), the model for the inland route (red line) arrives clearly too late to regions 5-8 (this is also seen in Suppl. Figs. 2a and 3, because there the simulated front has not yet arrived to Portugal at the average time of its genetic data). This is not surprising, because archaeological data and mathematical models have shown that these regions were reached by sea travel along the coast, with longer jumps per generation [46, 38, 39] than those of inland travel (the latter have lengths of about 50 km per generation according to ethnographic data for pre-industrial farmers [40]). As explained in the main paper, we develop a realistic sea model simply by running our homogeneous model with longer jumps, i.e. with square cells of longer sides (namely 70 km, which is the value necessary to attain agreement between simulations and data in Fig. 1b in the main paper [17]). However we must take into account that along a coast, the Neolithic wave of advance does not propagate in homogeneous space but following an irregular path, such as that shown in Suppl. Fig. 6a.[18]

For sea travel (Suppl. Fig. 6a) we cannot use Eqs. (S2)-(S7) because they are applied in simulations that are valid assuming that individuals can settle on any node of a two-dimensional grid. But farmers cannot settle on the sea, so for sea travel we will follow an approach motivated by a method that was introduced in Ref. [47] and called short-path distances. In our approach, distances are measured along the coast (estimated by the procedure explained in Suppl. Fig. 6a). In our simulations for regions reached by the sea route, the population in a coast cell (which corresponds to a side of our rectangular grid) can disperse to three neighboring cells (not four as in the case of inland cells), because the other cell is located on the sea. In these simulations the origin of the Neolithic wave of advance (i.e., the cell with an initial population of farmers) is a coast cell, i.e., it is located on one side of the rectangular grid. We determine the arrival time of the wave of advance to each location by considering a cell located along this side of the grid, and at the same distance from the origin as the distance to the considered location along the sea route (yellow line in Suppl. Fig. 6a) obtained using the free internet application sea-seek.com (https://www.sea-seek.com/tools/tools.php). Such a procedure has been recently applied to modelling and comparing to archaeological data [39] and is here extended to genetic data, i.e., we compute not only the arrival time at each region but also its percentage of haplogroup K (the latter is not estimated at the arrival time but at the average date of the genetic data available for each region).

[17] A previous model using a real geography (non-homogeneous model) and older archaeological databases found that the best agreement is attained for a very similar jump distance per generation, namely about 100 km (Suppl. Fig. 9 in Ref. [3]) instead of 70 km (Fig. 1b in our main paper). However, the results reported in Ref. [3] were based on jumps of 150 km because using 100 km it was observed that the front entered Italy from the North (Suppl. Fig. 10a in Ref. [3]) instead of from Albania, as observed archaeologically (Fig. 6 in Ref. [35]), despite the sea distance from Albania to Italy is only about 70 km. The reason was that one cell of the simulation grid was located on the sea between Albania and southern Italy (see Text S6 in [3]). This problem does not arise here because we use a homogeneous grid. This is an advantage of using a homogeneous model, in addition to the other two advantages mentioned in Sec. S2, point (iii). In our opinion, it is of interest to use homogeneous and non-homogeneous models in different papers and compare their results, advantages and drawbacks.

[18] It is worth to note that spatial interpolations of archaeological dates have shown previously [35] that the Neolithic front jumped from Albania to south-eastern Italy, which implies that sea travel took place with distances of at least 70 km. This jump is taken into account in Suppl. Fig. 6a (yellow line). Its distance is also consistent with the jumps of 70 km per generation used in our simulations. Such a capability for sea travel by early Neolithic farmers also follows from the fact that the Neolithic reached the island of Cyprus, which is separated by a similar distance from the continent.



Supplementary Figure 6a. An example (yellow line) of route used to estimate the distances ***along the coast*** for the regions reached mainly by sea (namely regions 5-9 and 13 in Figs. 1b and 3b in the main paper). The sea-seek.com free internet application (https://www.sea-seek.com/tools/tools.php) has been used to measure the distance of this route and google maps (www.google.es/maps/) has been used to visualize it. The route (yellow line) begins at Abu Hureyra (the oldest PPNB site in Syria [47], lower right in this map), which is the origin of the Neolithic spread in our simulations in the main paper. In this example, the end of the yellow line (lower left) corresponds to the location of the oldest Neolithic site in central Portugal (see Suppl. Table 4).

As explained in the main paper (Methods), we have used for the carrying capacities of farmers and HGs the values $p_{F\,max}$= 1.28 farmers/km^2 and $p_{HG\,max}$= 0.064 HGs/km^2, which imply that the maximum populations sizes per cell in the inland simulations (red line in main paper, Fig. 1b, cells with sides of 50 km) are $P_{F\,max} = 1.28\text{ farmers/km}^2 \cdot (50\text{ km})^2 = 3{,}200$ farmers/cell and $P_{HG\,max} = 0.064\text{ individuals/km}^2 \cdot (50\text{ km})^2 = 160$ individuals/cell. Analogously, since the cells of our sea-route simulations have sides of 70 km (blue line in main paper, Fig. 1b), we have taken into account that the maximum number of farmers per cell in these simulations are $P_{F\,max} = 1.28\text{ farmers/km}^2 \cdot (70\text{ km})^2 = 6{,}272$ farmers/cell and $P_{HG\,max} = 0.064\text{ HGs/km}^2 \cdot (70\text{ km})^2 = 314$ HGs/cell.

We have applied the dispersal threshold in Sec. S2-C (namely 0.06 farmers/km^2) to the following two models or simulation grids.

Firstly, in the simulation grid of the inland-route model (cells of 50 km x 50 km) there cannot be less than $0.06\text{ farmers/km}^2 \cdot (50\text{ km})^2 = 150$ farmers/cell. Therefore, for cells not located at the coast, farmers do not jump from any cell unless it has at least 970 farmers, due to the fact that in this way the number that jump in each direction is at least $NINT\big((1-0.38)\cdot 970/4\big) = 150$ farmers. Strictly, for cells located at the coast the threshold would be 725 farmers (not 970) because $NINT\big((1-0.38)\cdot 725/3\big) = 150$ farmers, but in practice this change (for coastal cells) is not necessary since it would not affect our results (due to the fact that this inland-route grid is applied only to regions located along the inland route, which are very distant from coastal cells).

Secondly, for the simulation grid of the sea-route model (cells of 70 km x 70 km) we apply that there cannot be less than $0.06\text{ farmers/km}^2 \cdot (70\text{ km})^2 = 294$ farmers/cell, and we have two cases. First, for cells not located at the coast, farmers do not jump from any cell unless it has at least 1,897 farmers, due to the fact that in this way the number that jump in each direction is at least $NINT\big((1-0.38)\cdot 1{,}897/4\big) = 294$ farmers. Second, for cells located at the coast, farmers do not jump from any cell unless it has at least 1,423 farmers, because $NINT\big((1-0.38)\cdot 1{,}423/3\big) = 294$ farmers.

In Fig. 1b in the main paper, for the *blue* error bars (regions reached by sea travel) the distances (horizontal axis) have been computed by using the procedure exemplified in Suppl. Fig. 6a[19]. For each of those regions, the final location of the sea path (Suppl. Fig. 6a, yellow line) is its oldest Neolithic site[20].

The procedure to estimate distances along the coast exemplified in Suppl. Fig. 6a has been also used in Fig. 3b in the main paper, but these distances are slightly different from those in Fig. 1b because in Fig. 3b for each region we have used the average location of the Neolithic individuals whose mtDNA has been measured (Suppl. Tables 1-2), not its oldest archaeological site as in Fig. 1b (Suppl. Table 4). The reason is that in Fig. 3b we compare to genetic data (squares and error bars, from Suppl. Table 3), not to archaeological data as in Fig. 1b.

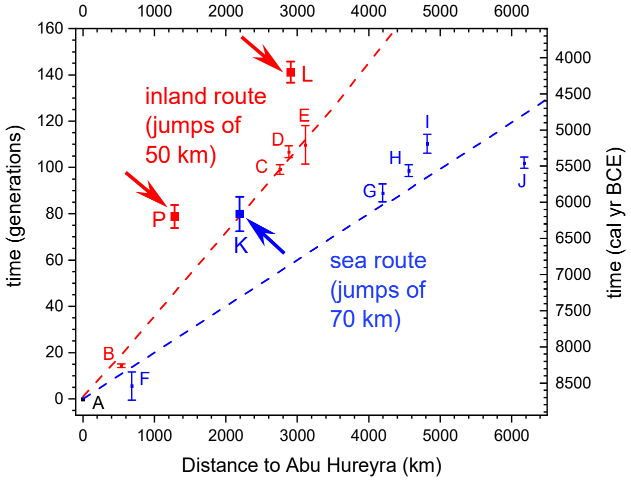
S3-C. Delayed and advanced regions

We have not included some regions, e.g., K (Italy), L (Sweden) and P (Bulgaria) in Fig. 1b (main paper) because they were affected by delays. To see this, Fig. 1b in the main paper is repeated in Suppl. Fig. 6b by adding the date of the oldest Neolithic site in regions K, L and P (large squares with error bars and arrows). We see that each of these 3 error bars is substantially more recent (by about 1,000 yr) than expected from the simulations (line of the same color). In fact, these delays are well-known in the archaeological literature. For region L (Sweden) it has been shown that there was a rapid warming at about 4,000 cal. yr BCE, simultaneous with the arrival of farming and a rise in population density in southern Scandinavia. Thus a possible explanation for this delay is that the observed warming could have moved the limit of cereal suitability to the North (see p. 164 in [23]). Similarly, for Bulgaria (region P) the delay has been attributed to a climatic event of cold and wet conditions that could have created a barrier to the spread of farming north of the Aegean zone (p. 78 in [23]). The delay in region K (Italy) has been also noted and discussed in the archaeological literature (p. 108 in [23]). Additional delays have been also detected, e.g., in Greece and between central and western Anatolia (p. 64 in [23]). Independently of the explanations, we think that it is not worth to complicate the model used in Fig. 1b by simulating all of such delays because, for our purposes, their only effect in these regions would be to yield a slightly different value for the %K of the pioneering populations of farmers, due to the evolution of the cline of %K during about 1,000 yr. But Suppl. Fig. 6c shows that during the first 1,000 yr after the arrival of the first farmers, the change in the %K is very small (less than 1% K), so a substantially more complicated model would lead to much the same results and conclusions concerning the cline of %K (Figs. 3a-b in the main paper).[21]

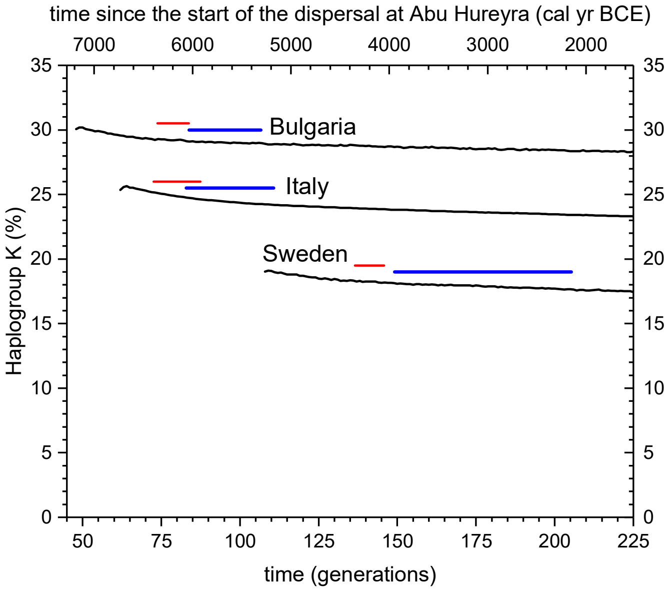
[19] A percentage of the path from Abu Hureyra is necessarily inland (especially for the regions Cyprus and Spain Navarre) but most of it is on the sea (even in those two cases), so we have applied the sea model (Suppl. Fig. 6a). We have assumed that Cyprus was reached from the North (closest mainland location, at about 70 km) for consistency with the sea travel distance used in the simulations (70 km). If it were reached from the East, the distance travelled by sea would be substantially larger. Then its error bar in Fig. 1b would be moved slightly to the left, so the results and conclusions would not change.

[20] A difference between an early model [35] and ours is that the former did not compare to the oldest site per region but with all early Neolithic sites, which is less realistic for the purpose of comparing to the arrival date computed by the simulations. Another difference is that Ref. [35] assumed jumps with several distances along the coast (but it has been argued that using a single distance is more realistic for the western Mediterranean, because it can lead to multiple entrances [38]). In any case, we admit that other models are possible, including substantially more complicated ones (e.g., with a longer jump distance for the West [38] than for the East Mediterranean and/or other regional features), but such complications are unlikely to change our conclusions, because the latter follow from an observed genetic cline at the continental level (not in a specific region such as the West Mediterranean).

[21] Many reasonable models can be envisaged. We have assumed the simplest possible one to obtain a realistic estimation the percentage of farmers involved in interbreeding along the inland and sea routes. In our model, the spread rate is constant (but different inland than along the coast). Thus, it is unavoidable that the Neolithic arrived somewhat earlier (according to archaeological data) than expected from the simulations for some regions, and somewhat later in others. An example of advanced region is Portugal (error bar J in Fig. 1b or Suppl. Fig. 6b), somewhat earlier than according to the simulations (blue line). In this case, we cannot compute the %K at the mean date of the early farmers whose mt DNA haplogroup is known



Supplementary Figure 6b. This figure is the same as Fig. 1b in the main paper but including also 3 large squares with arrows for regions K (Italy), L (Sweden) and P (Bulgaria). They have been omitted in Fig. 1b in the main paper because in those regions the arrival of the Neolithic was delayed. Lines are the Neolithic arrival times from the simulations, and error bars are the oldest Neolithic site per region. Blue color corresponds to the sea route (jumps of 70 km on the simulation grid) and red color to the inland route (jumps of 50 km on the simulation



Supplementary Figure 6c. Evolution of the percentage of farmers with haplogroup K in three delayed regions according to the simulations (black curves). The inland model (red line in Suppl. Fig. 6b) has been used for Bulgaria and Sweden, and the sea model (blue line in Suppl. Fig. 6b) has been applied for Italy. In all cases $\eta = 0.06$. Each curve begins when the first farmers arrive at the average location of farmers with known mt DNA in the region considered (Suppl. Table S1), not to its oldest site (so direct comparison of this initial date to Suppl. Fig. 6b is not possible). The conclusion from this figure is that the %K changes very slowly (by less than 1% K) during 1,000 years after each arrival. For each region, the horizontal blue line gives the range of the dates of farmers whose mt DNA has been determined (Suppl. Table 1) and each red line gives the radiocarbon range of the oldest site (so the red lines here corresponds to the error bar in Suppl. Fig. 6b). The red lines begin later than the black one (simulations) for each of these three regions. In this sense, the arrival of the Neolithic to them was delayed (see also Suppl. Fig. 6b). Note that the location of the oldest site per region (red line) is not exactly the same as the average location of farmers with known mt DNA in that region (blue line). Also, each farmer included to calculate the blue line has a slightly different location. In spite of this, these lines are illustrative of regional trends.

because the simulated front has not arrived yet, so we have computed it later (at 130 generations) to obtain the corresponding values of the %K (blue lines in Fig. 3b, at the distance corresponding to error bar 16). We think this is a reasonable approximation, because the %K is rather stable during hundreds of years (see Suppl. Fig. 6c).

S4. Effect of the initial genetic conditions

In this section (and in the main paper) we use the model with integer population sizes, dispersal threshold and random dispersal (Sec. S2-D and Suppl. Fig. 5). In contrast with Sec. S2, here our aim is not to compare to Ref. [3], so we will no longer use Ras Shamra as origin (as in Ref. [3]) but Abu Hureyra (as in our main paper). For the same reason, here we will not use the observed percentages and error bars for haplogroup K in Ref. [3] but those updated with more recent data (Suppl. Tables 1 and 3) and used in our main paper.

In this section we have applied, as in Fig. 3 in the main paper, the parameter values $p_{F\,min} = 0.06$ farmers/km^2, $R_{0,F} = 2.45$, $p_{F\,max}$ =1.28 farmers/km^2 and $p_{HG\,max} = 0.064$ HGs/km^2.

From the haplogroups of early Neolithic individuals in region 1 (Suppl. Table 1), the observed percentage of farmers with haplogroup K (%K) is 47.4%K. However, this percentage has a substantial uncertainty due to the small number of individuals available in region 1 (northern Mesopotamia). As reported in Suppl. Table 3, its lower bound is 31.6%K and its upper bound is 63.2%K (both at confidence level 80% and obtained using bootstrap resampling). In order to analyze the effect of this uncertainty on our results, in Suppl. Fig. 7a (arrow) the initial fraction %K has been set to its lower bound (31.6%K) and in Suppl. Fig. 7b (arrow) to its upper bound (63.2%K). The lines are the simulation results. Each square and error bar is the observed %K in the region considered, as used already in Fig. 3 in the main paper and given in Suppl. Table 3.

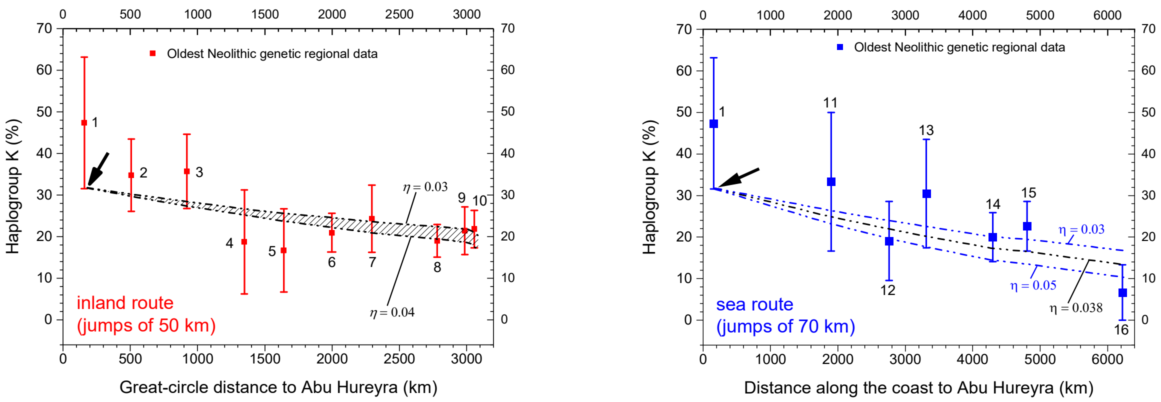
First we consider the lower bound (31.6%K) of the percentage of haplogroup K in region 1 (arrow in Suppl. Figs. 7a). As explained in the caption to Suppl. Fig. 7a, along the inland route (left) the consistency range is $0.03 \leq \eta \leq 0.04$, whereas along the sea route (right) consistency is attained only for $\eta = 0.038$. Thus both ranges overlap, and the common value of η implied by both routes is $\eta \approx 0.038$ assuming that the initial percentage of haplogroup K in Syria was 31.6%K. Note the difference with the corresponding value ($\eta = 0.07$) obtained if assuming that the initial percentage is equal to the observed value, i.e. 47.4%K (main paper, Fig. 3).

Next we consider the upper bound (63.2%K) of the percentage of haplogroup K in Syria (arrow in Suppl. Figs. 7b). As explained in the caption to Suppl. Fig. 7b, along the inland route (left) the consistency range is $0.10 \leq \eta \leq 0.12$ and along the sea route (right), the consistency range is $0.07 \leq \eta \leq 0.10$. Thus the ranges of η implied by both routes are compatible, and their common value is $\eta \approx 0.10$ assuming that the initial percentage of haplogroup K in Syria was 63.2%K. Again, note the difference with the corresponding value ($\eta = 0.07$) obtained if assuming that the initial percentage is equal to the observed value, i.e. 47.4%K (main paper, Fig. 3).

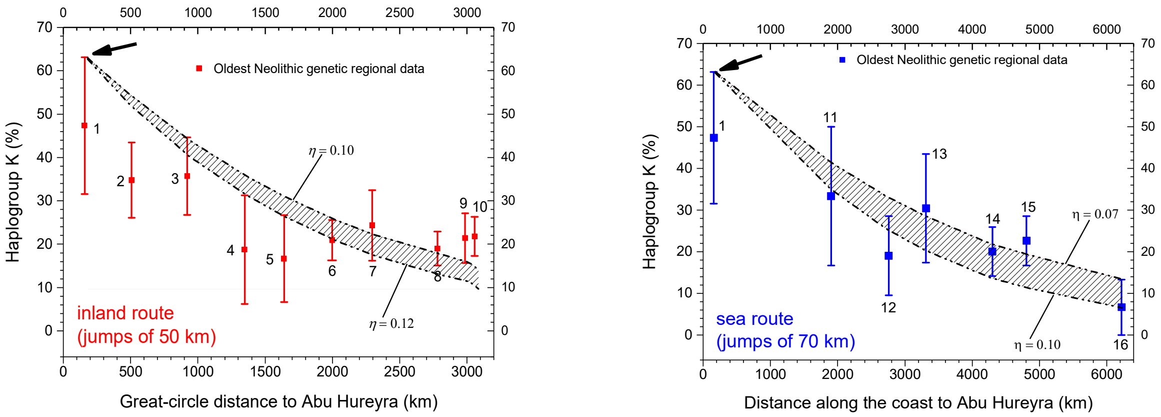
From the previous two paragraphs we conclude that if the initial genetic conditions are changed (within the range implied by the data), the range of η changes consistently along both routes, in such a way that: (1) both routes always yield similar ranges of η; and (2) both routes are always consistent with a single range of η. In our main paper, Fig. 3 assumes that the initial percentage of haplogroup K in Syria was 47.4%K (which corresponds to square 1 in Suppl. Figs. 7a-b). For that initial percentage of haplogroup K (47.4%K), Fig. 3 in the main paper leads to the conclusion that the interaction between farmers and HGs along both routes was very similar ($\eta \approx 0.07$). The results in the previous two paragraphs imply that this conclusion is maintained, and refined by the range $0.03 \leq \eta \leq 0.10$.

It is worth to note that in in the future the mt haplogroup of additional early farmers in northern Mesopotamia is determined, it may be possible to estimate a narrower confidence interval for the frequency of theses farmers than that implied by current data (31.6%K−63.2%K) and, therefore, it may be also possible to obtain a more precise range for the percentage of early farmers that interbred with HGs or acculturated them.

An analysis of the effect of the uncertainty in the initial %K taking into account also the uncertainty in the parameter values is included in Secs. S6-B and S6-C.



Supplementary Figure 7a. Observed and simulated Neolithic percentages of haplogroup K (%K) among early farmers. These plots are the same as those in Fig. 3 in the main paper, but here simulations have been performed by setting the initial value in northern Mesopotamia (region 1) to its lower bound (31.6%K, arrow) rather than to its mean value (47.4%K, square 1). Error bars (Suppl. Table 3) are the same as in Fig. 3 in the main paper. Simulations begin at the same date as in the main paper. Left: the clines for $\eta = 0.03$ and $\eta = 0.04$ cross all error bars. Right: the cline for $\eta = 0.038$ crosses all error bars.



Supplementary Figure 7b. Observed and simulated Neolithic percentages of haplogroup K (%K) among early farmers. These plots are the same as those in Fig. 3 in the main paper, but here simulations have been performed by setting the initial value in northern Mesopotamia (region 1) to its upper bound (63.2%K, arrow) rather than to its mean value (47.4%K, square 1). Error bars (Supplementary Table 3) are the same as in Fig. 3 in the main paper. Simulations begin at the same date as in the main paper. Left: for $\eta = 0.10$ and $\eta = 0.12$ the clines cross all error bars except four of them. Right: the clines for $\eta = 0.07$ and $\eta = 0.10$ cross all error bars except one.

S5. Sensitivity analysis of the cline of haplogroup K

In this section we report the results of a one-at-a-time sensitivity analysis in the sense that, for each parameter tested, we run the model first using its lowest value and then its higher value, while keeping each of the other parameters at its intermediate value (i.e., that used in Fig. 3 and Suppl. Figs. 7a and 7b). The parameters varied are the dispersal threshold ($P_{F\,min}$), the net fecundity of farmers ($R_{0,F}$), the carrying capacity of farmers ($p_{F\,max}$) and the carrying capacity of hunter-gatherers ($p_{HG\,max}$). The generation time and persistence have a negligible effect, as long as we use realistic values ethnographically [48, 40]. Similarly the dispersal distance is tightly constrained by the archaeological data (Fig. 1b). The one-at-a-time analysis in this section will be useful to develop a simultaneous sensitivity analysis of all of these parameters (Sec. S6).

We use again the same model as in the main paper, i.e., we apply integer numbers, a dispersal threshold and random dispersal (Sec. S2-D and Suppl. Fig. 5) for the inland and sea routes (homogeneous geography with cells of 50 km and 70 km, respectively).

S5-A. Effect of the dispersal threshold ($p_{F\,min}$)

As explained in Sec. S2-C above, archaeological data have been used to estimate the minimum population density of Early Neolithic Banderkeramik farmers and the result is $p_{F\,min}$ = 0.06 ± 0.01 individuals/km^2. The mean value of this range (i.e., 0.06 individuals/km^2) has been used to obtain Fig. 3 in the main paper. In this subsection we will use its lower and upper bounds (0.05 individuals/km^2 and 0.07 individuals/km^2, respectively) to determine if our results change appreciably or not.

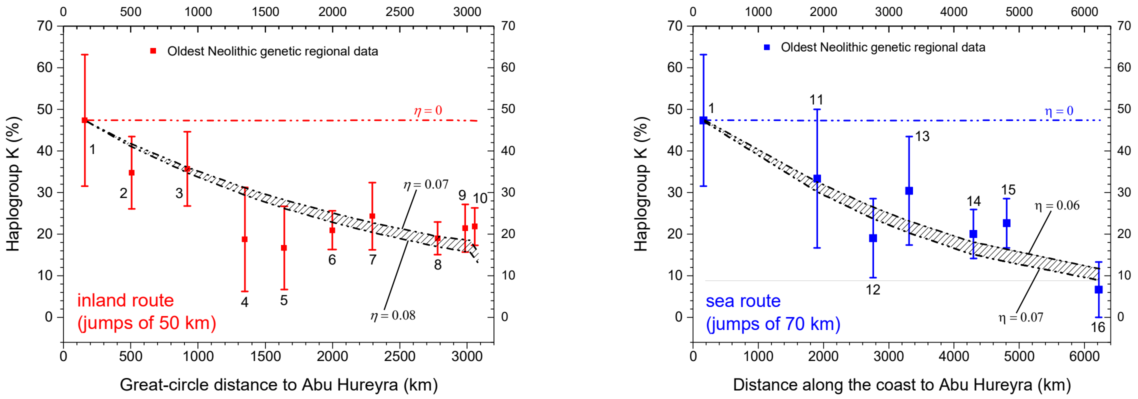
Firstly we repeat the calculations in Secs. S2-C (inland route) and S3-B (sea route) but now using the lower bound (Suppl. Fig. 8a). Then in the simulation grid of the inland-route model (cells of 50 km x 50 km) there cannot be less than $0.05\ \text{farmers/km}^2 \cdot (50\ \text{km})^2 = 125$ farmers/cell (i.e., the first group of arriving farmers must include at least 125 individuals). Thus, for cells not located at the coast, farmers do not jump from any cell unless it has at least 806 farmers, due to the fact that in this way the number that jump in each direction is at least $NINT\big((1-0.38)\cdot 806/4\big) = 125$ farmers. Thus, the lower bound of the dispersal threshold is 806 farmers (as mentioned in Sec. S3-B, the threshold for coastal cells has no effect on the results along the inland route). This value has been used to obtain the left plot in Suppl. Fig. 8a. Similarly, for the simulation grid of the sea-route model (cells of 70 km x 70 km) we apply that there cannot be less than $0.05\ \text{farmers/km}^2 \cdot (70\ \text{km})^2 = 245$ farmers/cell, and we have two cases. First, for cells not located at the coast, farmers do not jump from any cell unless it has at least 1,581 farmers, because in this way the number that jump in each direction is at least $NINT\big((1-0.38)\cdot 1{,}581/4\big) = 245$ farmers. Second, for cells located at the coast, farmers do not jump from any cell unless it has at least 1,185 farmers, because $NINT\big((1-0.38)\cdot 1{,}185/3\big) = 245$ farmers. These values have been used to obtain the right plot in Suppl. Fig. 8a.

Secondly we repeat the calculations in Secs. S2-C (inland route) and S3-B (sea route) but now using the upper bound (Suppl. Fig. 8b). Then in the simulation grid of the inland-route model (cells of 50 km x 50 km) there cannot be less than $0.07\ \text{farmers/km}^2 \cdot (50\ \text{km})^2 = 175$ farmers/cell (i.e., the first group of arriving farmers must include at least 175 individuals). Thus, for cells not located at the coast, farmers do not jump from any cell unless it has at least 1,129 farmers, due to the fact that in this way the number that jump in each direction is at least $NINT\big((1-0.38)\cdot 1{,}129/4\big) = 175$ farmers. Thus, the lower bound of the dispersal threshold is 1,129 farmers (as mentioned in Sec. S3-B, the threshold for coastal cells has no effect on the results along the inland route). This value has been used to obtain the left plot in Suppl. Fig. 8b. Similarly, for the simulation grid of the sea-route model (cells of 70 km x 70 km) we apply that there cannot be less than $0.07\ \text{farmers/km}^2 \cdot (70\ \text{km})^2 = 343$ farmers/cell, and we have two cases. First, for cells not located at the coast, farmers do not jump from any cell unless it has at least 2,213 farmers, because in this way the number that jump in each direction is at least $NINT\big((1-0.38)\cdot 2{,}213/4\big) = 343$ farmers. Second, for cells located at the coast, farmers do not jump from any cell unless it has at least 1,660 farmers, because $NINT\big((1-0.38)\cdot 1{,}660/3\big) = 343$ farmers. These values have been used to obtain the right plot in Suppl. Fig. 8b.

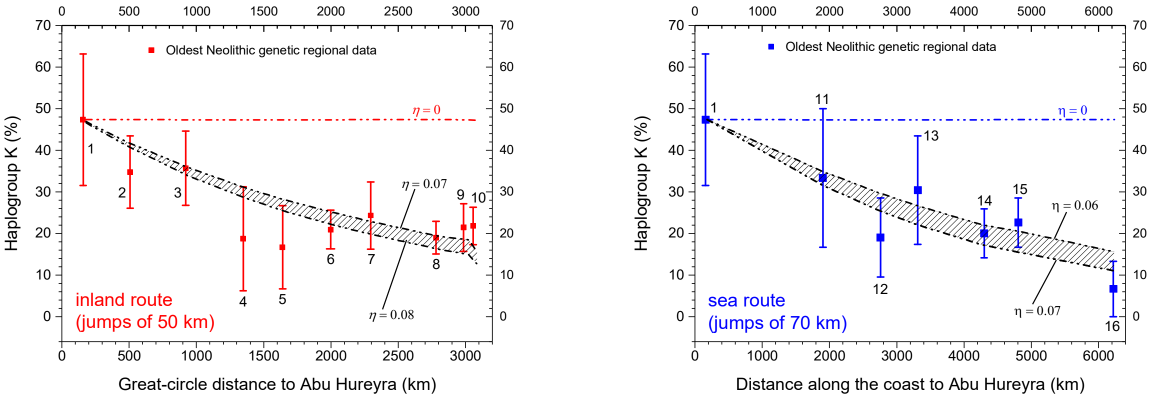
For the lower value of the dispersal threshold (caption of Suppl. Fig. 8a) and the inland route (left) the clines for $0.07 \leq \eta \leq 0.08$ agree with the data, whereas for the sea route (right) this agreement holds for $0.06 \leq \eta \leq 0.07$. So the value $\eta = 0.07$ agrees with both routes.

For the upper value of the dispersal threshold (caption of Suppl. Fig. 8b) and the inland route (left) the clines for $0.07 \leq \eta \leq 0.08$ agree with the data, whereas for the sea route (right) the same happens for $0.06 \leq \eta \leq 0.07$. So again the value $\eta = 0.07$ agrees with both routes. This confirms the main conclusion of our paper, i.e., that the interbreeding/acculturation behaviour of early farmers was the same along the inland and sea routes.

In Suppl. Figs. 8a-b we note that the effect of the dispersal threshold is small because it does not affect the values of η that agree best with the data. However, some changes in the cline can be seen along the sea route. This is reasonable because for sea cells the minimum number of farmers in a cell for dispersal to take place is substantially larger than for inland cells (see Sec. S3-B or paragraphs 2 and 3 in this Sec. S5-A).



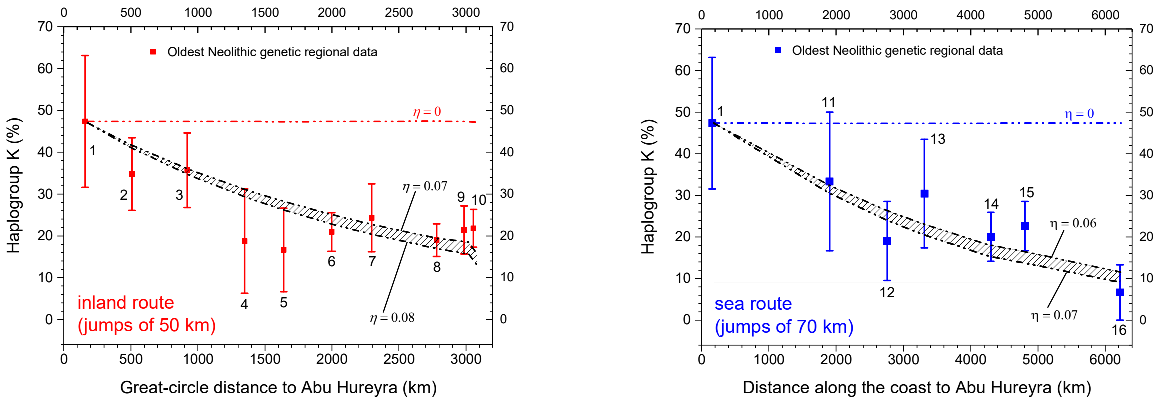
Supplementary Figure 8a. Observed and simulated Neolithic percentages of haplogroup K among early farmers. These plots are the same as those in Fig. 3 in the main paper, but here simulations have been performed by setting the dispersal threshold to its lower bound ($p_{F\,min}$ = 0.05 farmers/km^2) rather than to its mean value. Simulations begin at the same date as in the main paper. Left: the hatched area is defined by clines that cross all error bars but one. Right: the hatched are is defined in the same way.



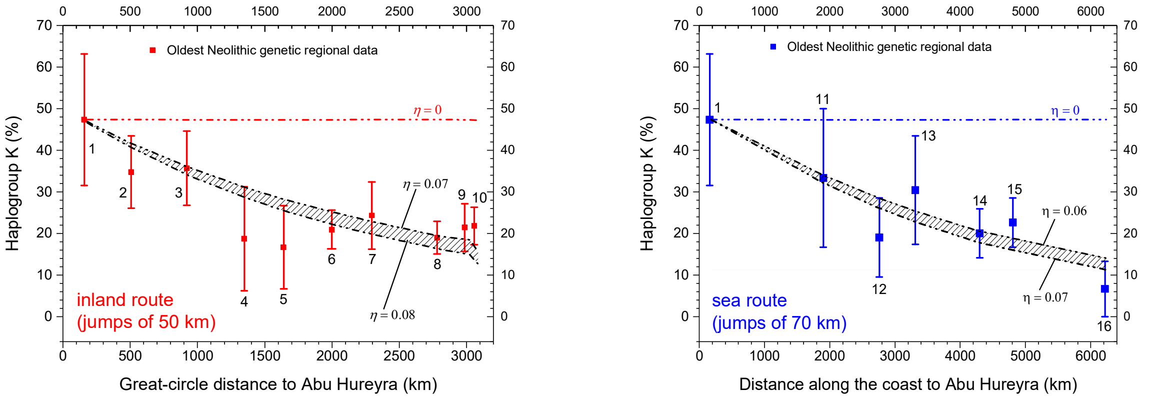
Supplementary Figure 8b. Observed and simulated Neolithic percentages of haplogroup K among early farmers. These plots are the same as those in Fig. 3 in the main paper, but here simulations have been performed by setting the dispersal threshold to its upper bound ($p_{F\,min}$ = 0.07 farmers/km^2) rather than to its mean value. Simulations begin at the same date as in the main paper. Left: the hatched area is defined by clines that cross all error bars but one. Right: the hatched are is defined in the same way.

S5-B. Effect of the net fecundity of farmers ($R_{0,F}$)

In the main paper we have applied for the net fecundity of farmers $R_{0,F} = 2.45$. This value has been obtained by applying that $R_{0,F} = e^{a_F T}$ (see Eq. (184) in Ref. [49]), where $T = 32$ yr is the mean age difference between a parent and one of his/her children [48] and we have used the mean value of $a_F = 0.028 \pm 0.005\, yr^{-1}$ (80% confidence-level interval), which is called the initial growth rate and has been estimated from ethnographic data [50]. The uncertainty in the value of T has a very small effect on the propagation of the front [48]. Thus, a lower bound of the net fecundity can be estimated by using the above equation with $a_F = 0.023\, yr^{-1}$ and $T = 32\, yr$, which yields $R_{0,F} = 2.09$. Similarly, an upper bound can be obtained from $a_F = 0.033\, yr^{-1}$ and $T = 32\, yr$, which yields $R_{0,F} = 2.87$. We plot the observed and simulated percentages of haplogroup K versus distance for the lower bound of the net fecundity (Suppl. Fig. 9a) and for its upper bound (Suppl. Fig. 9b). For given values of η and distance, the simulated %K is higher for the upper bound of $R_{0,F}$ (Suppl. Fig. 9b) because only farmers can have haplogroup K. However Suppl. Figs. 9a-b are almost identical to Fig. 3 in the main paper, so the shape of the cline is essentially independent of the net fecundity of farmers $R_{0,F}$ (if realistic values for the latter are used). We also note that the same percentage of farmers interacted with HGs along both routes (because $\eta \approx 0.07$ for both of them), so our main conclusion is maintained.



Supplementary Figure 9a. Observed and simulated Neolithic percentages of haplogroup K among early farmers. These plots are the same as those in Fig. 3 in the main paper, but here simulations have been performed by setting the net fecundity of farmers to its lower bound ($R_{0,F} = 2.09$) rather than to its mean value. Simulations begin at the same date as in the main paper. Left: the hatched area is defined by clines that cross all error bars but one. Right: the hatched are is defined in the same way.

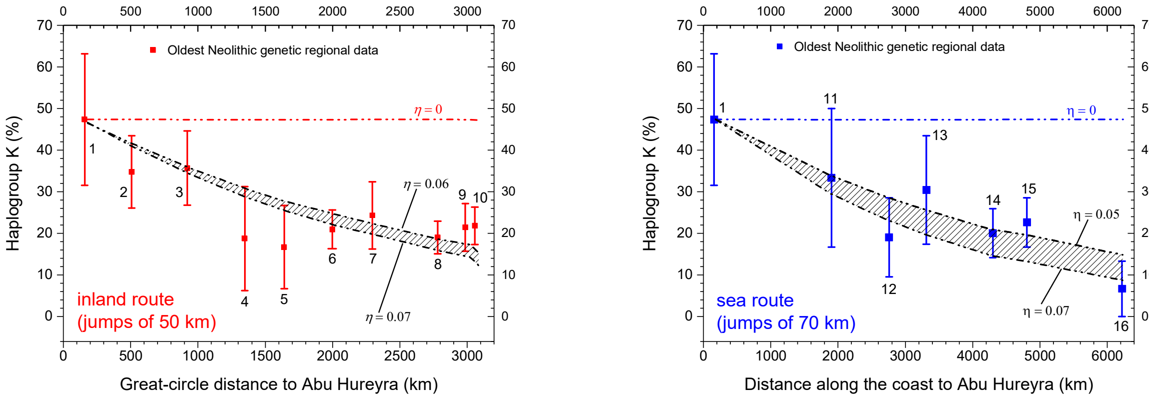


Supplementary Figure 9b. Observed and simulated Neolithic percentages of haplogroup K among early farmers. These plots are the same as those in Fig. 3 in the main paper, but here simulations have been performed by setting the net fecundity of farmers to its upper bound ($R_{0,F} = 2.87$) rather than to its mean value. Simulations begin at the same date as in the main paper. Left: the cline for $\eta = 0.07$ crosses all error bars but one, and that for $\eta = 0.08$ crosses all error bars but two. Right: the hatched area is defined by clines that cross all error bars but one.

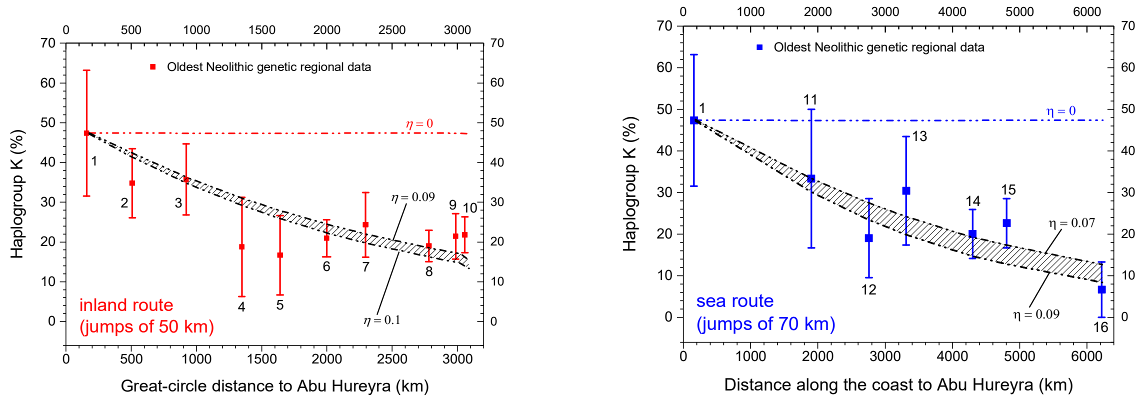
S5-C. Effect of the carrying capacity of farmers ($p_{F\,max}$)

In the main text and Sec. S3-B we have used $p_{F\,max} = 1.28$ individuals/km^2 [6, 35, 3]. This is an intermediate value within the range suggested by a statistical analysis of 9 archaeological estimations of Early Neolithic population densities in south-eastern Europe (6,500-5,500 BCE) and central Europe (5,500-4,300 BCE), namely $p_{F\,max} = 0.96 - 1.86$ individuals/km^2 [51]. Thus, we use for the lower bound $p_{F\,max}$= 0.96 individuals (Suppl. Fig. 10a) and for the upper bound $p_{F\,max}$= 1.86 individuals/ km^2 (Suppl. Fig. 10b). For given values of η and distance, the simulated %K is higher for the upper bound of $P_{F\,max}$ (Suppl. Fig. 10b) than for its lower bound (Suppl. Fig. 10a), as expected because there are more farmers and only them can have haplogroup K.

As explained in the caption to Suppl. Fig. 10a, for the inland route (left) the clines for $0.06 \leq \eta \leq 0.07$ agree with the data, and for the sea route (right) this agreement holds for $0.05 \leq \eta \leq 0.07$, so the ranges for both routes overlap ($0.06 \leq \eta \leq 0.07$). As explained in the caption to Suppl. Fig. 10b, for the inland route (left) the clines for $0.09 \leq \eta \leq 0.1$ agree with the data, and for the sea route (right) the same happens for $0.07 \leq \eta \leq 0.09$, so both routes are consistent with $\eta = 0.09$. This confirms our main conclusion, namely that the farmer-HG interaction was essentially the same along both routes.



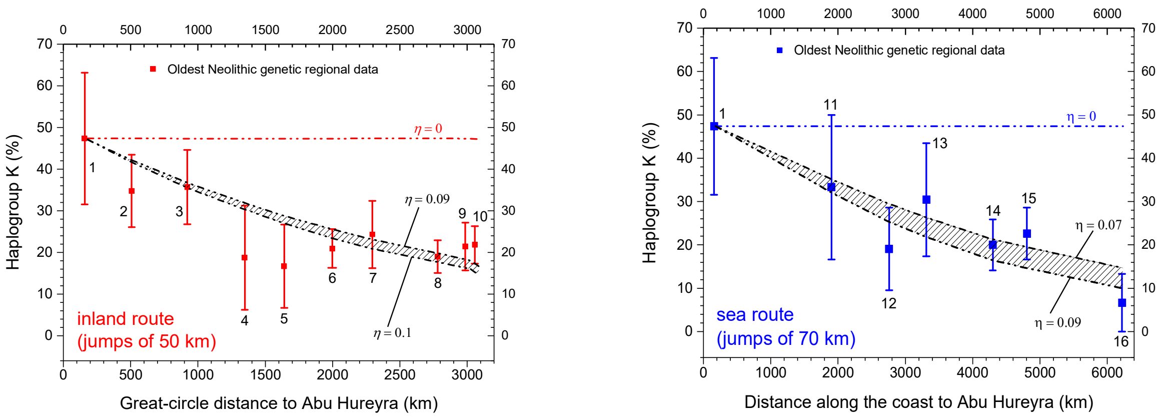
Supplementary Figure 10a. Observed and simulated Neolithic percentages of haplogroup K among early farmers. These plots are the same as those in Fig. 3 in the main paper, but here simulations have been performed by setting the carrying capacity of farmers to its lower bound ($p_{F\,max}$ = 0.96 individuals/km^2). Simulations begin at the same date as in the main paper. Left: the hatched area is defined by clines that cross all error bars but two. Right: the hatched area is defined by clines that cross all error bars but one.



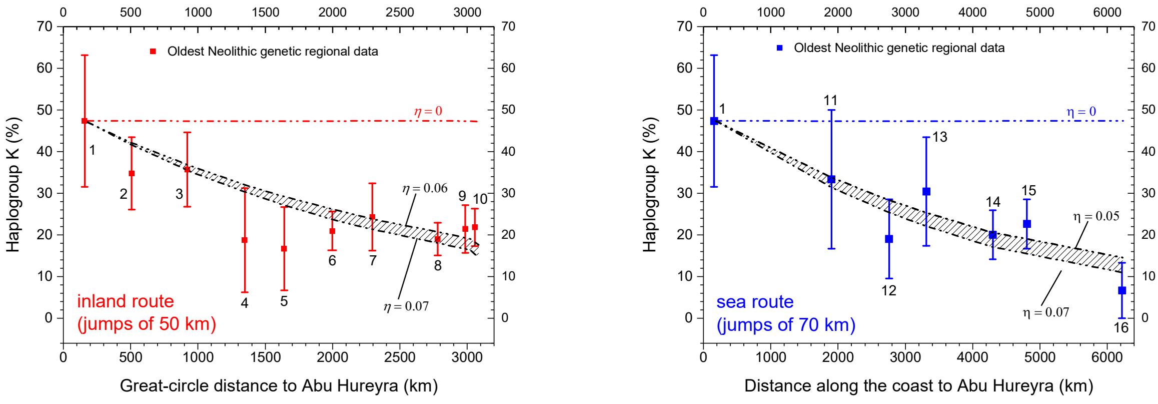
Supplementary Figure 10b. Observed and simulated Neolithic percentages of haplogroup K among early farmers. These plots are the same as those in Fig. 3 in the main paper, but here simulations have been performed by setting the carrying capacity of farmers to its upper bound ($p_{F\,max} = 1.86$ individuals/km^2). Simulations begin at the same date as in the main paper. Left: the hatched area is defined by clines that cross all error bars but two. Right: the cline for $\eta = 0.07$ crosses all error bars, and that for $\eta = 0.09$ crosses all error bars but one.

S5-D. Effect of the carrying capacity of hunter-gatherers ($p_{HG\ max}$)

In the main text and Sec. S3-B we have used the value $p_{HG\ max}$= 0.064 individuals/km^2 [52, 6, 3]. This is an intermediate value within the range of medians for several environments reported by Steele et al. [52] in their table 1, namely 0.047−0.072 individuals/ km^2 (excluding values for arctic/subarctic, tropical/subtropical and steppe environments, because we are dealing with Europe and the Near East). In Suppl. Figs. 11a-b we use this lower and upper bound, respectively. We note from Suppl. Figs. 11a-b that $p_{HG\ max}$ has the opposite effect that $p_{F\ max}$, i.e., that for given values of η and the distance, a higher value of $p_{HG\ max}$ leads to a lower percentage of haplogroup K in farmers. This is as expected because all HGs lack haplogroup K. As explained in the caption to Suppl. Fig. 11a, for the inland route (left) the clines for $0.09 \leq \eta \leq 0.1$ agree with the data, and the same happens for the sea route (right) if $0.07 \leq \eta \leq 0.09$, so both routes are consistent with η = 0.09. Similarly, as explained in the caption to Suppl. Fig. 11b, for the inland route (left) there is good agreement if $0.06 \leq \eta \leq 0.07$ and for the sea route (right) if $0.05 \leq \eta \leq 0.07$, so both routes are consistent with $0.06 \leq \eta \leq 0.07$. This confirms the main conclusion in our paper, namely that the interaction behavior between farmers and HGs was essentially the same along both routes.



Supplementary Figure 11a. Observed and simulated Neolithic percentages of haplogroup K among early farmers. These plots are the same as those in Fig. 3 in the main paper, but here simulations have been performed by setting the carrying capacity of HGs to its lower bound ($p_{HG\ max}$ = 0.047 individuals/km^2). Simulations begin at the same date as in the main paper. Left: the hatched area is defined by clines that cross all error bars but one. Right: the hatched are is defined in the same way.



Supplementary Figure 11b. Observed and simulated Neolithic percentages of haplogroup K among early farmers. These plots are the same as those in Fig. 3 in the main paper, but here simulations have been performed by setting the carrying capacity of HGs to its upper bound ($p_{HG\ max}$ = 0.072 individuals/km^2). Simulations begin at the same date as in the main paper. Left: the cline for η = 0.06 crosses all error bars but three (but almost crosses two of these), and that for η = 0.07 crosses all error bars but two. Right: the hatched area is defined by clines that cross all error bars but one.

Overall, this sensitivity analysis shows that the conclusions in the main paper do not change with the parameter values, provided that the latter lie within reasonable ranges according to the ethnographic and archaeological data available.

In the next section we generalize the sensitivity analysis above to take into account the effect of these parameters simultaneously.

S6. Envelopes on the simulation outputs

We know from the one-at-a-time sensitivity analysis in the previous section that higher values of the dispersal threshold $p_{F\,min}$, the net fecundity of farmers $R_{0,F}$ and/or the carrying capacity of farmers $p_{F\,max}$ lead to higher values of the percentage of haplogroup K (%K). On the other hand, lower values of the carrying capacity of HGs $p_{HG\,max}$ also increase the % K. In this section we perform a simultaneous analysis of the effect of these parameters, which leads to a more accurate prediction for the range of the interbreeding intensity η that is consistent with the data. As mentioned in Sec. S5, the generation time and the persistence have a negligible effect [48, 40], and similarly the dispersal distance is tightly constrained by the archaeological data (Fig. 1b in the main paper).

In all figures in this section, we will apply two sets of parameter values.

First, according to the results in the previous section, the realistic parameter values that lead to the maximum possible value of the %K (for given values of η, the initial %K and the distance) are $p_{F\,min} = 0.07$ farmers/km^2, $R_{0,F} = 2.87$, $p_{F\,max}$ =1.86 farmers/km^2 and $p_{HG\,max} = 0.047$ HGs/km^2. Obviously this parameter set will lead to the upper curve of the simulation output envelopes.

Second, the realistic parameter values that lead to the minimum possible value of the %K (for given values of η, the initial %K and the distance) are $p_{F\,min}$=0.05 farmers/km^2, $R_{0,F} = 2.09$, $p_{F\,max} = 0.96$ farmers/km^2 and $p_{HG\,max} = 0.072$ HGs/km^2. Obviously this parameter set will lead to the lower curve of the simulation output envelopes.

In all figures in this section, we shall display the simulation output envelope as a hatched area between the upper and lower curves, obtained by using these two set of parameter values respectively. This envelope or hatched area gives, for each distance, the possible values of %K obtained from the simulations using realistic parameter values (i.e., $0.05 \leq p_{F\,min} \leq 0.07$ farmers/km^2, $2.09 \leq R_{0,F} \leq 2.87$, $0.96 \leq p_{F\,max} \leq 1.86$ farmers/km^2 and $0.072 \geq p_{HG\,max} \geq 0.047$ HGs/km^2).

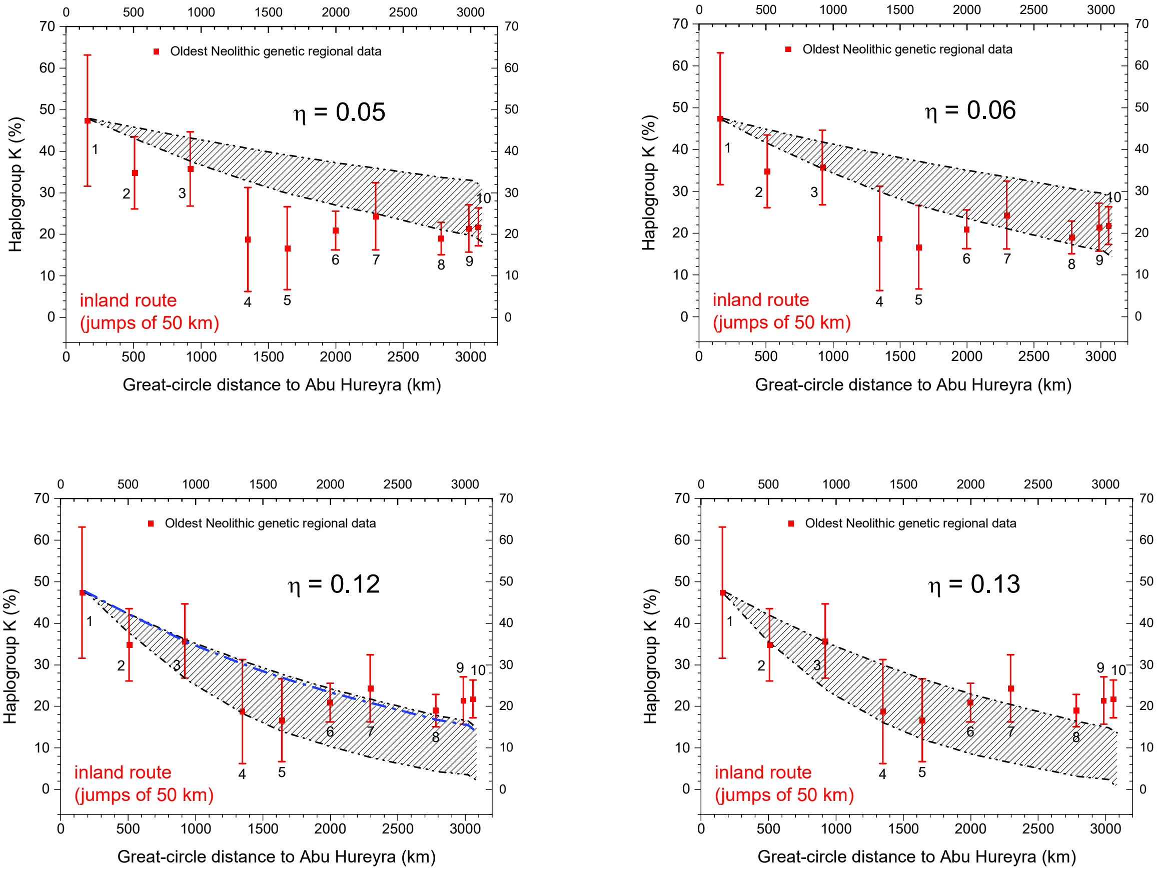
All plots in this section have been obtained in the same way as Fig. 3 and Suppl. Fig. 7, but here the value of η is fixed in each panel so that we can examine the simultaneous effects of the uncertainties in the other parameter values, and for each value of η we have an envelope rather than a line.

S6-A Envelopes for the observed value of the initial frequency of haplogroup K

In this subsection we use the observed value for the initial frequency of haplogroup K, namely 47.4% (square 1 in Fig. 3 and Suppl. Fig. 12a, from Suppl. Table 3).

Inland route (for 47.4%K in region 1)

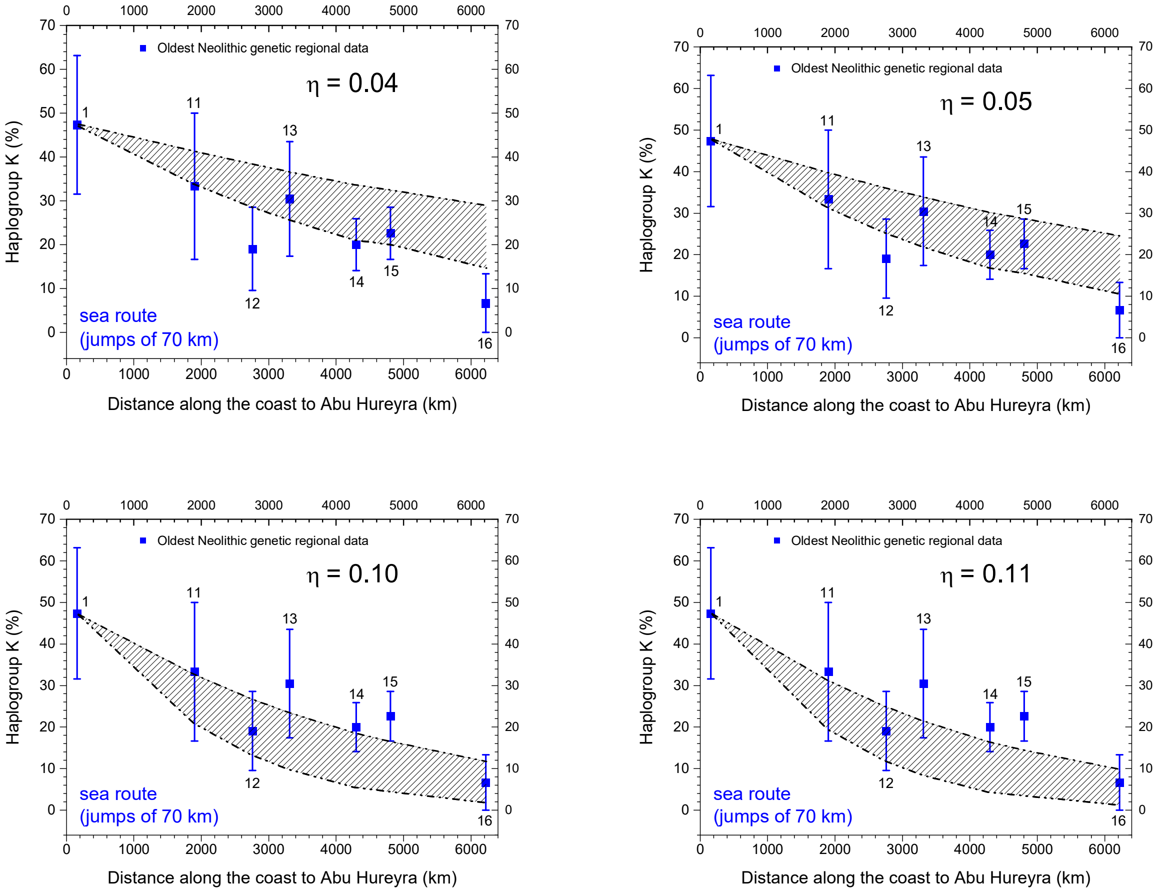
In the main paper, Fig. 3a, we have noted that for the inland route there are no simulation outputs that cross all error bars, but some simulation outputs cross all error bars except one. The panel for $\eta = 0.05$ in Suppl. Fig. 12a shows that for $\eta \leq 0.05$ the simulations cannot agree with the data in this sense, because it is not possible to find any simulation output that crosses all error bars except one (note from Suppl. Fig. 12a that the %K increases with decreasing values of η). In contrast, the panel for $\eta = 0.06$ shows that there is at least one simulation output (e.g., the lower black line) that crosses all error bars except one. Thus $\eta = 0.06$ is the minimum value of η consistent with the data along the inland route. Also in Suppl. Fig. 12a, the panel for $\eta = 0.13$ shows that for $\eta \geq 0.13$ the simulations cannot agree with the data, because it is not possible to find any simulation output that crosses all error bars except one. In contrast, the panel for $\eta = 0.12$ shows that there is at least one simulation output (e.g., the blue line) that crosses all error bars except one. Thus $\eta = 0.12$ is the maximum value of η consistent with the data along the inland route. We conclude that assuming that the initial percentage of haplogroup K is equal to its observed value (47.4%K), consistency between the genetic data and the simulations in the inland route is possible only if $0.06 \leq \eta \leq 0.12$. This refines the range found in the main paper ($0.07 \leq \eta \leq 0.08$, from Fig. 3a, where the uncertainties in the parameter values are not taken into account).



Supplementary Figure 12a. This is Fig. 4 in the main paper without boxes. Observed and simulated Neolithic percentages of haplogroup K among early farmers along the inland route assuming initially 47.4%K (square 1). For the parameter values used to obtain the upper and lower curves, see the first paragraphs in Sec. S6. Figure 3a in the main paper has been obtained by using intermediate values of those parameters. In panel $\eta = 0.12$, the blue line has been obtained for $p_{F\,min} = 0.07$ farmers/km^2, $R_{0,F} = 2.87$, $p_{F\,max}$ =1.65 farmers/km^2 and $p_{HG\,max} = 0.047$ HGs/km^2.

Sea route (for 47.4%K in region 1)

In the main paper, Fig. 3b, we have noted that for the sea route there are simulation outputs that cross all error bars. The panel for $\eta = 0.04$ in Suppl. Fig. 12b shows that for $\eta \leq 0.04$ the simulations cannot agree with the data in this sense, because it is not possible to find any simulation output that crosses all error bars. In contrast, the panel for $\eta = 0.05$ shows that there are simulation outputs that cross all error bars. Thus $\eta = 0.05$ is the minimum value of η consistent with the data along the sea route. Also in Suppl. Fig. 12b, the panel for $\eta = 0.11$ shows that for $\eta \geq 0.11$ the simulations cannot agree with the data, because it is not possible to find any simulation output that crosses all error bars. In contrast, the panel for $\eta = 0.10$ shows that there is one simulation output (the upper black line) that crosses all error bars. Thus $\eta = 0.10$ is the maximum value of η consistent with the data along the inland route. We conclude that assuming that the initial percentage of haplogroup K is equal to its observed value (47.4%K), consistency between the genetic data and the simulations in the sea route is possible only if $0.05 \leq \eta \leq 0.10$. This refines the range found in the main paper ($0.06 \leq \eta \leq 0.07$, from Fig. 3b, where the uncertainties in the parameter values are not taken into account).



Supplementary Figure 12b. Observed and simulated Neolithic percentages of haplogroup K among early farmers along the sea route assuming initially 47.4%K (square 1). For the parameter values used to obtain the upper and lower curves, see the first paragraphs in Sec. S6. Figure 3b in the main paper has been obtained by using intermediate values of those parameters.

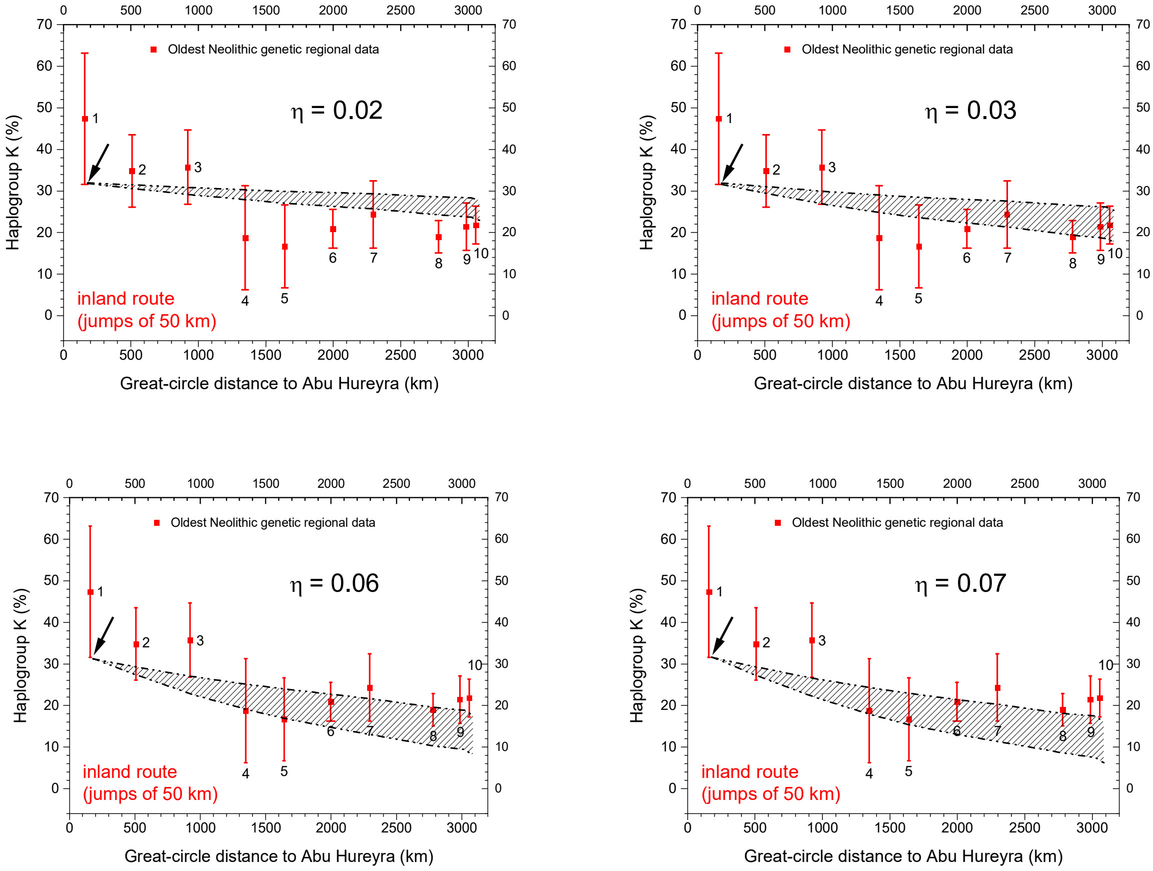
The result for the sea route taking into account the uncertainties in the parameters ($0.05 \leq \eta \leq 0.10$, from Suppl. Fig. 12b) overlaps widely with the corresponding one for the inland route ($0.06 \leq \eta \leq 0.12$, from Suppl. Fig. 12a). This gives additional support to our conclusion in the main paper that the value of η was similar along both routes. The common range ($0.06 \leq \eta \leq 0.10$) assuming that the initial percentage of haplogroup K in Syria was equal to the observed value (47.4%K) refines the corresponding estimation ($\eta \approx 0.07$) obtained without taking into account the uncertainties in the parameter values (main paper, Fig. 3).

S6-B Envelopes for the lower bound of the initial frequency of haplogroup K

In this subsection we use the lower value of the error bar for the frequency of haplogroup K in region 1, namely 31.6% (lower bound of error bar 1 and arrow in Suppl. Figs. 7a and 12a, obtained by bootstrap resampling).

Inland route (for 31.6%K in region 1)

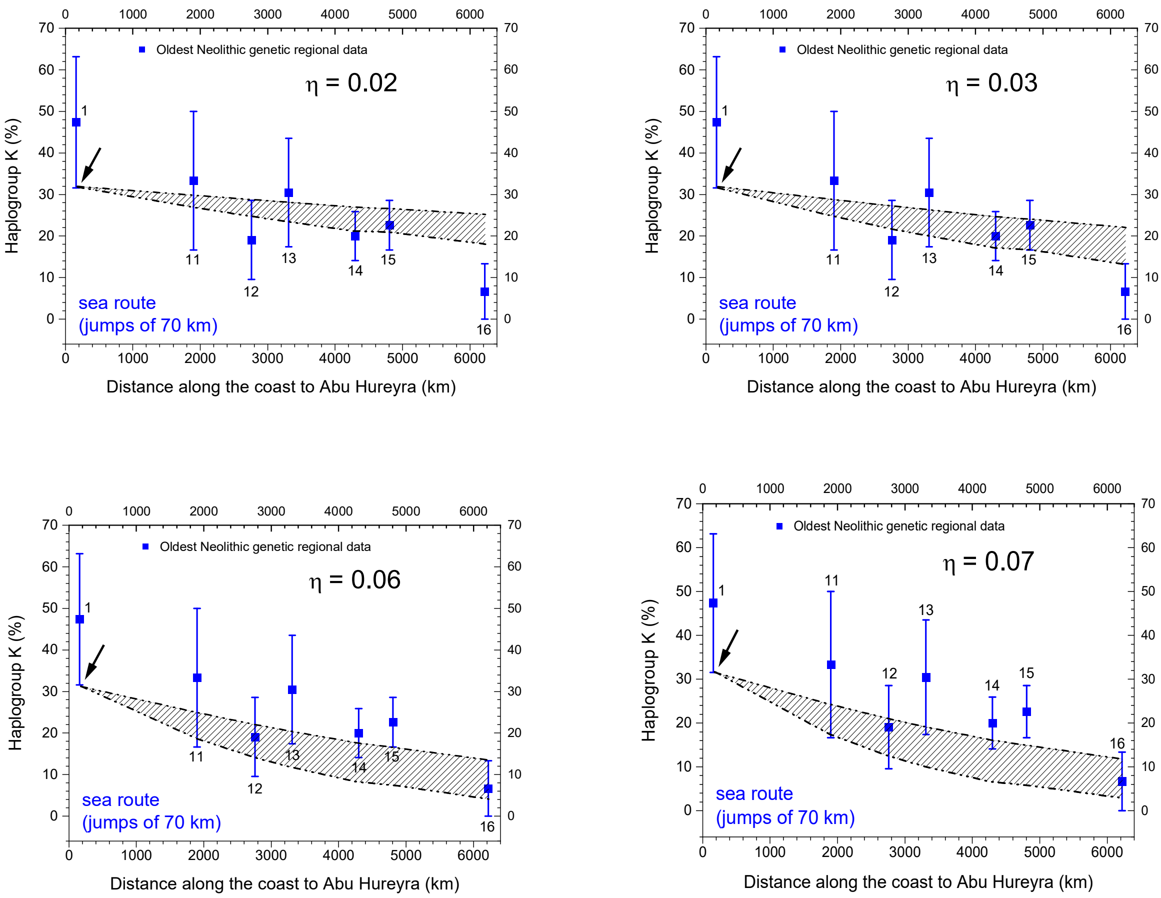
The panel for $\eta = 0.02$ in Suppl. Fig. 13a shows that for $\eta \leq 0.02$ it is not possible to find any simulation output that crosses all error bars. In contrast, the panel for $\eta = 0.03$ shows that there is at least one simulation output (e.g., the lower black line) that crosses all error bars. Thus $\eta = 0.03$ is the minimum value of η consistent with the data along the inland route. Also in Suppl. Fig. 13a, in the panel for $\eta = 0.07$ the upper black line does not cross error bar 10 and we can be sure that for $\eta \geq 0.07$ no simulation output crosses all error bars. In contrast, the panel for $\eta = 0.06$ shows that there is at least one simulation output (the upper black line) that crosses all error bars. Thus $\eta = 0.06$ is the maximum value of η consistent with the data along the inland route. We conclude that consistency between the genetic data and the simulations in the inland route is possible only if $0.03 \leq \eta \leq 0.06$. This refines the range found in Suppl. Fig. 7a (left), where the uncertainties in the parameter values are not taken into account ($0.03 \leq \eta \leq 0.04$).



Supplementary Figure 13a. Observed and simulated Neolithic percentages of haplogroup K among early farmers along the inland route assuming initially 31.6%K (arrow). For the parameter values used to obtain the upper and lower curves, see the first paragraphs in Sec. S6. Supplementary Figure 7a (left) has been obtained by using intermediate values of those parameters.

Sea route (for 31.6%K in region 1)

The panel for $\eta = 0.02$ in Suppl. Fig. 13b shows that for $\eta \leq 0.02$ it is not possible to find any simulation output that crosses all error bars (because the %K increases with decreasing values of η). In contrast, the panel for $\eta = 0.03$ shows that there is one simulation output (the lower black line) that crosses all error bars. Thus $\eta = 0.03$ is the minimum value of η consistent with the data along the inland route. Also in Suppl. Fig. 13b, the panel for $\eta = 0.07$ shows that for $\eta \geq 0.07$ the simulations do not cross all error bars. In contrast, the panel for $\eta = 0.06$ shows that there is one simulation output (the upper black line) that crosses all error bars. Thus $\eta = 0.06$ is the maximum value of η consistent with the data along the sea route. We conclude that assuming that the initial percentage of haplogroup K is equal to the lower bound of its error bar (31.6%K), consistency between the genetic data and the simulations in the sea route is possible only if $0.03 \leq \eta \leq 0.06$. This refines the estimation found in Suppl. Fig. 7a (right), where the uncertainties in the parameter values are not taken into account ($\eta \approx 0.038$).



Supplementary Figure 13b. Observed and simulated Neolithic percentages of haplogroup K among early farmers along the sea route assuming initially 31.6%K (arrow). For the parameter values used to obtain the upper and lower curves, see the first paragraphs in Sec. S6. Supplementary Figure 7a (right) has been obtained by using intermediate values of those parameters.

The result for the sea route taking into account the uncertainties in the parameters ($0.03 \leq \eta \leq 0.06$, from Suppl. Fig. 13b) is the same that the corresponding one for the inland route ($0.03 \leq \eta \leq 0.06$, from Suppl.

Fig. 13a). This gives additional support to our conclusion in the main paper that the value of η was similar along both routes. It also refines the corresponding estimation without taking into account the uncertainty in the parameter values ($\eta \approx 0.038$, from Suppl. Fig. 7a).

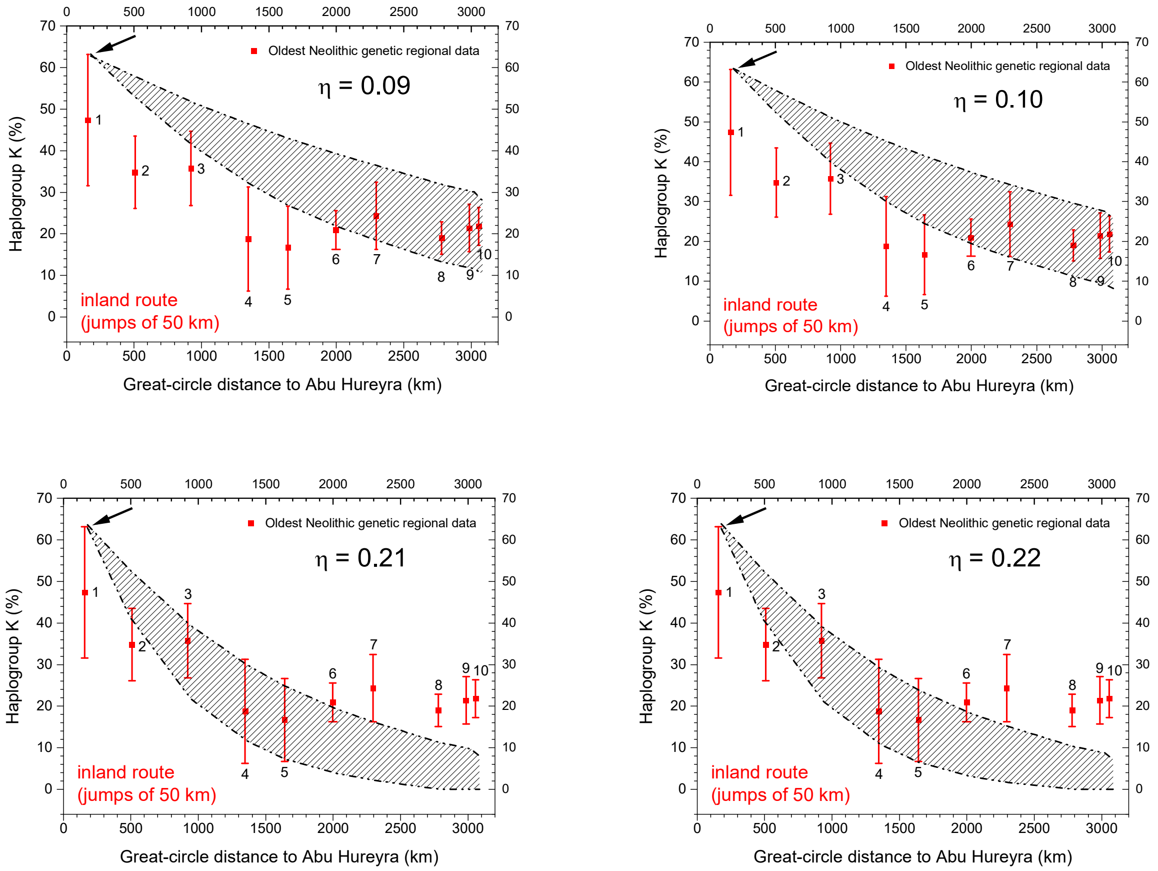
S6-C Envelopes for the upper bound of the initial frequency of haplogroup K

In this subsection we use the upper value of the error bar for the frequency of haplogroup K in region 1, namely 63.2% (upper bound of error bar 1 and arrow in Suppl. Figs. 7a and 14a, obtained from bootstrap resampling).

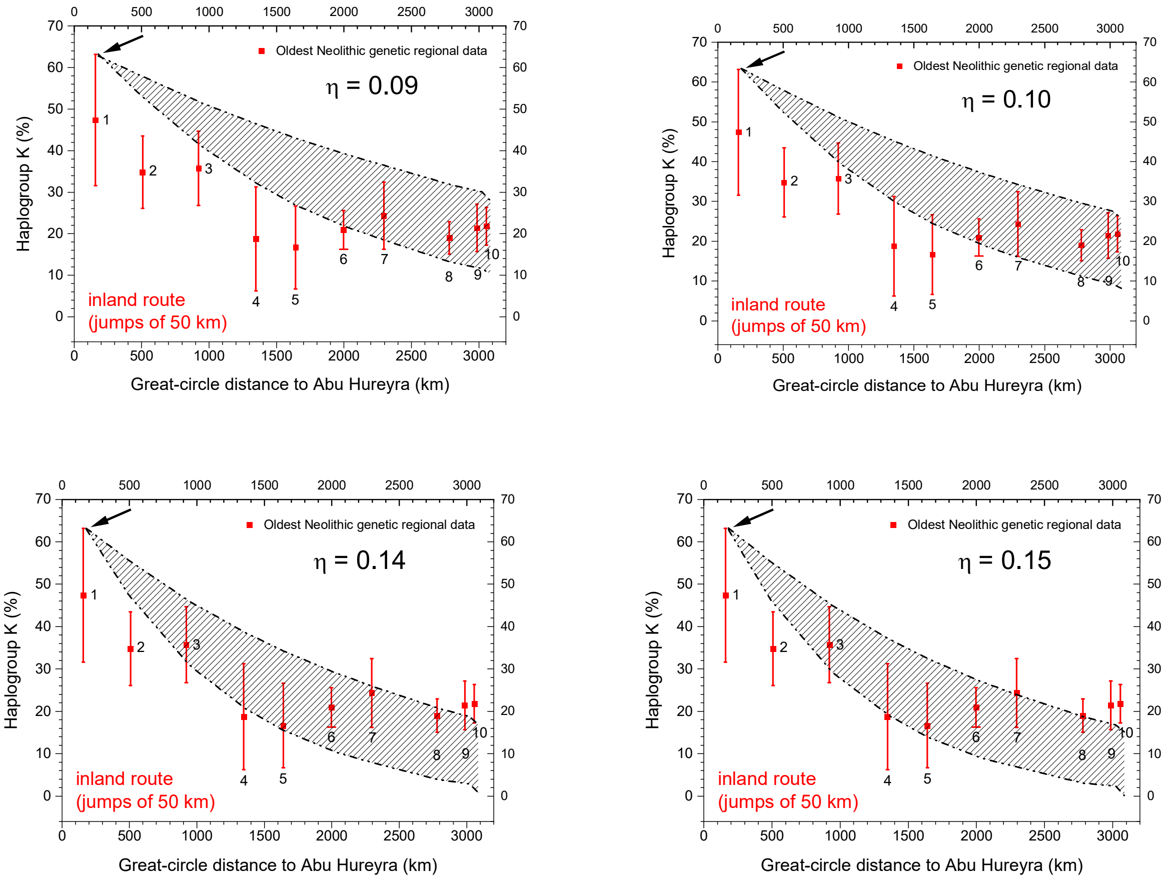
Inland route (for 63.2%K in region 1)

Let us now consider the inland route and assume an initial value of 63.2%K. Supplementary Figures 7b (left) and 14a show that no matter which simulated cline we consider, it will not cross at least four error bars. At first sight, an adequate approach could be to note from Suppl. Fig. 14a that for $\eta \leq 0.09$ or $\eta \geq 0.22$ there is not any simulated cline that crosses all error bars except four of them. In contrast, for $0.10 \leq \eta \leq 0.21$ there is always at least one cline that crosses all error bars except four (e.g., the lower curve in the panel $\eta = 0.10$ and the upper curve in panel $\eta = 0.21$). From this we could be tempted to conclude that for $0.10 \leq \eta \leq 0.21$ there is fair overall consistency between the simulations and the observed genetic data. However, note that for the panel $\eta = 0.10$ in Suppl. Fig. 14a the agreement between simulations (hatched area) and data (error bars) is clearly better than for the panel $\eta = 0.21$, because the latter displays rather poor agreement for the most distant regions (error bars 8, 9 and 10).

An alternative approach is to note from Suppl. Fig. 14b that for $\eta \leq 0.09$ or $\eta \geq 0.15$ there are at least two error bars that do not cross the hatched area. In contrast, for $0.10 \leq \eta \leq 0.14$ only one error bar does not cross the hatched area. Thus in this sense, consistency between the genetic data and the simulations is attained for $0.10 \leq \eta \leq 0.14$. This refines the range $0.10 \leq \eta \leq 0.12$, estimated without taking into account the uncertainties in the parameter values from Suppl. Fig. 7b (left). We think that this approach is more reasonable than that in the previous paragraph for two reasons. First, non-homogeneous parameter values can lead to many different clines (but all of them will be within the hatched areas in Suppl. Fig. 14b), so it is reasonable to consider each hatched area as a whole rather than only specific clines for homogeneous parameter values (e.g., the upper and lower curves of each hatched area). Second, in Suppl. Fig. 14b the plots for $\eta = 0.10$ and $\eta = 0.14$ agree similarly well with the error bars, so this approach (Suppl. Fig. 14b) seems more reasonable than that in the previous paragraph (Suppl. Fig. 14a). Anyway the results are similar ($0.10 \leq \eta \leq 0.21$ for Suppl. Fig. 14a and $0.10 \leq \eta \leq 0.14$ for Suppl. Fig. 14b), so the conclusions obtained by using either of both approaches would be much the same.



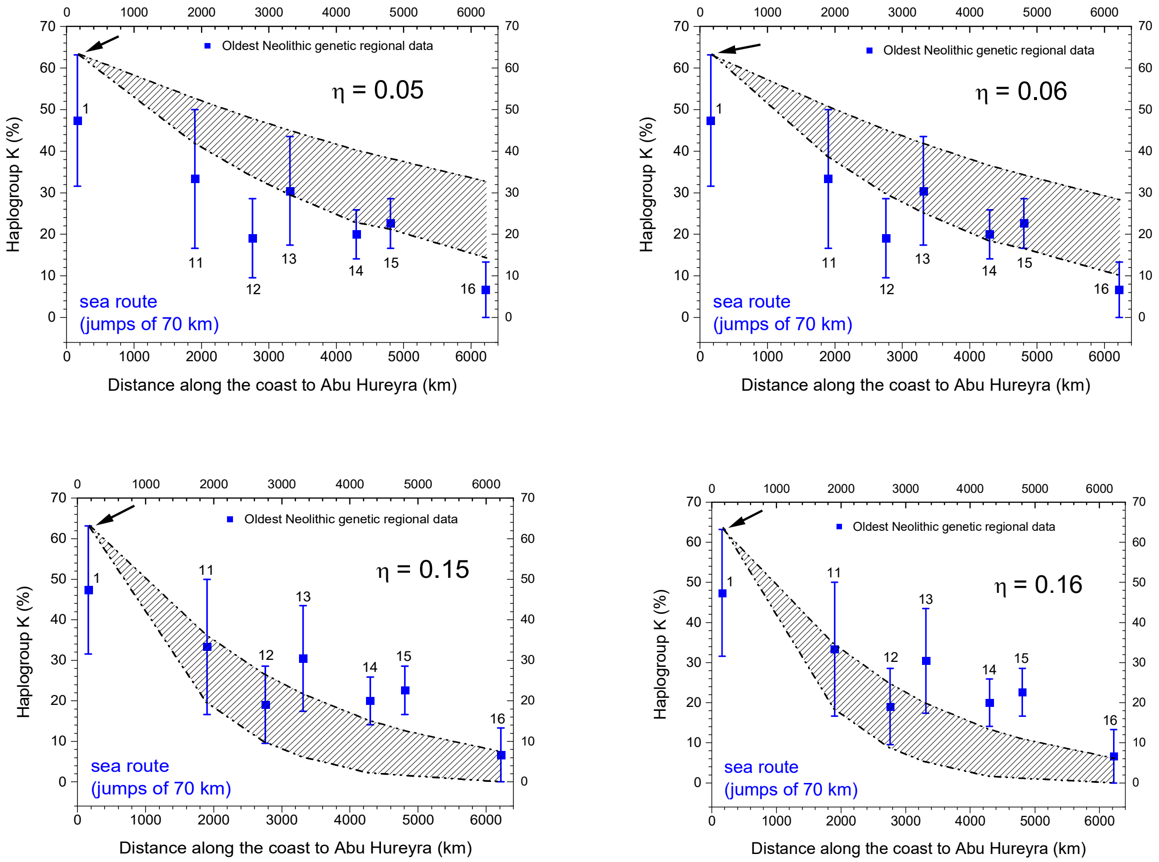
Supplementary Figure 14a. Observed and simulated Neolithic percentages of haplogroup K among early farmers along the inland route assuming initially 63.1%K (arrow). For the parameter values used to obtain the upper and lower curves, see the first paragraphs in Sec. S6. Supplementary Figure 7b (left) has been obtained by using intermediate values of those parameters. These plots illustrate a possible approach, but we prefer that in Supplementary Figure 14b (as explained in the text).



Supplementary Figure 14b. Observed and simulated Neolithic percentages of haplogroup K among early farmers along the inland route assuming initially 63.1%K (arrow). For the parameter values used to obtain the upper and lower curves, see the first paragraphs in Sec. S6. Supplementary Figure 7b (left) has been obtained by using intermediate values of those parameters. As explained in the text, these plots seem more appropriate than those in Supplementary Figure 14a.

Sea route (for 63.2%K in region 1)

In Suppl. Fig. 7b (right) we have noted that for the inland route there are no simulation outputs that cross all error bars, but some simulation outputs cross all error bars except one. The panel for $\eta = 0.05$ in Suppl. Fig. 14c shows that for $\eta \leq 0.05$ the simulations cannot agree with the data along the sea route in this sense, because it is not possible to find any simulation output that crosses all error bars except one (note from Suppl. Figs. 14c that the %K increases with decreasing values of η). In contrast, the panel for $\eta = 0.06$ shows that there is at least one simulation output (e.g., the lower black line) that crosses all error bars except one. Thus $\eta = 0.06$ is the minimum value of η consistent with the data along the inland route. Also in Suppl. Fig. 14c, the panel for $\eta = 0.16$ shows that for $\eta \geq 0.16$ the simulations cannot agree with the data, because it is not possible to find any simulation output that crosses all error bars except one. In contrast, the panel for $\eta = 0.15$ shows that there is at least one simulation output (e.g., the upper black line) that crosses all error bars except one. Thus $\eta = 0.15$ is the maximum value of η consistent with the data along the inland route. We conclude that assuming that the initial percentage of haplogroup K is equal to its upper value (63.2%K), consistency between the genetic data and the simulations in the sea route is possible only if $0.06 \leq \eta \leq 0.15$. This refines the range found without taking into account the uncertainties in the parameter values ($0.07 \leq \eta \leq 0.10$) in Suppl. Fig. 7b (right).



Supplementary Figure 14c. Observed and simulated Neolithic percentages of haplogroup K among early farmers along the sea route assuming initially 63.2%K (arrow). For the parameter values used to obtain the upper and lower curves, see the first paragraphs in Sec. S6. Supplementary Figure 7b (right) has been obtained by using intermediate values of those parameters.

The result for the sea route taking into account the uncertainties in the parameters ($0.06 \leq \eta \leq 0.15$, from Suppl. Fig. 14c) overlaps widely with the corresponding one for the inland route ($0.10 \leq \eta \leq 0.14$, from Suppl. Fig. 14b). This gives additional support to our conclusion in the main paper that the value of η was similar along both routes. The common range ($0.10 \leq \eta \leq 0.14$) assuming that the initial percentage of haplogroup K in Syria was equal to its upper value (63.2%K) refines the corresponding estimation obtained without taking into account the uncertainties in the parameter values ($\eta \approx 0.10$, from Suppl. Fig. 7b).

In the main paper, Methods, we derive Eq. (14) and use it to convert these ranges of η into ranges of the percentage of pioneering farmers that interbred with HGs or acculturated them.

S7. The spread rates along the inland and sea routes are different

In the main paper we have noted that our simulations (full lines in Fig. 1b) yield a spread rate along the sea route that is twice faster than that along the inland route. In order to test such a demonstrable difference directly from the archeological data, we performed linear regressions of the error bars in Fig. 1b. For the inland route we find that the spread rate is 0.89 km/yr and its range is 0.69-1.01 km/yr with 95% confidence level (CL). For the sea route the spread rate is 1.64 km/yr and its range is 1.27-2.01 km/yr with 95% CL. These ranges do not overlap, which confirms that there is a statistically significant difference between both dispersal rates.

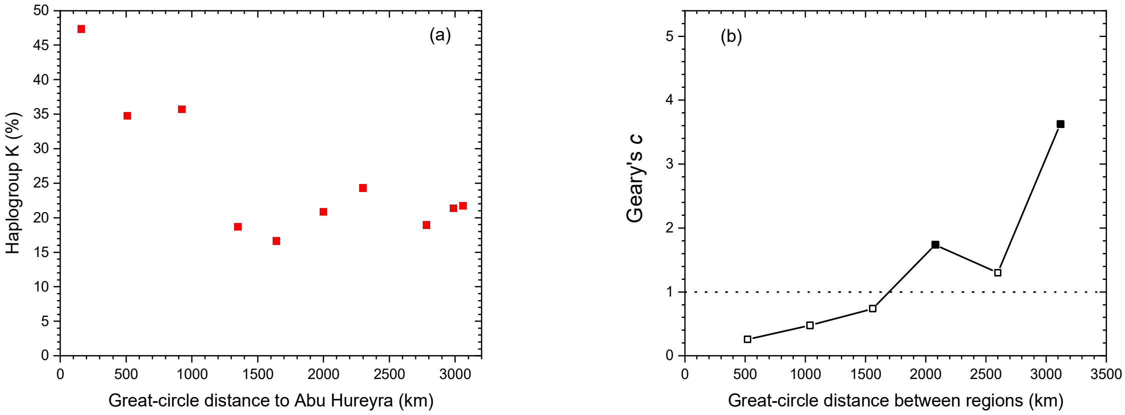
S8. Formal test of the existence of the genetic cline along the inland and sea routes

In the main paper we have argued that Figs. 3a-b show the existence of a genetic cline of haplogroup K along both the inland and sea routes of Neolithic spread in Europe. Here we test this statement formally, first for the observed values (squares in Figs. 3a-b) and then by taking into account the uncertainty in the data (as illustrated by the error bars in Figs. 3a-b).

Geary's c correlogram and the Bonferroni technique provide a simple way to test formally the existence of clines [53], and their results agree with those of alternative approaches [54]. The first step is to compute the distances between all pairs of locations or regions where the percentage of haplogroup K (%K) has been measured. These distances are then grouped in several distance intervals or classes. For each class, a measure of the dissimilarity (Geary's coefficient c) between the values of the %K for the corresponding pairs is computed. Finally, in the correlogram each point has ordinate equal to the value of Geary's coefficient c for a class and abscissa equal to its distance. If there is a gradual spatial variation of the %K, nearby locations (i.e., those separated by small distances) will have similar values of the %K and Geary's coefficient c will be small. On the other hand, for large distances the differences in %K will be important and Geary's coefficient c will be large. Thus a cline or gradient (of either gradually decreasing or increasing values) of the %K can be detected as an increasing trend in Geary's coefficient c as a function of distance. In contrast, for a random spatial distribution the correlogram will be flat and have an expected mean value of Geary's coefficient $c = 1$ (see Ref. [53], Fig. 13.5h).

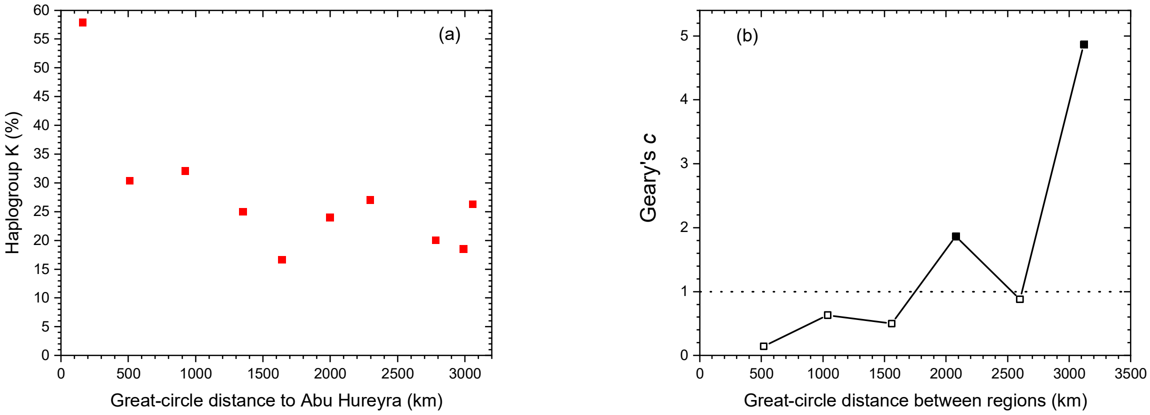
Inland route

For the inland route, Suppl. Fig. 15a shows the observed values of the %K as a function of their distance to Abu Hureyra (i.e., the squares in Fig. 3a in the main paper, from Suppl. Table 3). Since there are 10 regions along the inland route (Suppl. Table 1), we compute all 45 great-circle distances between pairs of regions using a free internet application [55] that applies the Haversine equation (S8). For each region, we use the mean location (latitude and longitude) of the individuals whose mt haplogroup is known (Suppl. Table2). We have grouped these distances (between pairs of regions) in 6 distance classes in agreement with Sturge's rule (Eq. (13.3) in Ref. [53]) and obtained the corresponding correlogram (Suppl. Fig. 15b) using PASSaGE software [56]. As expected, Geary's *c* displays an increasing trend with increasing distances in Suppl. Fig. 15b. This correlogram is significant over the entire range of classes ($P<0.05$ Bonferroni corrected [53, 57]). Therefore we can reject the null hypothesis that the values of Geary's *c* coefficient for the whole set of classes are equal to the value expected under a random spatial distribution ($c = 1$), and this confirms formally the existence of the cline [53, 57] for the observed values of the %K along the inland route.



Supplementary Figure 15. (a) Observed frequencies of haplogroup K among early farmers along the inland route. (b) The corresponding correlogram, with data points plotted at the upper distance of each class. In (b) black squares are class-specific significant values. In (b) Geary's *c* has an increasing trend, as expected for a cline in the original data (a) and the correlogram is significant over the entire range of classes ($P<0.05$ Bonferroni corrected). Under a random spatial distribution (absence of a genetic cline) in (a), the correlogram (b) would be flat and centered about the dashed horizontal line ($c = 1$).

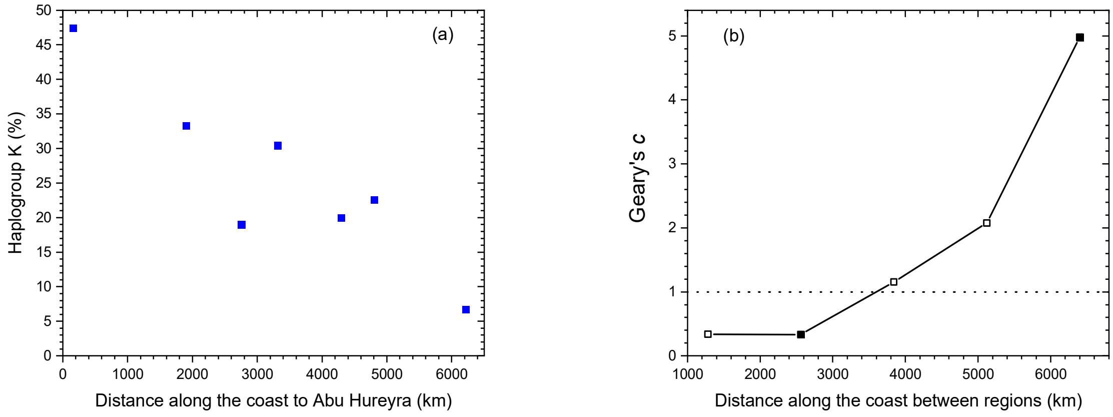
Next we take the data uncertainty in the regional values of the %K into account for the inland route. The first step is a bootstrap resampling with replacement (as also used to obtain the error bars in Fig. 3a in the main paper). This means for example that if a region has 30 early farmers whose mt haplogroup is known, and 10 of them have haplogroup K, we choose at random one of 30 balls (10 of them black and 20 white), record its color, put it again with the other balls, choose a second ball, and repeat this procedure until we have extracted 30 balls. In general, this will give a %K different from the original one (note that the original one is 10/30 or 33.33% in this example). Supplementary Figure 16a shows an example of such bootstrap frequencies, as obtained by applying this bootstrap procedure to all inland regions. Supplementary Figure 16b shows the corresponding correlogram. Again Geary's *c* displays an increasing trend with increasing distances, and the correlogram is significant over the entire range of classes ($P<0.05$ Bonferroni corrected). We repeated this procedure many times and computed the corresponding correlograms (similar to those in Suppl. Figs. 15b and 16b). In all of them, we observed that Geary's *c* increases with increasing distances, as expected for a genetic cline, and more than 80% of correlograms are significant over the entire range of classes ($P<0.05$ Bonferroni corrected). Thus in this sense, with 80% confidence level we can reject the null hypothesis that the values of Geary's *c* coefficient for the whole set of classes are equal to the value expected under a random spatial distribution ($c = 1$). This confirms formally the existence of the cline [53, 57] along the inland route, taking the uncertainty in the values of the %K into account.



Supplementary Figure 16. (a) An example of haplogroup K frequencies among early farmers along the inland route, obtained by bootstrap resampling. (b) The corresponding correlogram, with data points plotted at the upper distance of each class. In (b) black squares are class-specific significant values. In (b) Geary's c has an increasing trend, as expected for a cline in the original data (a) and the correlogram is significant over the entire range of classes (P<0.05 Bonferroni corrected). Under a random spatial distribution (no genetic cline) in (a), the correlogram (b) would be flat and centered about the dashed horizontal line ($c = 1$).

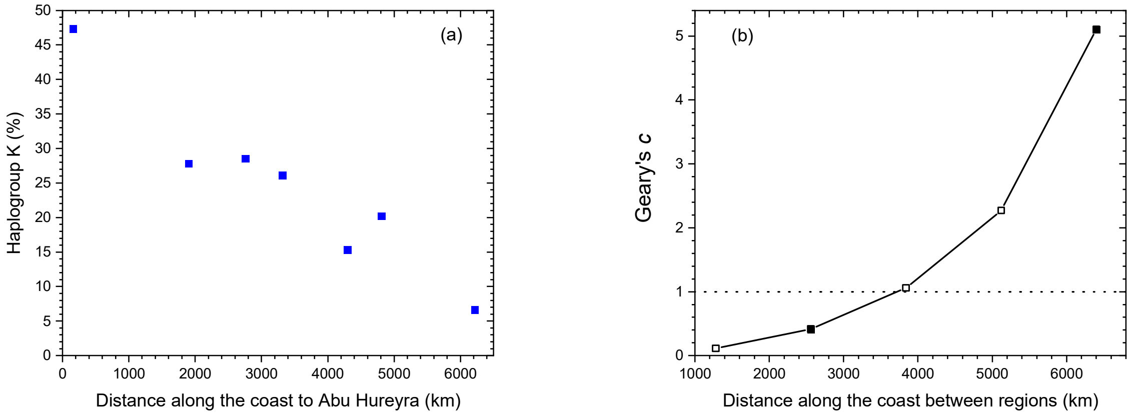
Sea route

For the sea route, we have followed the same approach. Supplementary Figure 17b shows the observed values of the %K as a function of their distance to Abu Hureyra (i.e., the squares in Fig. 3b in the main paper, from Supp. Table 3). Since there are 7 regions along the sea route (Supp. Table 1), we have computed all 21 distances along the coast between pairs of regions and grouped them in 5 distance classes in agreement with Sturge's rule (Eq. (13.3) in Ref. [53]). The corresponding correlogram (Suppl. Fig. 17b) shows that, as expected, Geary's c displays an increasing trend with increasing distances and it is significant over the entire range of classes (P<0.05 Bonferroni corrected [53]). Thus we can reject the null hypothesis that the values of Geary's c coefficient for the whole set of classes are equal to the value expected under a random spatial distribution ($c = 1$). This confirms formally the existence of the sea cline for the observed values of the %K along the sea route.



Supplementary Figure 17. (a) Observed frequencies of haplogroup K among early farmers along the sea route. (b) The corresponding correlogram, with data points plotted at the upper distance of each class. In (b) black squares are class-specific significant values. In (b) Geary's c has an increasing trend, as expected for a cline in the original data (a) and the correlogram is significant over the entire range of classes (P<0.05 Bonferroni corrected). Under a random spatial distribution (absence of a genetic cline) in (a), the correlogram (b) would be flat and centered about the dashed horizontal line ($c = 1$).

Next we take the uncertainty in the regional values of the %K into account for the sea route. Supplementary Figure 15a shows an example of a bootstrap frequencies, and Suppl. Fig. 18b shows the corresponding correlogram. Again Geary's *c* increases with increasing distances, and the correlogram is significant over the entire range of classes (P<0.05 Bonferroni corrected). We repeated this procedure many times and computed the corresponding correlograms (similar to those in Suppl. Figs. 17b and 18b). In all of them Geary's *c* displays an increasing trend with increasing distances, as expected for a cline, and at least 80% of correlograms are significant over the entire range of classes (P<0.05 Bonferroni corrected). In this sense, with 80% confidence level we can reject the null hypothesis that the values of Geary's *c* coefficient for the whole set of classes are equal to the value expected under a random spatial distribution ($c = 1$). This confirms formally the existence of the cline [53, 57] along the sea route, taking the uncertainty in the values of the %K into account.



Supplementary Figure 18. (a) An example of haplogroup K frequencies among early farmers along the sea route, obtained by bootstrap resampling. (b) The corresponding correlogram, with data points plotted at the upper distance of each class. In (b) black squares are class-specific significant values. In (b) Geary's *c* has an increasing trend, as expected for a cline in the original data (a) and the correlogram is significant over the entire range of classes (P<0.05 Bonferroni corrected). Under a random spatial distribution (absence of a genetic cline) in (a), the correlogram (b) would be flat and centered about the dashed horizontal line ($c = 1$).

S9. The cline of haplogroup K is not due to random effects

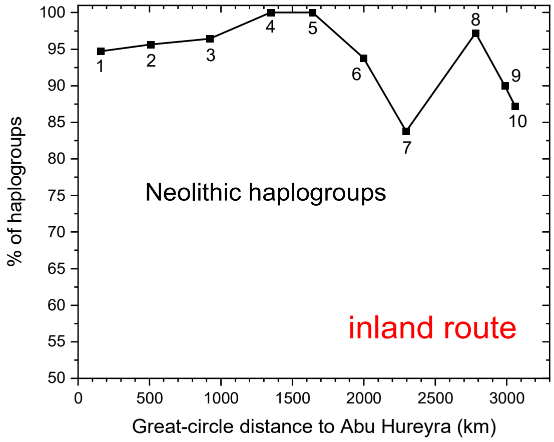
In this section we test the possibility that random effects were so strong that they were responsible for the spatial distribution of haplogroup K. If this were the case, the clines that we have analyzed (Fig. 3 in the main paper) would not be due to interbreeding (as assumed by our model) but just to random effects.

Inland route

Some mt haplogroups of early farmers can be considered Neolithic in the sense that they were essentially absent in HGs. This includes haplogroups K, T2, N1a, J, HV, V, W, X [58] and, along the inland route, also haplogroup H [59, 60][22]. We used our databases of early farmers (Suppl. Tale 1) and hunter-gatherers (Suppl. Table 5) to check these claims and identify additional haplogroups. The results show that these 9 mt haplogroups were indeed present in early farmers and essentially absent in HGs, and so were 12 additional ones, namely R0, L3, N, U1, N1b, U3, T1, D1/G1a1, C5, R1, T and I1. They all have frequencies in HGs below 2% (Suppl. Table 6), so they have negligible effects (Sec. S1-D). For each region along the inland route, we computed the frequency of each of these Neolithic haplogroups and added them up (Suppl. Table

[22] A detailed discussion on haplogroup H is included in Sec. S10.

7a). The line in Suppl. Fig. 19 shows that in all regions the great majority of early farmers (between 84% and 100%) have these 21 haplogroups that were essentially absent in HGs.

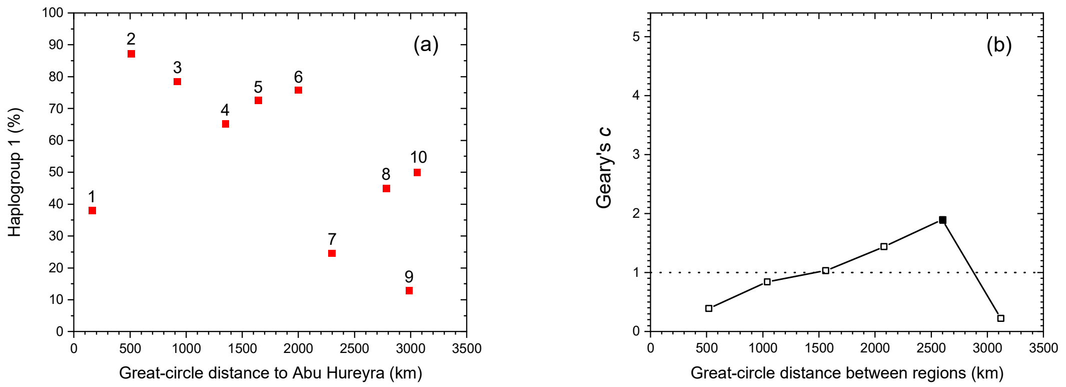


Supplementary Figure 19. Total percentage of Neolithic haplogroups (K, T2, N1a, J, HV, V, W, X, H, R0, L3, N, U1, N1b, U3, T1, D1/G1a1, C5, R1, T and I1) among early farmers in the 10 regions of the inland route. Frequencies from Supp. Table 7a.

We shall use Suppl. Fig. 19 to generate random genetic clines, as follows. According to Suppl. Fig. 19 or Suppl. Table 7, in region 1 a percentage of 94.7% (i.e., a proportion of 0.947) of early farmers have Neolithic haplogroups. We generate a random number between 0 and 0.947. The result (e.g., 0.380) will be the proportion of our first simulated haplogroup (Suppl. Fig. 20a, region 1). Analogously, 95.7% of early farmers in region 2 have Neolithic haplogroups (Suppl. Fig. 19 or Suppl. Table 7), so for region 2 we generate a random number between 0 and 0.957 and the result (e.g., 0.873) gives the proportion in Suppl. Fig. 20a, region 2. Following the same procedure for regions 3, 4, ..., 10 we have obtained Suppl. Fig. 20a. In order to determine if there is a cline in Suppl. Fig. 20a we apply Geary's correlogram and the Bonferroni technique, as already explained and applied in Sec. S8. In Suppl. Fig. 20b we see that the correlogram is flat (compared to those in Suppl. Figs. 15b, 16b, 17b and 18b) and centered about the dashed horizontal line $c = 1$. Moreover the condition of statistical significance for a cline to exist (P<0.05 Bonferroni corrected) is not satisfied. Thus there is no statistically significant cline in the original data shown in Suppl. Fig. 20a (which have been generated at random, as explained above). This in contrast to the correlograms in Suppl. Figs. 15b, 16b, 17b and 18b, all of which display an increasing trend and are statistically significant (P<0.05 Bonferroni corrected), as expected for a cline in their original data (i.e., if nearby values are similar rather than purely random).

In order to generate a second random cline, we subtract the proportion of all Neolithic haplogroups in region 1 (0.947) minus the value generated randomly above for the first cline (0.380), i.e., 0.947-0.380=0.567 and generate a random number between 0 and 0.567. We proceed analogously for all regions. The third, fourth, etc. clines are obtained in the same way, until almost all frequencies are very small (below e.g. 2%). Then, in order to obtain more random clines, we begin again from Suppl. Fig. 19.

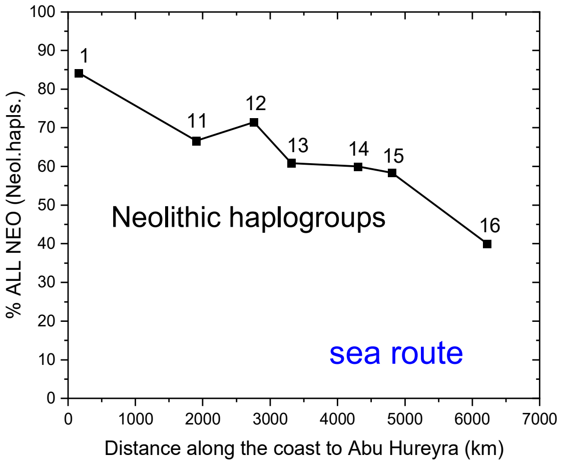
We have generated many clines in this way and found no statistically significant cline (the condition P<0.05 Bonferroni corrected is not satisfied) in more than 90% of the cases. Thus in this sense, with 90% confidence level we can conclude that a spatial decrease similar to that of haplogroup K is not obtained by generating clines at random along the inland route.



Supplementary Figure 20. (a) Percentages of an hypothetical haplogroup ("haplogroup 1") among early farmers in the 10 regions of the inland route, obtained at random. (b) The corresponding correlogram, with data points plotted at the upper distance of each class. Black squares are class-specific significant values.

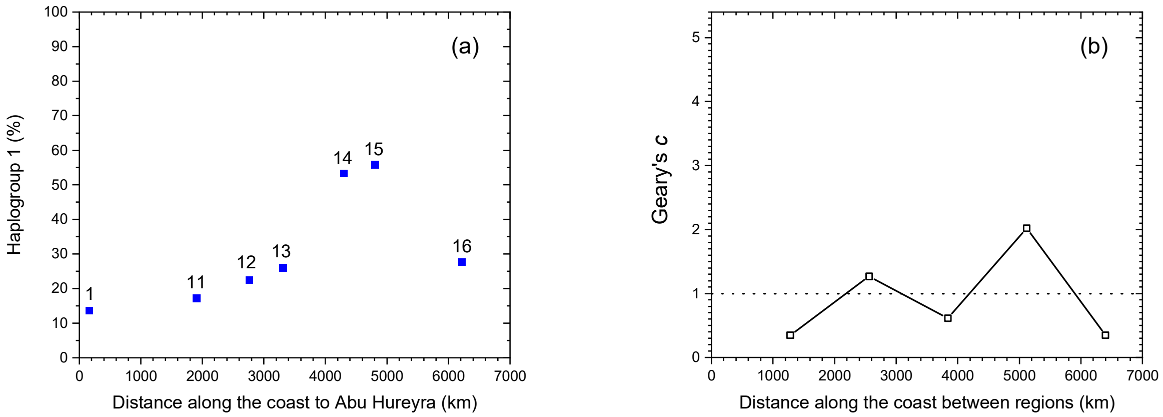
Sea route

As mentioned in the previous subsection, we have defined and identified Neolithic haplogroups as those present in early farmers and essentially absent in HGs. More precisely, we have considered as Neolithic those haplogroups with frequencies below about 2% in HGs (Suppl. Table 6) because percentages up to 2% in HGs have negligible effects (Sec. S1-D). Along the sea route, haplogroup H cannot be considered a Neolithic haplogroup because its frequency in HGs is 5.8% (Suppl. Table 6). In fact haplogroup H is widely considered as a HG marker along the sea route, and clades H1 and H3 are thought to have spread from a glacial Iberian refugium [60]. The Neolithic haplogroups that are present in at least one early farmer along the sea route are K, T2, N1a, J, HV, V, W, X, R0, L3, N, U1, T1, U3 and I (Suppl. Table 7). For each region along the inland route, we computed the frequency of each of these Neolithic haplogroups and added them up (Suppl. Table 7b) in Suppl. Fig. 21.



Supplementary Figure 21. Total percentage of Neolithic haplogroups (K, T2, N1a, J, HV, V, W, X, R0, L3, N, U1, T1, U3 and I) among early farmers in the 7 regions of the sea route. Frequencies from Suppl. Table 7b.

Proceeding exactly as in the previous section but using Suppl. Fig. 21 (instead of Suppl. Fig. 19), we have generated many random clines. Supplementary Figure 22a shows an example, and Suppl. Fig. 22b the corresponding Geary's correlogram. We see that the correlogram is flat (compared to those in Suppl. Figs. 15b, 16b, 17b and 18b) and centered about the dashed horizontal line $c = 1$. Moreover the condition of statistical significance for a cline to exist (P<0.05 Bonferroni corrected) is not satisfied. Thus there is no statistically significant cline in the original data shown in Suppl. Fig. 22a (which have been generated at random, as explained above). This in contrast to the correlograms in Figs. 15b, 16b, 17b and 18b, all of which display an increasing trend and are statistically significant (P<0.05 Bonferroni corrected), as expected for a cline in their original data (i.e., if nearby values are similar rather than purely random). We have generated many random clines and found no statistically significant cline (the condition P<0.05 Bonferroni corrected is not satisfied) in more than 90% of the cases. Thus in this sense, with 90% confidence level we can conclude that a spatial decrease similar to that of haplogroup K is not obtained by generating clines at random along the sea route.



Supplementary Figure 22. (a) Percentages of an hypothetical haplogroup ("haplogroup 1") among early farmers in the 7 regions of the sea route, obtained at random. (b) The corresponding correlogram, with data points plotted at the upper distance of each class.

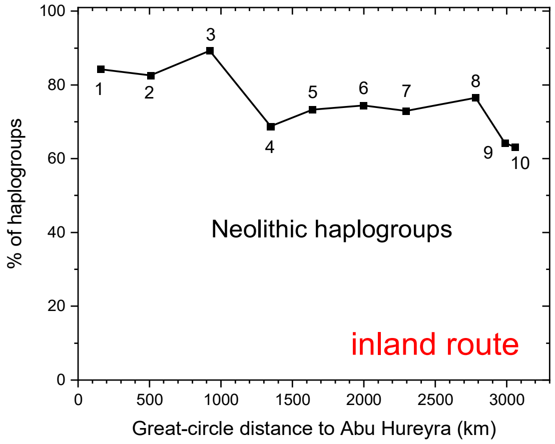
Just to summarize, for the sea route we have reached the same conclusion as for the inland route, namely that a spatial decrease similar to that of haplogroup K is not obtained (with 90% confidence level) by generating clines at random. This indicates that for haplogroup K the formation of the cline is driven by a non-random process that dominates over purely random effects. In our model this process is interbreeding and/or acculturation (several other processes are discussed and dismissed in Sec. S1-B).

S10. The effect of mitochondrial haplogroup H

Along the sea route (as mentioned in Sec. S8) mitochondrial haplogroup H is widely considered a HG marker because it has considerable frequencies in western Mediterranean HGs (13% in Spain and 17% in Portugal, from our Suppl. Table 5). Moreover clades H1 and H3 are thought to have spread from a glacial Iberian refugium [60].

Along the inland route, in the previous section we have considered haplogroup H as Neolithic because its frequency in HGs is below 2% along the inland route (Suppl. Table 6). This agrees with some previous papers that have featured haplogroup H as Neolithic in central and northern Europe [59, 60]. In contrast, however, some authors have argued that haplogroup H cannot be unambiguously ascertained to Neolithic farmers neither hunter-gatherers in central Europe [58]. Those studies were published more than 10 years ago. More recently haplogroup H has been found in one HG in Germany dated 6,200-5,400 cal BCE (Suppl. Table 5), which overlaps only very slightly with the dates of early LBK farmers in Germany (5,500-5,000 cal

BCE, from Suppl. Table 1) so we cannot exclude that haplogroup H was present in HGs in Germany before the arrival of the first farmers. Furthermore, in two other regions of the inland route (Romania and Denmark) HGs with haplogroup H have now been found (with frequencies 7.1% and 2.6%, respectively) and dated before 7,000 cal. BCE (Suppl. Table 5), i.e., much older than the earliest farmers in those regions (Suppl. Table 1). This proofs conclusively the presence of haplogroup H in HGs along the inland route. Although its overall frequency was below 2% (Suppl. Table 6), in some regions it was as high as about 7%. For this reason, it is important to see that our results would not change if haplogroup H were considered a HG haplogroup along the inland route (besides the sea route). The only change of this assumption would be that Suppl. Fig. 19 would be replaced by Suppl. Fig. 23. We see that the main difference with Suppl. Fig. 19 is the substantial decrease of Neolithic haplogroups in region 4 (Bulgaria) in Suppl. Fig. 23. This is due to an important increase of haplogroup H (from 7% to 31%) in region 4 (Suppl. Table 7c). None of the 6 HGs from region 4 (Bulgaria) whose mt haplogroup are known display haplogroup H but, unfortunately, they are all about 45,000 yr old (Suppl. Table 5). For such an old date, we cannot assume that the haplogroup distribution was similar to that at the arrival of the Neolithic to Bulgaria (c. 8,000 yr ago). Thus more data are indeed needed before we can conclude if haplogroup H was present or not in HGs in Bulgaria when the first farmers arrived[23]. Again, this shows that at present we lack the data necessary to definitely consider haplogroup H as a Neolithic or a HG haplogroup (i.e., as essentially absent or not in HGs). Nevertheless, for the purposes of our paper, the only necessary point is the following. We have repeated the analysis above (Suppl. Fig. 20), using Suppl. Fig. 23 instead of Suppl. Fig. 19, and reached the same conclusion (also with 90% confidence level), namely that a spatial decrease similar to that of haplogroup K is not obtained by generating clines at random.

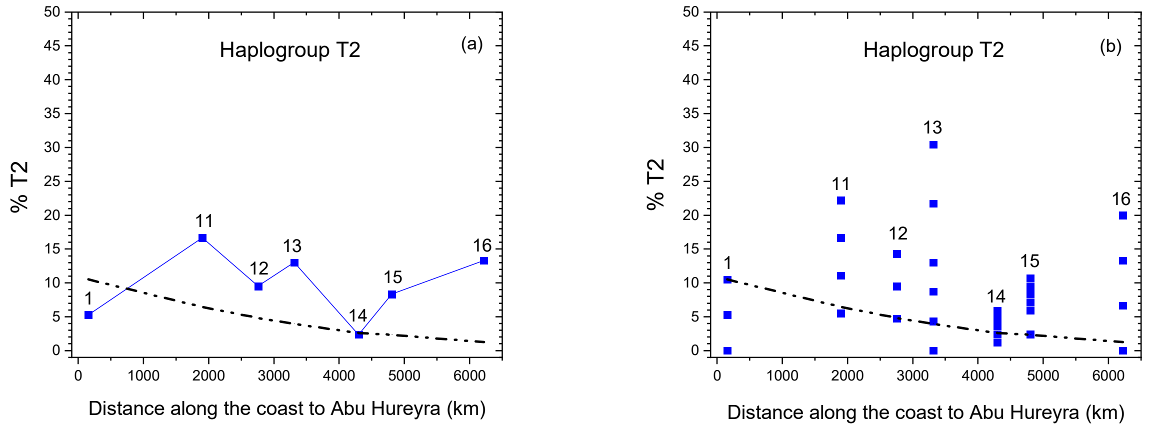


Supplementary Figure 23. Total percentage of Neolithic haplogroups (K, T2, N1a, J, HV, V, W, X, R0, L3, N, U1, N1b, U3, T1, D1/G1a1, C5, R1, T and I1) among early farmers in the 10 regions of the inland route (Suppl. Table 7c). The only difference with Suppl. Fig. 19 is that haplogroup H is not included.

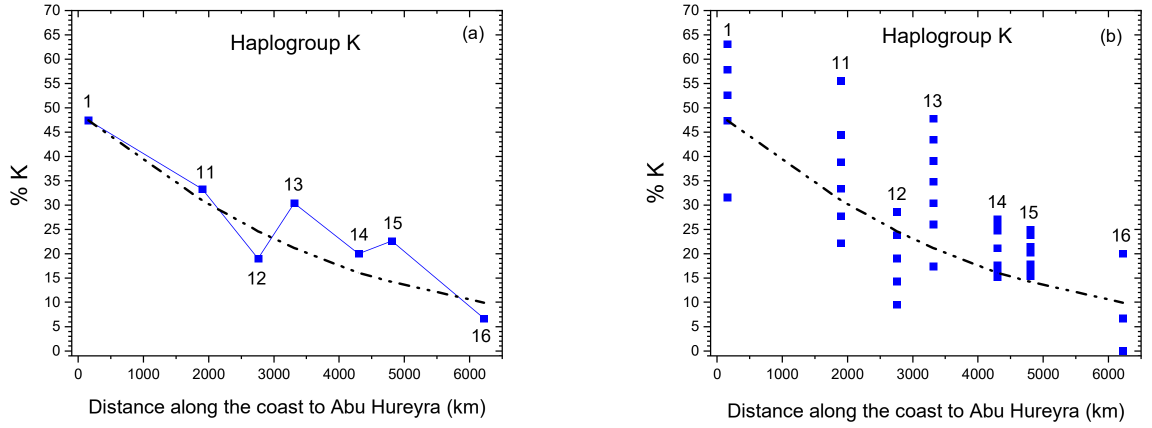
[23] Settling this question, i.e., using new data to decide if Suppl. Fig. 19 or Suppl. Fig. 23 (or another one) is realistic, will be also necessary to develop future quantitative models of HG haplogroup clines in early farmers (or equivalently, of clines for the % of all Neolithic haplogroups, since this % plus that of HG haplogroups in early farmers adds up to 100% in Suppl. Table 7). To develop such models, more complete databases of early farmers and specially of HGs are needed. Indeed, they are necessary to identify conclusively which haplogroups can be considered as Neolithic and which ones as HG haplogroups.

S11. Low-frequency haplogroups

We have considered mitochondrial haplogroup K because of two main reasons. First, it is essentially absent in HGs. This implies that we do not need data on its frequency in HGs in different regions, which would introduce additional uncertainties. The second main reason to consider haplogroup K is that it is the one with highest frequency in the region from where the Neolithic spread (region 1). In this section we explain the rationale behind this second reason by using two illustrative haplogroups (Suppl. Figs. 24-25). First of all, according to Suppl. Table 7 we know the haplogroups of 19 individuals in region 1. Of these, 9 have haplogroup K (47.4%) and all other haplogroups have low frequencies (they are present in only 1 of 2 individuals).



Supplementary Figure 24. (a) Observed frequencies of mt haplogroup T2 along the sea route (from Suppl. Table 7b). The regions are the same as for all figures on the sea route, namely 1 Northern Mesopotamia, 11 Greece and Northern Macedonia, 12 Croatia, 13 Italy, 14 Southern France, 15 Spain and 16 Portugal. (b) Frequencies obtained by bootstrap resampling 10 times with replacement of the data in each region (some trials yield the same frequency, so there are less than 10 points per region in panel b). There is less dispersion in regions 14 and 15 because they have more early farmers whose haplogroup is known (85 and 86 individuals, respectively). In contrast, regions 1, 11, 12, 13 and 16 have 19, 18, 21, 23 and 15 individuals, respectively (Suppl. Table 7). The dashed-dotted line is the same in both plots and has been obtained from our simulations for η = 0.07 and a frequency of 10.5% in region 1.



Supplementary Figure 25. (a) Observed frequencies of mt haplogroup K along the sea route (i.e., the same as the squares in Fig. 3b in the main paper). The regions are the same as for all figures on the sea route, namely 1 Northern Mesopotamia, 11 Greece and Northern Macedonia, 12 Croatia, 13 Italy, 14 Southern France, 15 Spain and 16 Portugal. (b) Frequencies obtained by bootstrap resampling 10 times with replacement of the data in each region (some trials yield the same frequency, so there are less than 10 points per region in panel b). There is less dispersion in regions 14 and 15 because they have more early farmers whose haplogroup is known (85 and 86 individuals, respectively). In contrast, regions 1, 11, 12, 13 and 16 have 19, 18, 21, 23 and 15 individuals, respectively (Suppl. Table 7). The dashed-dotted line is the same in both plots and has been obtained from our simulations for η = 0.07 and a frequency of 47.4% in region 1 (this line is also included in Fig. 3b in the main paper).

The question is if our model can be applied to haplogroups with low frequencies or not. In principle it can, but here we shall see that for low-frequency markers, the uncertainties of the data available at present are too large to reach any useful conclusion. In order to see this, we show an example of a low-frequency haplogroup in Suppl. Fig. 24a. We see that there is not apparently any clear spatial trend. Moreover any simulation from our model (the dashed-dotted line is an example) will show a decreasing cline due interbreeding, but such a cline is apparently inconsistent with the data. However, in order to illustrate the effect of the data uncertainty we have performed a bootstrap resampling with replacement 10 times for each region. This technique has been already used in Fig. 3 in the main paper (to obtain the error bars) and in Sec. S8, but it is worth to summarize it here again. For example, in region 1 there is 1 early farmer with haplogroup T2 and 18 early farmers without it. So the frequency of T2 is (1/19)·100=5.3%, as shown in Suppl. Fig. 24a. Bootstrap resampling with replacement works as follows. Imagine that we have 19 balls (one of them white and the other 18 black), we take a ball at random, record its color, and mix it again with the rest of balls. We then take a second ball, record its color, mix it with the rest, take a third ball, etc., until we have taken 19 balls. Possibly the number of white balls will be different from one, so the frequency will not be necessary be 1/19=5.3%. Supplementary Figure 24b shows the results for 10 such bootstrap runs for each region. We note that there is a huge variation, up to the point that the decreasing cline from our model (dashed-dotted black line) should not be regarded as inconsistent with the data. However, the dispersal is so large that the data are consistent with either a decreasing, increasing or uniform cline (among many others).

The case of haplogroup K is completely different, because its frequency reaches high values (about 50% in region 1). Supplementary Figure 25a shows its frequencies as estimated directly from the data. Even when the data uncertainty is taken into account (Suppl. Fig. 25b), an overall decreasing cline is clearly seen (as also proofed formally in Sec. S8). It is obvious by comparing Suppl. Fig. 25b to Suppl. Fig. 24b that the clearly decreasing cline for haplogroup K can be recognized due to the fact that this haplogroup displays a large variation in frequencies (from about 50% in region 1 to about 5% in region 16, according to Suppl. Fig. 25a). In contrast, for haplogroup T2 the frequency varies in a substantially narrower range, namely 5-15% (Suppl. Fig. 24a) and this is why the data uncertainty makes it impossible to detect any clear cline (Suppl. Fig. 24b). These two examples show very clearly that in order to apply our model to low-frequency clines we would need many more data. Indeed, as noted in the caption of Suppl. Fig. 24b, regions 14 and 15 have more data and therefore less dispersion, but it is still rather large. Therefore, with the data available at present we can only detect clear clines for markers that display high frequencies (e.g., Suppl. Fig. 25). Since according to the data there is not any haplogroup except K with a high frequency (about 40% or more) in any region (Suppl. Table 7), in practice this means that we have to use high-frequency haplogroups and the only one is haplogroup K because all other haplogroups display low frequencies (below 20%, see Supp. Tables 7b-c).

The other main justification to use haplogroup K is the fact that it is essentially absent in HGs, as already discussed at the beginning of this section.

The conclusion of this section is that with data available at present sampling effects, on their own, cause such a substantial data uncertainty that they make it impossible to detect clines for low-frequency markers.

S12. Application of our model to Y-chromosome data

Our model was developed to analyze clines of mitochondrial haplogroups. In contrast to mitochondrial DNA, the Y chromosome is present only in males. Thus before applying our model to Y-chromosome data, we have to analyze whether this difference makes it necessary to modify our equations or not. Our model has 3 populations, namely hunter-gatherers (HG), farmers with the haplogroup considered (N) and farmers without the haplogroup considered (X). In a previous paper we developed a model with 6 populations, i.e., a sub-population of males and another one of females for each of the 3 populations mentioned above (Text S11 in [3]). Such a more complicated approach could be used, but we think that it is not necessary. To see this, first we note that if we deal with only the 3 populations considered above, we have to include both males and females in each population if we want to apply the carrying capacities of farmers and HGs given in the main paper (Methods) because those carrying capacities refer to complete populations (i.e., including both males and females). Thus we define population N as male farmers with the haplogroup considered, and an equal number of female farmers. Similarly we define population X as male farmers without the haplogroup considered, and an equal number of female farmers. Finally, since we consider an haplogroup absent in HGs, we define population HG as all HGs, i.e. male HGs (all of which lacking the haplogroup considered), and an equal number of female HGs.

Equations (1), (2) and (6) in the main paper can be applied because they were derived in Ref. [9] from cultural transmission theory for complete populations (i.e., including both males and females). Equations (3)-(5) are valid because they are just definitions, and Eqs. (7)-(8) are also valid because they follow directly from them. Analogously to what we have done in the main paper (Methods), we assume that in half of the mixed matings (HN and XN) the father belongs to group N, so we can apply that 50% of the sons will be of type N and therefore Eqs. (9)-(10) also hold. Finally Eqs. (11)-(13) follow straightforwardly from the previous equations (Methods), so we conclude that the equations used in our simulations are also valid to simulate clines of Y-chromosome haplogroups (as done to obtain the curves in Fig. 5 in the main paper).

Author contributions. Secs. S1 and S6-S11: JF. Secs. S2-S5: figures by JP-L and JF, text by JF.

References

[1] Ammerman, A. J. & Cavalli-Sforza, L. L., "Measuring the rate of spread of early farming in Europe," *Man,* vol. 6, pp. 674-688, 1971.

[2] Sgaramella-Zonta, L. & Cavalli-Sforza, L. L., "Method for the detection of a genetic cline," in *Genetic structure of populations*, Honolulu, University of Hawaii Press, 1973, pp. 128-135.

[3] Isern, N., Fort, J. & Rioja, V. L., "The ancient cline of haplogroup K implies that the Neolithic transition in Europe was mainly demic," *Sci. Rep.,* vol. 7, no. 11229, 2017.

[4] Rendine, S., Piazza, A. & Cavalli-Sforza, L. L., "Simulation and separation by principal components of multiple demic expansions in Europe," *American Naturalist,* vol. 128, pp. 681-706, 1986.

[5] Barbujani, G., Sokal, R. D. & Oden, N. L., "Indo-European origins: a computer simulation test of five hypothesis," *Am. J. Phys. Anthropol.,* vol. 96, pp. 109-132, 1995.

[6] Currat, M. & Excoffier, L., "The effect of the Neolithic expansion on European molecular diversity," *Proc. Roy. Soc. B,* vol. 272, pp. 679-688, 2005.

[7] Currat, M. & Excoffier, L., "Modern humans did not admix with Neanderthals during their range expansion into Europe," *PLoS Biology,* vol. 2, pp. 2264-2274, 2002.

[8] Rasteiro, R., Bouttier, P.-A., Sousa, V. C. & Chikhi, L., "Investigating sex-biased migration during the Neolithic transition in Europe, using an explicit spatial simulation framework," *Proc. R. Soc. B,* vol. 279, pp. 2409-2416, 2012.

[9] Fort, J., "Vertical cultural transmission effects on demic front propagation: Theory and application to the Neolithic transition in Europe," *Phys Rev E,* vol. 056124, no. 83, pp. 1-10, 2011.

[10] Cavalli-Sforza, L. L. & Feldman, M. W., Cultural transmission and evolution: A quantitative approach, New Jersey: Princeton University Press, 1981.

[11] Fort, J., "Synthesis between demic and cultural diffusion in the Neolithic transition in Europe," *PNAS,* vol. 109, pp. 18669-18673, 2012.

[12] Belle, E. M. S., Landry, P.-A. & Barbujani, G., "Origins and evolution of the Europeans genome: evidence from multiple microsatellite loci," *Proc. R. Soc. B,* vol. 273, p. 1595–1602, 2006.

[13] Klopfstein, S., Currat, M. & Escoffier, L., "The fate of mutations surfing on the wave of a range expansion," *Mol. Biol. Evol.,* vol. 23, pp. 482-490, 2006.

[14] Ammerman, A. J. & Cavalli-Sforza, L. L., The neolithic transition and the genetics of populations in Europe, New Jersey: Princeton University Press, 1984.

[15] Fix, A. G., "Gene frequency clines in Europe: demic diffusion or natural selection?," *J. Roy. Anthrop. Inst.,* vol. 2, pp. 625-643, 1996.

[16] van der Walt J M, et al., "Mitochondrial Polymorphisms Significantly Reduce the Risk of Parkinson Disease," *Am. J. Hum. Genet.,* vol. 72, p. 804–811, 2003.

[17] Chang X, et al., "Mitochondrial DNA haplogroups and risk of attention deficit and hyperactivity disorder in European Americans," *Transl. Psychiatry,* vol. 10, no. 370, 2020.

[18] Jensen, T. Z. T., et al., "A 5700 year-old human genome and oral microbiome from chewed birch pitch," *Nature Commun.,* vol. 10, no. 5520, 2020.

[19] Silva, N. M., et al., "Ancient mitochondrial diversity reveals population homogeneity in Neolithic Greece and identifies population dynamics along the Danubian expansion axis," *Sci. Rep.,* vol. 12, no. 13474, 2022.

[20] Lazaridis, I. et al., "Genomic insights into the origin of farming in the ancient Near East," *Nature,* vol. 536, pp. 419-424, 2016.

[21] Lazaridis, I., et al., "Ancient DNA from Mesopotamia suggests distinct Pre-Pottery and Pottery Neolithic migrations into Anatolia," *Science,* vol. 377, p. 982–987, 2022.

[22] Mathieson, I., et al., "The genomic history of southeastern Europe," *Nature,* vol. 555, pp. 197-203, 2018.

[23] Shennan, S., The first farmers of Europe. An evolutionary perspective, Cambridge: Cambridge University Press, 2018.

[24] Gurova, M, & Bonsall, C., "'Pre-Neolithic' in southeast Europe: a Bulgarian perspective," *Docum. Praehist. 41, 95-109,* 2014.

[25] Brami M., et al., "Was the Fishing Village of Lepenski Vir Built by Europe's first farmers?," *J. World Prehist.,* vol. 35, pp. 109-133, 2022.

[26] Brace, S. et al., "Ancient genomes indicate population replacement in Early Neolithic Britain," *Nature Ecology & Evolution,* vol. 3, p. 765–771, 2019.

[27] Sánchez-Quinto, F., Malmström, H., Fraser, M., Girdland-Flink, L., Svensson, E. M., Simoes, G. et al., "Megalithic tombs in western and northern Neolithic Europe were linked to a kindred society," *Proc. Natl. Acad. Sci. U.S.,* vol. 116, pp. 9469-9474, 2019.

[28] Cassidy L. M. et al., "A dynastic elite in monumental Neolithic society," *Nature,* vol. 582, pp. 384-388, 2020.

[29] Cassidy, L. M., et al., "Neolithic and Bronze Age migration to Ireland and establishment of the insular Atlantic genome," *PNAS,* vol. 130, pp. 368-373, 2016.

[30] Olalde, I. et al., "The Beaker phenomenon and the genomic transformation of northwest Europe," *Nature,* vol. 555, p. 190–196, 2018.

[31] Sheib, C.L., et al., "East Anglian early Neolithic monument burial linked to contemporary Megaliths," *Ann. Hum. Biol.,* vol. 46, pp. 145-149, 2019.

[32] Fowler, C., "A high-resolution picture of kinship practices in an Early Neolithic tomb," *Nature,* vol. 601, pp. 584-587, 2022.

[33] Allentoft, M. E., et al., "Population genomics of post-glacial western Eurasia," *Nature,* vol. 625, pp. 301-310, 2024.

[34] Jones, E. R., et al., "The Neolithic transition in the Baltic was not driven by admixture with early european farmers," *Curr. Biol.,* vol. 27, pp. 576-582, 2017.

[35] Fort, J., Pujol, T. & vander Linden, M., "Modelling the Neolithic transition in the Near East and Europe," *Am. Antiq.,* vol. 77, pp. 203-220, 2012.

[36] Baird, D., et al., "Agricultural origins on the Anatolian plateau," *PNAS,* vol. 115, pp. E3077-E3086, 2018.

[37] Smith, A., Oechsner, A., Rowley-Cowny, P. & Moore, A. M. T., "Epipalaeolithic animal tending to Neolithic herding at Abu Hureyra, Syria (12,800-7,800 cal BP): deciphering dung spherulites," *PLoS One,* vol. 17, no. e0272947, 2022.

[38] Isern, N., Zilhao, J., Fort, J. & Ammerman, A. J., "Modeling the role of voyaging in the coastal spread of the Early Neolithic in the West Mediterranean," *PNAS,* vol. 114, pp. 897-

902, 2017.

[39] Fort, J., "Dispersal distances and cultural effects in the spread of the Neolithic along the northern Mediterranean coast," *Archaeol. Anthropol. Sci.,* vol. 14, no. 153, 2022.

[40] Fort, J., Pérez-Losada, J. & Isern, N., "Fronts from integrodifference equations and persistence effects on the Neolithic transition," *Phys. Rev. E,* vol. 76, no. 031913, 2007.

[41] Tkachenko, M., Weissmann, J. D., Petersen, W. P., Lake, G., Zollikofer, C. P. R., Callegari, S., "Individual-based modelling of population growth and diffusion in discrete time," *PLoS One,* vol. 12, no. e0176101, 2017.

[42] Wobst, M., "Boundary contitions for Paleolithic social systems: a simulation approach," *Am. Antiq.,* vol. 39, pp. 147-178, 1974.

[43] Zimmerman, A., Hilpert, J. & Wendt, K. P., "Estimations of population density for selected periods between the Neolithic and AD 1800," *Human Biology,* vol. 81, pp. 357-380, 2009.

[44] "https://desktop.arcgis.com/es/arcmap/10.3/guide-books/map-projections/about-map-projections.htm".

[45] De Smith, M. J., Goodchild, M. F. & Longley, P. A., Geospatial analysis: a comprehensive guide to principles, techniques and software tools, 8th. ed., Whinchelsea: The Whinchelsea press, 2018. Available at http://www.spatialanalysisonline.com (accessed 2020 October 15).

[46] Zilhao, J., "Radiocarbon evidence for maritime pioneer colonization at the origins of farming in west Mediterranean Europe," *PNAS,* vol. 98, pp. 14180-14185, 2001.

[47] Pinhasi, R., Fort, J. & Ammerman, A. J., "Tracing the origin and spread of agriculture in Europe," *PLoS Biol.,* vol. 3, no. e410, 2005.

[48] Fort, J., Jana, D. & Humet, J. M., "Multidelayed random walks: theory and application to the Neolithic transition in Europe," *Phys. Rev. E,* vol. 70, no. 031913, 2004.

[49] Fort, J. & Pujol, T., "Progress in front propagation research," *Rep. Prog. Phys.,* vol. 71, no. 086001, 2008.

[50] Isern, N., Fort, J. & Pérez-Losada, J., "Realistic dispersion kernels applied to cohabitation reaction-dispersion equations," *J. Stat. Mech. Theor. Exp.,* vol. 10, p. P10012, 2008.

[51] Müller, J. & Diachenko, A., "Tracing long-term demographic changes: the issue of spatial scales," *PLoS One,* vol. 14, no. e0208739, 2019.

[52] Steele, J. M., Adams, J. & Sluckin, T. J., "Modeling Paleoindian dispersals," *World Archaeology,* vol. 30, pp. 286-305, 1998.

[53] Legendre, P. & Legendre, L., Numerical Ecology, Amsterdam: Elsevier, 2012.

[54] Oden, N. L., "Assessing the significance of a spatial correlogram," *Geogr. Anal.,* vol. 16, pp. 1-16, 1984.

[55] https://www.movable-type.co.uk/scripts/latlong.html.

[56] Rosenberg, M. S. & Anderson, C. D., "PASSaGE: pattern analysis, spatial statistics and geographic exegesis. Version 2," *Meth. Ecol. Evol.,* vol. 2, pp. 229-232, 2011.

[57] Legendre, P. & Fortin, M.-J., "Spatial pattern and ecological analysis," *Vegetatio,* vol. 80, pp. 107-138, 1989.

[58] Brandt, G., Haak, W., Adler, C. J., Roth, C., Szécsényi-Nagy, A., Karimnia, S., et al., "Ancient DNA reveals key stages in the formation of central European mitochondrial genetic diversity," *Science,* vol. 342, pp. 257-261, 2013.

[59] Fu, Q., Rudan, P., Pääbo, S. & Krause, J., "Complete Mitochondrial Genomes Reveal Neolithic Expansion into Europe," *PLoS One,* vol. 7, no. e32473, 2012.

[60] Brotherton, P., Haak, W., Templeton, J., Brandt, G., Soubrier, J., Adler, C. J., et al., "Neolithic mitochondrial haplogroup H genomes and the genetic origins of Europeans," *Nature Comm.,* vol. 4, no. 1764, 2013.

[61] Bramanti, B., et al., "Genetic discontinuity between local hunter-gatherers and central Europe's first farmers," *Science,* vol. 326, pp. 137-140, 2009.

[62] Eriksson, G., et al., "Same island, different diet: cultural evolution on food practice on Öland, Sweden, from the Mesolithic to the Roman period," *J. Anthropol. Archaeol.,* vol. 27, pp. 520-543, 2008.

[63] Malmer, J. P., The Neolithic of south Sweden: TRB, GRK, and STR, Stockholm: Royal Swedish Academy of Letters, History and Antiquities, 2002.

[64] Skoglund, P., et al., "Genomic diversity and admixture differs for stone-age Scandinavian foragers and farmers," *Science,* vol. 344, pp. 747-750, 2014.

[65] Hofmanová, Z., et al., "Early farmers from across Europe directly descended from Neolithic Aegeans," *PNAS,* vol. 113, pp. 6886-6891, 2016.

[66] González-Fortes, G., et al., "Paleogenomic evidence for multi-generational mixing between Neolithic farmers and Mesolithic hunter-gatherers in the lower Danube basin," *Curr. Biol.,* vol. 27, pp. 1801-1810, 2017.

[67] Betti, L., et al., "Climate shaped how Neolithic farmers and European hunter-gatherers interacted after a major slowdown from 6,100 BCE to 4,500 BCE," *Nature human behaviour,* vol. 4, pp. 1004-1010, 2020.

[68] Feldman, M., et al., "Late Pleistocene human genome suggests a local origin for the first farmers of central Anatolia," *Nature Commun.,* vol. 10, no. 1218, 2019.

[69] Malmström, H., et al., "Ancient DNA reveals lack of continuity between Neolithic hunter-gatherers and contemporary Scandivians," *Curr. Biol.,* vol. 19, pp. 1758-1762, 2009.

[70] Malmström, H., et al., "Ancient mitochondrial DNA fron the northern fringe of the Neolithic farming expansion in Europe sheds light on the dispersion process," *Phil. Trans. R. Soc. B,* vol. 370, no. 20130373, 2015.