Kinship structures create persistent channels for language transmission

Lansing et al.

Supporting Information (SI)

Genetic data. The sample dataset comprises 982 men from the eastern Indonesian islands of Sumba ($n = 505$) and Timor ($n = 477$), with associated genetic, cultural and linguistic information. Of the 25 villages in this study, 14 are on Sumba and 11 on Timor. All Sumba villages are patrilocal. Of the 11 Timor villages, two follow patrilocal kinship practice (Fatukati and Umaklaran), while the remainder are matrilocal. The diversity of mtDNA and Y chromosome haplogroups found in each sampled village is shown in Fig. S1. The genetic trees were constructed using the method described below, which is freely available as an open source R application at https://carpntrees.shinyapps.io/shinyapp/ [last accessed August 2017].

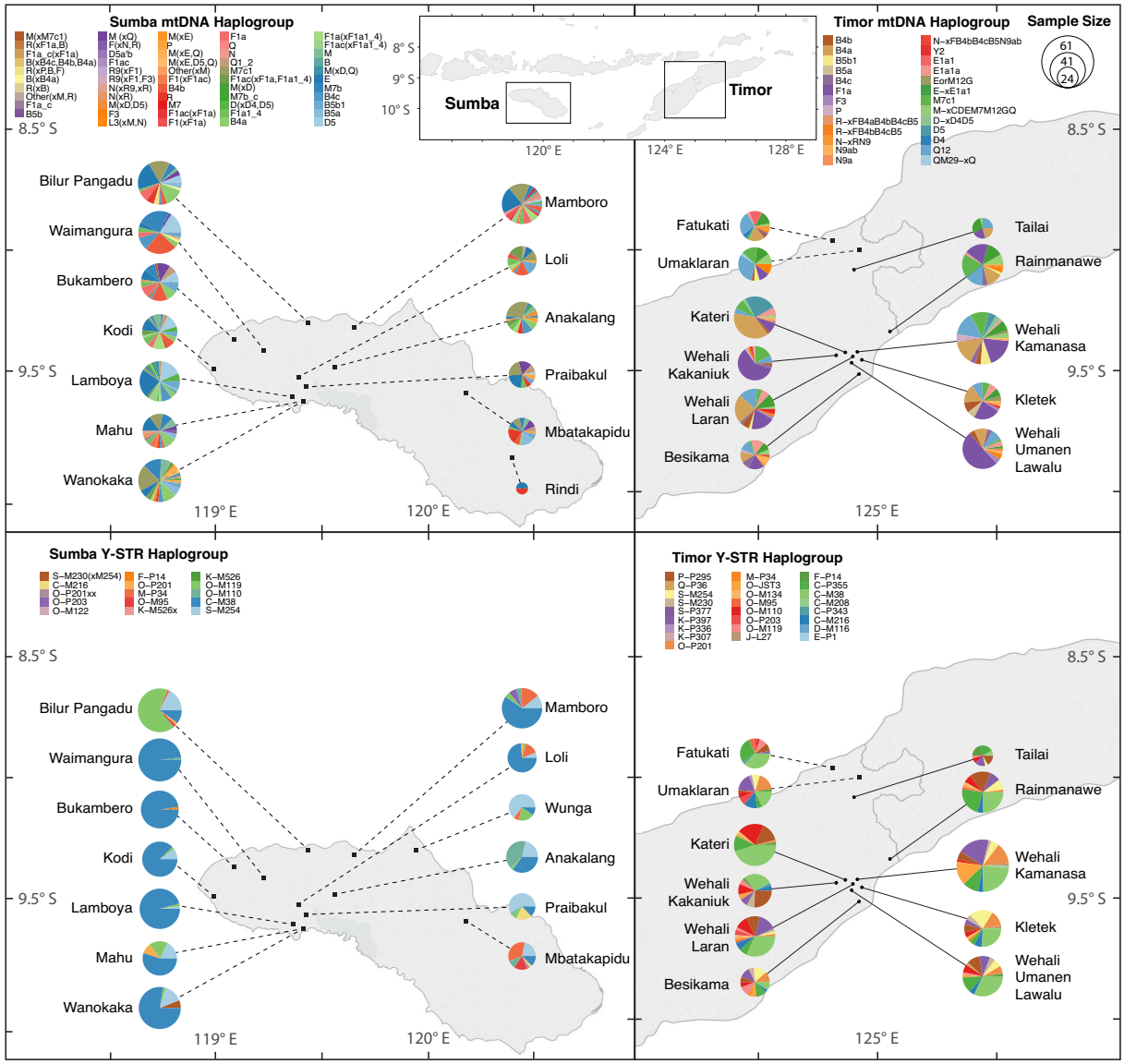

Fig. S1. Genetic diversity on Sumba and Timor. Diversity of mtDNA haplogroups (top) and Y chromosome haplogroups (bottom) in villages on Sumba (left) and Timor (right). ▪ indicates patrilocal villages; • indicates matrilocal villages. Pie charts, scaled by sample size, show the distribution of haplogroups within each village.

mtDNA tree and distances. A tree of female lineages was constructed using mitochondrial DNA, with a combination of HVS-I sequences (540 base pairs) and SNP-defined haplogroup information. While the sequence data show recent variation between individuals, the haplogroups (determined by hierarchically genotyped SNPs) permit greater resolution of deeper parts of the phylogeny. By combining both information sources, a phylogenetic tree can be built that represents the true phylogeny far better than one generated from a single data source alone.

The tree building approach involved several steps. First, a tree was built using the haplogroup of each individual, with the typology constrained to match the consensus haplogroup tree from the Phylotree project [1]. This tree contains leaves representing individuals that share the same haplogroup. These leaves were then further refined using a neighbor-joining algorithm on the HVS-I sequences, using the most closely related haplogroup as an outgroup. Thus, we iteratively generate a final tree in which the deep nodes are determined by SNP-defined haplogroups and recent nodes by HVS-I variation. All individuals with the same haplogroup and HVS-I sequence are initially collapsed to speed up the calculations, later being separated at the tips after the tree reconstruction was complete. The final tree is shown in Fig. S2. Subtrees (e.g., for patrilocal villages on Timor, etc.) were extracted from this master tree.

The pairwise distance matrix used for IM model fitting and language transmission analysis was built using the number of pairwise differences in HVS-I sequences between each pair of individuals.

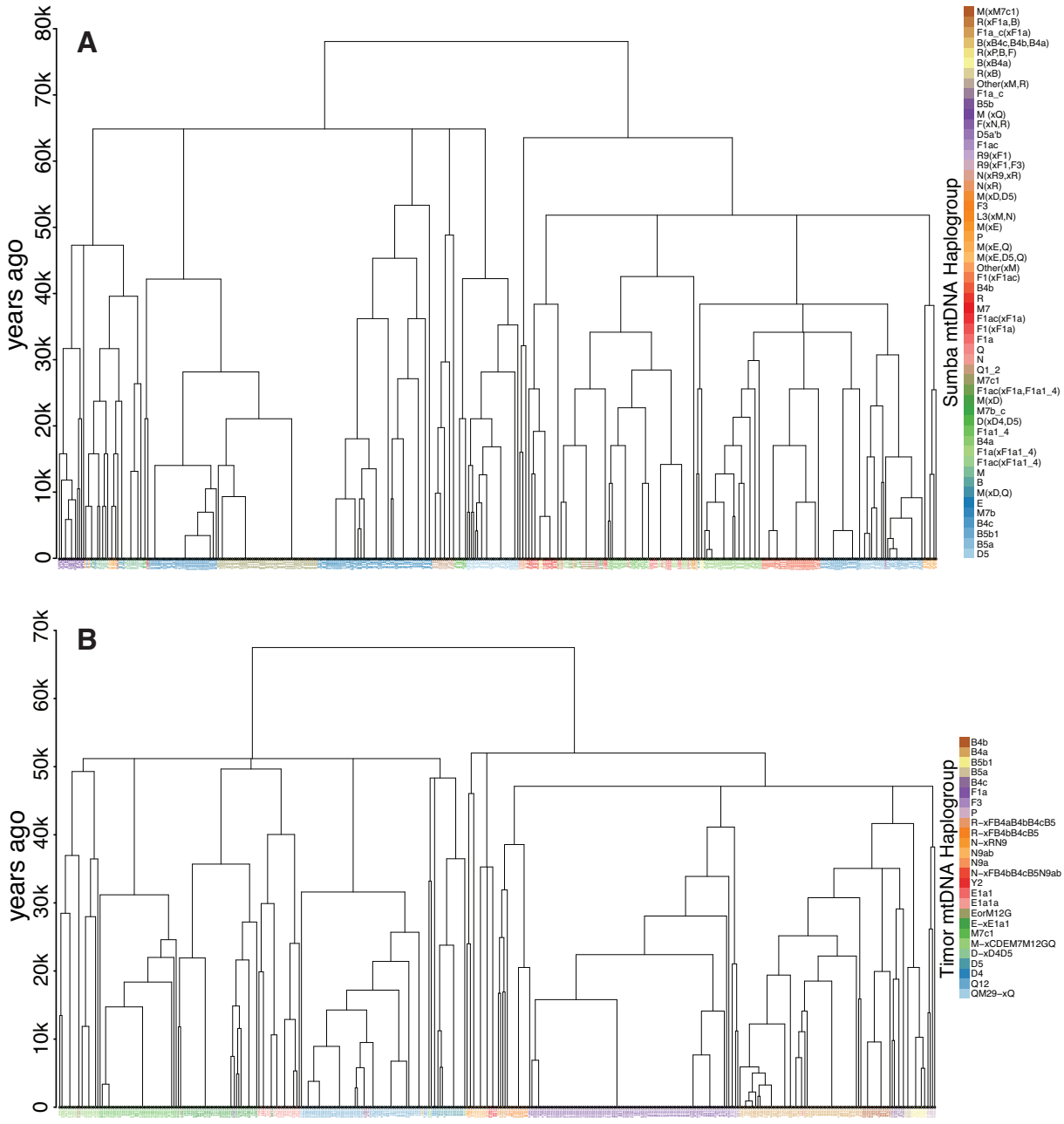

Fig. S2. mtDNA phylogeny. The female lineages of sampled men from (A) Sumba and (B) Timor, colored according to their mtDNA haplogroup.

Y chromosome tree and distances. The tree of male lineages was reconstructed using both Y chromosome haplogroup information and Y-STR markers. Using a similar approach as for mtDNA, both information sources were combined by first building a haplogroup tree, and then refining recent lineage branching using Y-STR variation. The haplogroup tree was constrained to the topology of the Y-DNA Haplogroup Tree defined by the International Society of Genetic Genealogy [2], with improvements as per Karafet et al. [3] to incorporate new markers specific to this geographical region. To refine relationships at the tips, Bruvo's distance [4] was calculated between each pair of individuals based on Y-STR variation. Leaves were then further refined using a neighbor-joining algorithm on the Bruvo distance of the Y-STRs. As before, all genetically identical individuals were initially collapsed and later separated out after the overall tree structure had been determined. The final tree is shown in Fig. S3. Individual population trees were extracted from this master tree.

The pairwise distance matrix used for IM model fitting and language transmission analysis was built using the Bruvo distance [4] between all pairs of individuals, as defined by their Y-STR diversity.

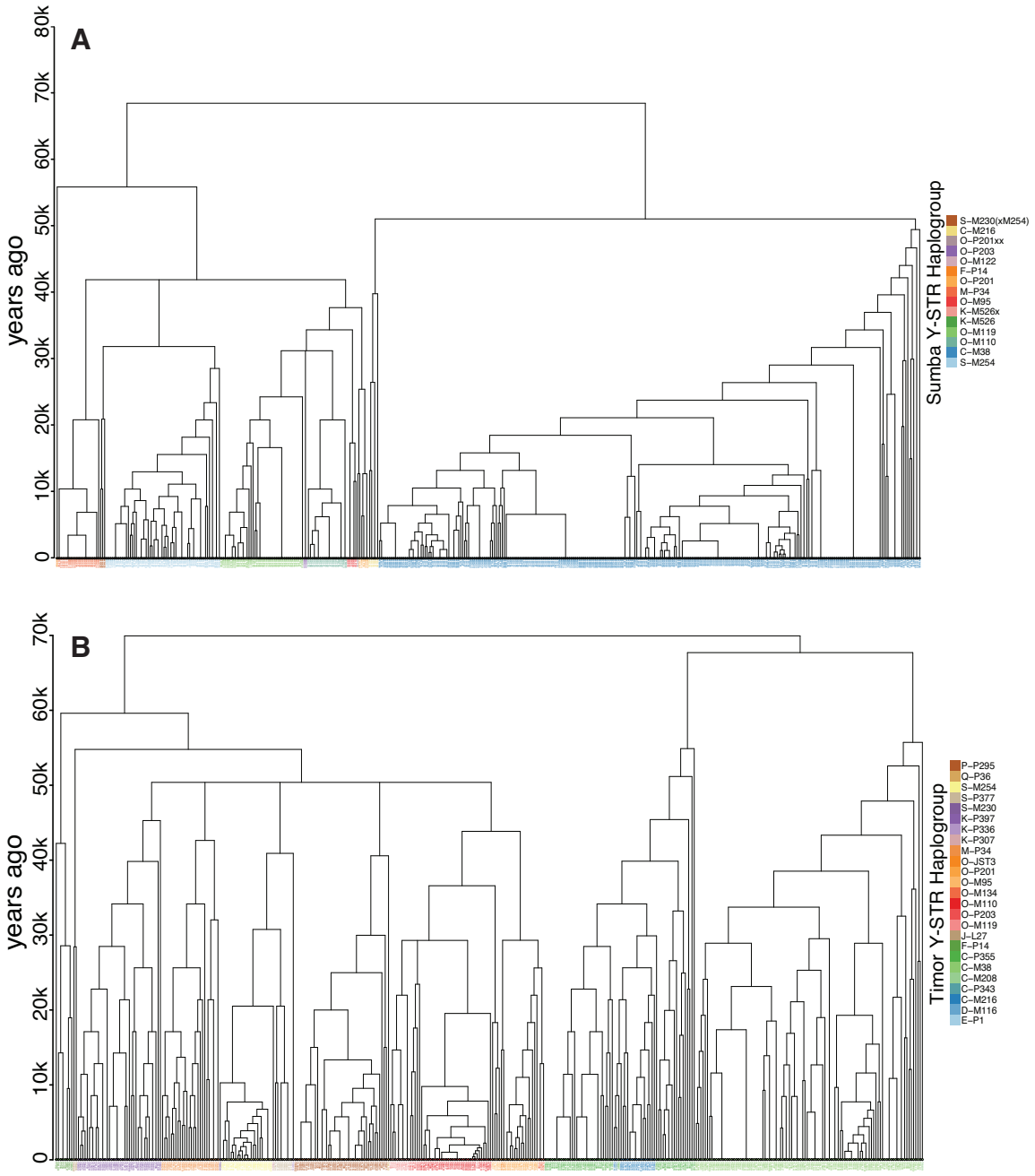

Fig. S3. Y chromosome phylogeny. The male lineages of sampled men from (A) Sumba and (B) Timor, colored according to their Y chromosome haplogroup.

Molecular dating. Previously dated haplogroups with confidence intervals were used to calibrate the trees. For the mtDNA phylogeny, three time points were used, as determined from whole mtDNA genome sequences by Fu et al. [5]. For the Y chromosome phylogeny, five time points were used, taken from Karafet et al. [3]. Molecular dating of both trees using these calibration points was performed using the *chronos* function in the R package *ape* [6].

Language data and tree. The linguistic data were organized and classified according to the principles of the traditional comparative method. As an independent check, the resulting language phylogenies for Sumba and Timor were compared against the language relationships given in Ethnologue [7]. The language phylogenies were then calibrated using penalized likelihood and a relaxed substitution rate model to estimate internal branching times [8]. The language tree for Sumba (Fig. S4A) was constructed by setting the root to 4,085 years before present, the date inferred by Xu *et al.* [9] for the arrival of Austronesian speakers on Sumba. The language tree for Timor (Fig. S4B) was constructed using two much less well supported priors: i) the Timorese Austronesian languages were set to split 5,000 years ago; and ii) the languages spoken on Timor were assumed to have a coalescent date no earlier than 10,000 years ago. Note that the depth of the Austronesian/non-Austronesian split is not known, and is purposely given an arbitrary date. The coalescence age of this external branch has little effect on the analyses.

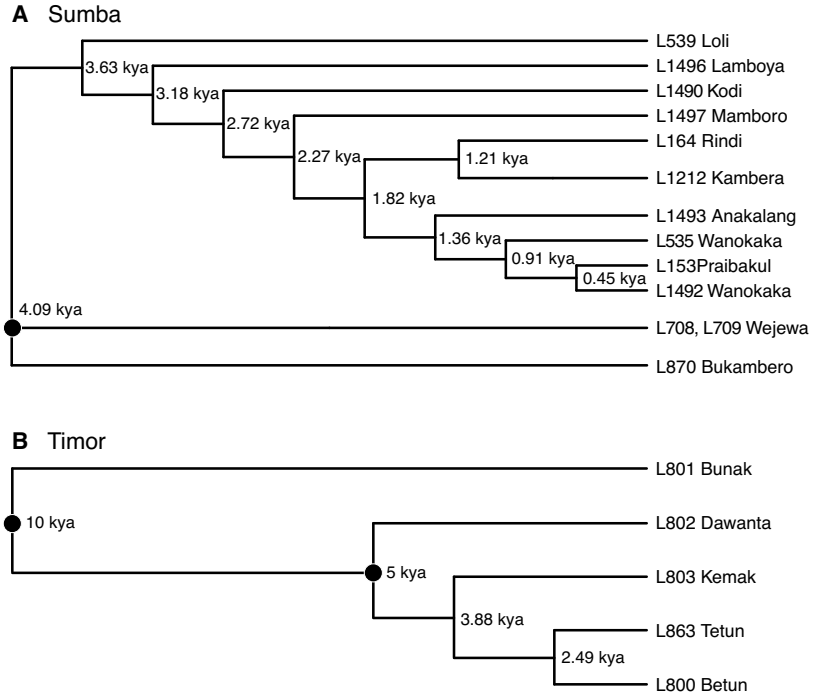

Fig. S4. Language phylogenies for (A) Sumba and (B) Timor. • indicates calibration points.

Isolation with Migration (IM) model. An Isolation with Migration model was used to capture the impact of sex-biased movements on genetic diversity [10]. The IM model describes a single panmictic population of size aN that splits into n subpopulations of size N at time $2N\tau$ generations in the past. Migration occurs at a rate m between subpopulations. To assess the effects of postmarital residence practices on mtDNA and the Y chromosome, we distinguish between the migration rates of women (m_{female}, for mtDNA) and men (m_{male}, for the Y chromosome), and run the model separately for each group. Both runs have the same values of N, n, a, τ and mutation rate μ, but can have different migration rates m_{female} and m_{male}. For example, matrilocal kinship systems lead to greater migration of men than women, such that $m_{\text{male}} > m_{\text{female}}$. The converse is true for patrilocal systems. A cultural preference for endogamy is reflected by low migration rates for both m_{female} and m_{male}.

Using a scaled (haploid) migration rate $M = 2Nm$ and $\theta = 2N\mu$, with mutation rate μ in mutations per generation for the locus, the probability that the number of nucleotide differences between a pair of individuals S_j is k is given by

$$P(S_j = k) = \sum_{r=1}^{2} A_{jr}\left(\frac{\lambda_r \theta^k}{(\lambda_r+\theta)^{k+1}}\left(1 - e^{-\tau(\lambda_r+\theta)}\sum_{l=0}^{k}\frac{(\lambda_r+\theta)^l \tau^l}{l!}\right) + \frac{e^{-\tau(\lambda_r+\theta)}(a\theta)^k}{(1+a\theta)^{k+1}}\sum_{l=0}^{k}\frac{\left(\frac{1}{a}+\theta\right)^l \tau^l}{l!}\right) \quad [1]$$

where $j = 0$ and $j = 1$ correspond to the cases of two samples being from the same or different villages, respectively. The remaining sub-equations are defined as

$$\lambda_1 = \frac{nM + n - 1 - \sqrt{D}}{2(n-1)}$$
$$\lambda_2 = \frac{nM + n - 1 + \sqrt{D}}{2(n-1)}, \quad [2]$$

$$D = (nM + n - 1)^2 - 4\,(n-1)\,M, \quad [3]$$

$$A_{01} = \frac{\lambda_2 - 1}{\lambda_2 - \lambda_1}$$
$$A_{02} = \frac{1 - \lambda_1}{\lambda_2 - \lambda_1}$$
$$A_{11} = \frac{\lambda_2}{\lambda_2 - \lambda_1}$$
$$A_{02} = \frac{-\lambda_1}{\lambda_2 - \lambda_1}. \quad [4]$$

This model can be used to make genetic predictions regarding the distribution of genetic differences in mtDNA and in the Y chromosome given sex-biased migration. To do so, we calculate the equations twice, applying a migration rate of $M = M_{\text{male}} = 2Nm_{\text{male}}$ to explore Y chromosome differences and $M = M_{\text{female}} = 2Nm_{\text{female}}$ to explore mtDNA difference, while keeping other parameters constant. We take this approach when exploring the phase space of the model, see Fig. S5.

It is important to be aware that such a model is a considerable simplification of the complex migration patterns generated by kinship systems. For instance, it assumes that i) movements of mtDNA and Y chromosomes are independent; ii) the effective population size N and demographic history, as described by n, τ and a, of the mtDNA and Y chromosome are the same (in practice, N is often lower for men due to their higher reproductive variance); and iii) standard coalescent assumptions hold, such as a relatively small sample size compared to the total population size, and exchangeability among sampled individuals. A further important assumption is that the measurement of pairwise differences between two individuals in a sample is independent of the number of pairwise differences between other individuals in the sample. This final point is relevant to our Approximate Bayesian Computation (ABC) modeling especially, and is elaborated on below.

Despite these limitations, this non-equilibrium model is complex enough to incorporate many demographic events that are expected to be reflected in the data. It closely fits the distribution of pairwise differences for distant relatives observed in the Sumba and Timor mtDNA and Y chromosome data sets (see main text Fig. 3), as well as differences between these loci at smaller genetic distances. We are especially interested in the quality of fitting at smaller genetic distances because these better reflect the impact of kinship systems over recent genetic history. While we do not suggest that the fitted parameters correspond directly to real-world properties of these populations, they do capture patterns that provide a qualitative indication of whether matrilocality or patrilocality is a likely cause of the observed data.

Parameter space of the IM model within and between villages. The mtDNA and Y chromosome pairwise distances within and between subpopulations are shown in Fig. S5, with other parameters $N = 200$, $n = 100$, $\tau = 5.0$ and $a = 2.0$ and m_{male} and m_{female} scaled between 0 and 0.05. The Kullback-Leibler divergence for between-village comparisons (Fig. S5B) is almost the same for all migration rate values. Differences in the quality of the model fit, as quantified by the Kullback-Leibler divergence, only become apparent when looking at within-village comparisons (Fig. S5A). In other words, it is only easy to observe kinship systems through differences in mtDNA and Y chromosome pairwise distances within subpopulations. This is an important future consideration when designing methods to detect the signature of kinship practices, and is leveraged in our ABC method.

Fitting the IM model using Approximate Bayesian Computation (ABC). ABC is a method to assess how much the observed data supports different parameter values in a model. The pairwise Bruvo distances [4] of the Y chromosome (values in range [0,1]) were first scaled such that the average observed distance was the same as the mtDNA pairwise differences. Given that the time depth of the Y chromosome and mtDNA trees are broadly similar for our samples, we consider this operation equivalent to 'scaling' the mutation rate of the Y chromosome and mtDNA such that, for the purposes of modeling, they can be considered equal (μ = per site mutation rate of 1.67×10^{-7} × generation time of 25 years × 540 mtDNA sites = 0.002214). The distribution of mtDNA pairwise distances and scaled Y chromosome Bruvo's distances are similar. After obtaining the observed distributions, we proceeded with the ABC iterations.

We use full coalescent simulations in our ABC model fitting, rather than simply sampling from the theoretical distribution of pairwise distances (Eq. (1)). This is slower but allows us to properly account for correlations in pairwise distances between samples (e.g., see [11]). The coalescent simulations were performed using the coalescent program *msms* [12] and implemented so as to replicate the IM model explored in Wilkinson-Herbots (2008) and described above.

In the simple form of ABC used, an iteration consists of i) proposing a combination of parameter values according to the parameter prior distributions; ii) using *msms* to simulate samples given these parameter values, both for the female (with migration rate m_{female}) and male (with migration rate m_{male}) lineages; iii) calculating the average pairwise distance between

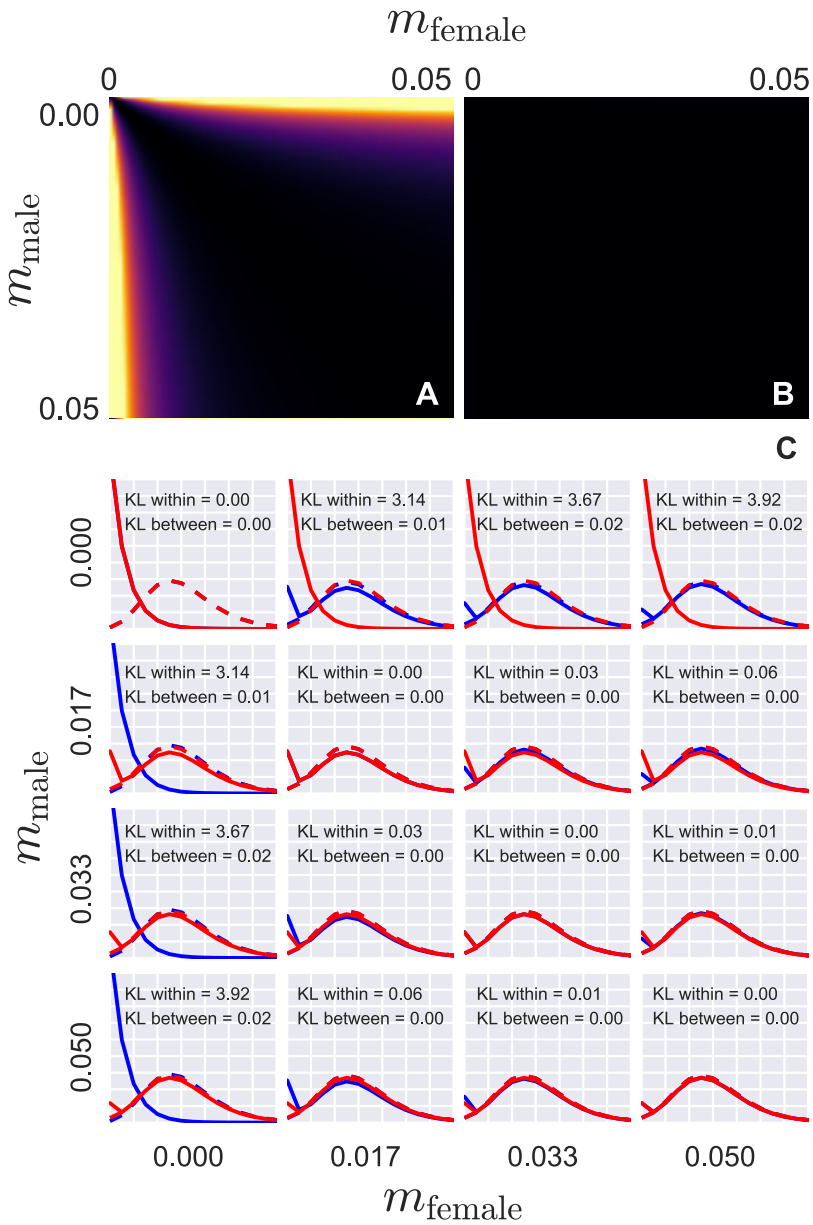

Fig. S5. Migration rate phase space. (A) Kullback-Leibler divergence of pairwise distance distributions (KL(Y|mtDNA) + KL(mtDNA|Y)) within subpopulations. (B) Kullback-Leibler divergence of pairwise distance distributions between subpopulations. (C) Example plots of mtDNA (blue) and Y chromosome (red) pairwise distance distributions within (solid lines) and between (dashed lines) subpopulations, focusing on the low-difference regime.

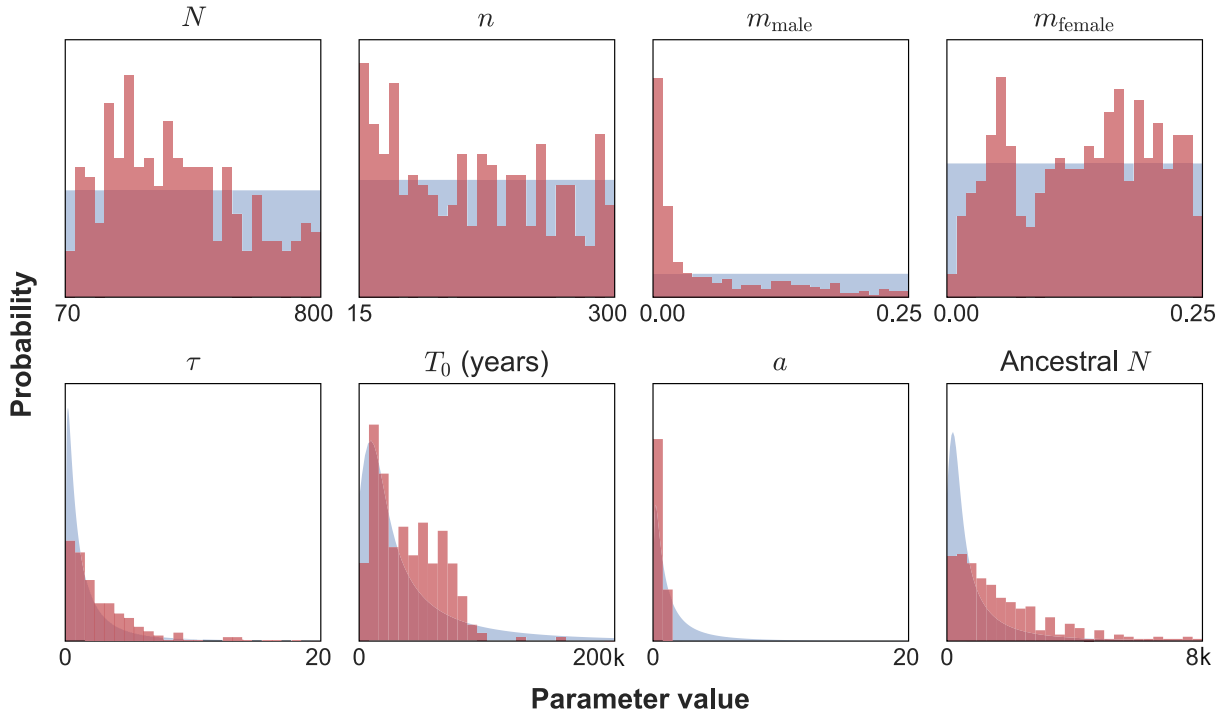

Fig. S6. Posterior distributions of IM model fitting for Sumba data, based on 300 best-fitting parameter values from three million samples of the prior distributions.

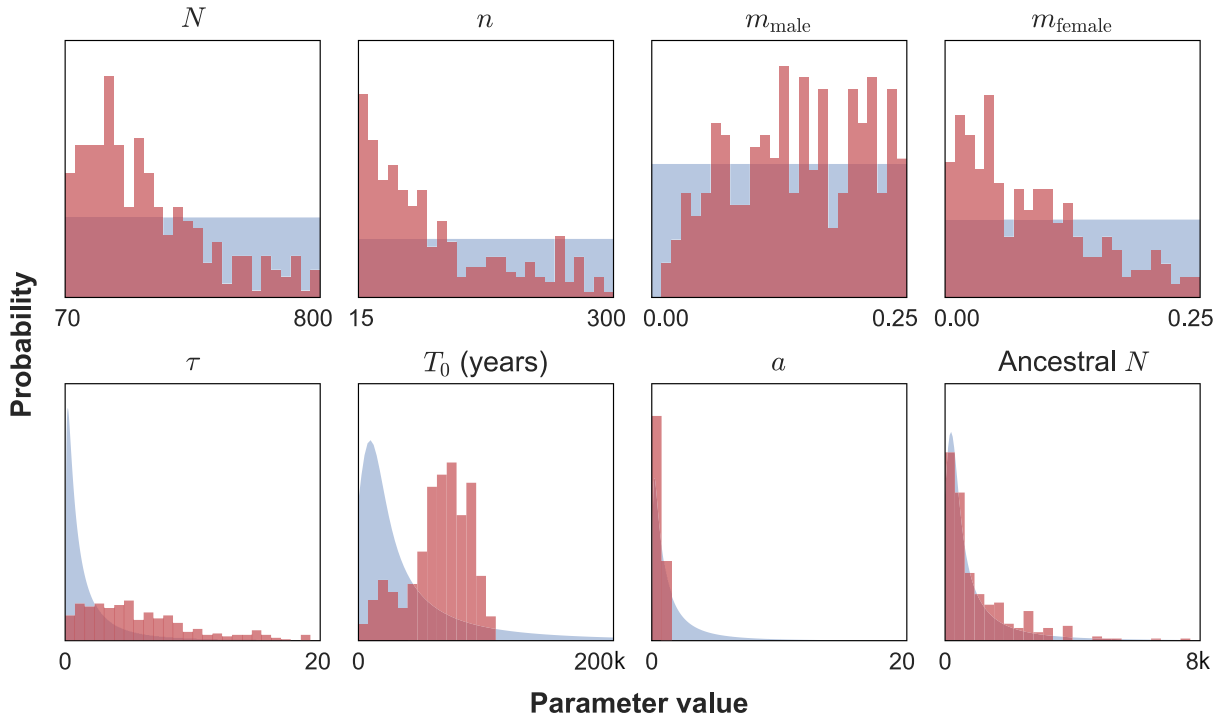

Fig. S7. Posterior distributions of IM model fitting for Timor data, based on 300 best-fitting parameter values from three million samples of the prior distributions.

samples, using only within-village comparisons (which are more informative regarding locality, see Fig. S5); and iv) calculating the distance metric (Kullback-Leibler divergence) by summing the KL-divergence of the simulated and observed mtDNA and Y chromosome distance distributions. For N, n and the migration rates we used uniform priors with bounds $70 \leqslant N \leqslant 80$, $15 \leqslant n \leqslant 300$, $0 \leqslant m_{\text{male}} \leqslant 0.25$, $0 \leqslant m_{\text{female}} \leqslant 0.25$. For τ and a, we used lognormal distributions with mean 1 and standard deviation 1.25, allowing us to sample large values of the derived parameters $T_0 = 2.0 \times 25 \times N \times \tau$ and $N_{\text{ancestral}} = a \times N$ while weighting sampling toward more plausible smaller parameter values. The number of individuals sampled from each village in the model was the same as the number of individuals in our data. This procedure was iterated three million times, keeping the parameters that generate a model output with the lowest distance from the data as being representative of the posterior distributions. The ABC posterior distributions of the IM model used for main text Fig. 3 are shown in Figs. S6 and Figs. S7. The prior distribution of T_0 and ancestral N are calculated from the uniform and lognormal priors of other parameters.

Kinship and language transmission in matrilocal Timor. In Fig. 4 of the main text, we show the genetic distances of the villages in Timor, including the 9 matrilocal villages and 2 patrilocal villages. In Fig. S8, only the genetic distances of the matrilocal villages on Timor are shown. The changes are minor when compared to Fig. 4B, D : the 'p' cluster does not disappear, and considering only matrilocal villages results in only a slight difference in the kernel density of pairwise genetic distances. We interpret this result as reflecting the relatively small number of pairwise patrilocal comparisons within the total sample. The presence of this weak 'p' cluster may be related to the high proportion of effectively ambilocal marriages in the closely integrated villages of the centuries-old Wehali village cluster (see Therik [13], but also noted in our informal surveys) with weak representation of Y-chromosome-specific drift in a minority of samples.

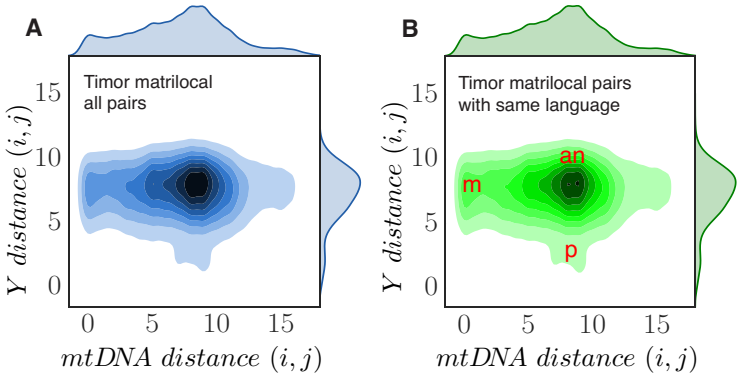

Fig. S8. Genetic distances on matrilocal villages of Timor, between all pairs of individuals (A) and only individuals who speak a common language (B). Considering only matrilocal villages in Timor does not significantly change pairwise genetic distances, and still show matrilocality (m), ambi- or neolocality (an), with a small patrilocal cluster (p) related to a high proportion of effectively ambilocal marriages.

Analysis of linguistic and genetic distances. In Fig. S9, we compare genetic distances along both matrilines and patrilines with linguistic distances, $1 - d_l(i, j)$, defined as

$$1 - d_l(i, j) = 1 - cos(\mathbf{l_i}, \mathbf{l_j}), \quad [5]$$

where $\mathbf{l_i}$ is the language vector of individual i and $\mathbf{l_j}$ is the language vector of individual j. A pair of individuals who speak the same set of languages will have $1 - d_l(i, j) = 0$. If they both speak at least one language in common, but one or both of them speak other languages as well, then $0 < 1 - d_l(i, j) < 1$. If they do not speak any language in common, $1 - d_l(i, j) = 1$.

On Sumba, we see the signal of monolinguality as pairs of individuals that either share all of their languages or do not share a language at all (Fig. S9A-B). The few pairs of individuals that speak the same language form a cluster with a mtDNA genetic distance of ~8 (Fig. S9A). On their Y chromosome (Fig. S9B), pair of individuals that speak the same languages are either closely related (~0) or distantly related to some degree (~5–10).

Multilinguality is evident in Timor (Fig. S9C-D) where there are pairs of individuals that share some, but not all, of their languages. Most individuals speak the same languages (for instance, nearly everyone speaks Upper Tetun) and they are closely related on their mtDNA (Fig. S9C), but have only distant relatedness on the Y chromosome (Fig. S9D).

Co-phylogeny of genes and languages. The co-phylogeny method has often been applied in functional ecology to find links between species traits and environmental variables, thus explaining observed biological processes in ecosystems [14]. More recently, similar methods have been applied in the context of host-parasite co-evolution [15] and cultural traits [16]. The evolution of language is analogous to parasites carried by their human speakers, whose genes evolve in a parallel process. Language traits evolve as the environmental and genetic conditions of the speech communities evolve. Using the co-phylogeny framework, we can test the significance of a global hypothesis of co-evolution between genes and language [15].

Co-phylogenetic analysis requires the specification of a binary association link matrix that indicates which of the locally spoken languages each sampled individual speaks (e.g., for Timor: Bunak, Betun, Dawanta, Kemak and/or Upper Tetun;

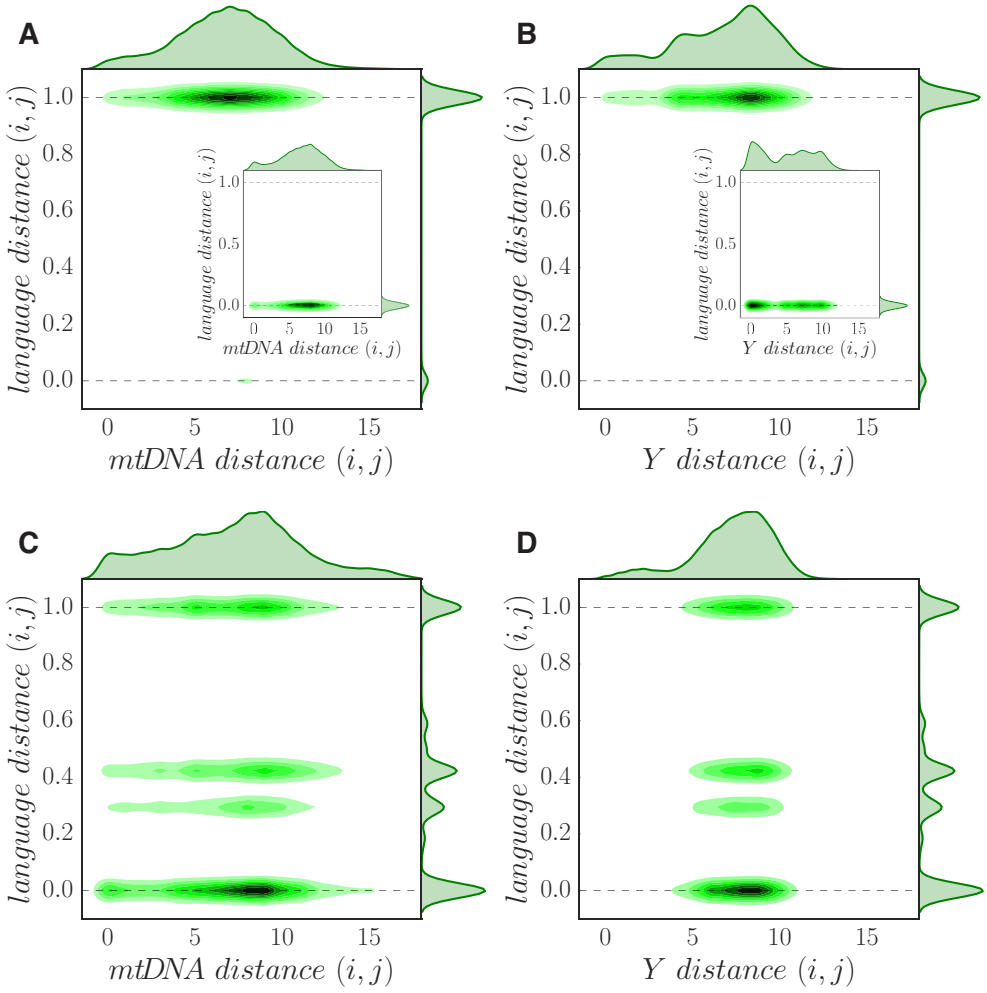

Fig. S9. Linguistic and genetic distances. The kernel density of the linguistic and genetic distances is shown for all pairs of individuals. (A) On Sumba, where villages are monolingual, only a few pairs with mtDNA distance ~8 speak the same language. (B) For the Y chromosome on Sumba, pairs of individuals who speak the same language form two clusters of close (~0) and slightly distant (~5–10) relatedness. (C) On Timor, there is a high probability of finding pairs of individuals that are closely related both on their mtDNA and in terms of the languages they speak. (D) Most pairs of individuals on Timor have a distant degree of relatedness on their Y chromosome, but speak a similar set of languages.

for Sumba: languages spoken in each monolingual village). On Timor, language fluency was assessed for each participant on a three-point ordinal scale ('Understand', 'Understand some', 'Do not understand') during a survey administered to each participant. The first two categories were then coded as a positive indication of ability with the corresponding language for the purposes of the cophylogenetic analysis.

Statistical test for gene-language coevolution. To test whether a significant association exists between the evolution of genes and languages, we employ statistics that are functions of the genetic and language phylogenies, together with their association links (i.e., the list of which individuals speak which languages) [15]. In this statistical test, we examine the significance of congruence, if any, between the gene and language trees. If genes and languages occupy corresponding positions in both phylogenetic trees with a significant degree of congruence, we reject the global null hypothesis that their evolution has been independent.

ParaFit statistic. Evaluating the global hypothesis requires three pieces of information: i) the gene phylogeny **G**, ii) the language phylogeny **L**, and iii) the association between genes and languages **LG**.

The cophylogeny matrices **L** and **G** (Fig. S10) represent the principal coordinates of the linguistic and genetic distances of individuals along their respective trees. The association matrix **LG** is a binary link specifying the languages spoken by each individual. To calculate the congruence of the gene and language trees, we define the fourth-corner statistics matrix **D** [15] as

$$D = G\ LG'\ L \qquad [6]$$

From **D**, we define the global association $ParaFitGlobal$ statistic, to test the gene-language coevolution hypothesis, as the sum of squares of d_{ij}

$$ParaFitGlobal = trace(\boldsymbol{D}'\boldsymbol{D}) = \sum ({d_{ij}}^2) \qquad [7]$$

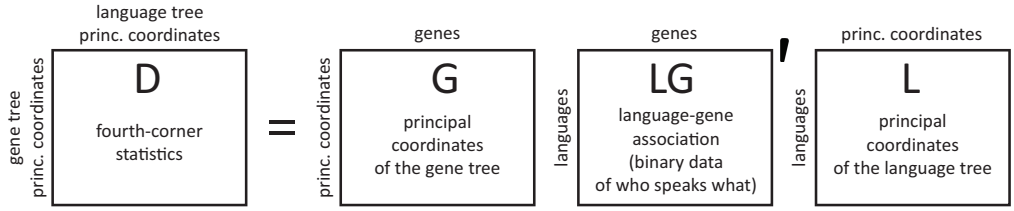

Fig. S10. Fourth-corner matrix. Statistics describing gene-language associations.

Permutation Model. To test whether the association between genes and languages is significant, we calculate $ParaFitGlobal$ on the original data, as well as on the permuted data. Permutation is performed by shuffling the columns of **LG** as shown in Fig. S11. Randomizing the columns is akin to shuffling the gene labels in the link matrix, which removes the association, but preserves the extent and degree of multilinguality.

The observed association between genes and languages is then determined using Z scores. The distribution of $ParaFitGlobal$ values of different permuted realizations of the data, $ParaFitGlobal^{perm}$, are taken and the number of standard deviations $ParaFitGlobal^{data}$ (i.e., the $ParaFitGlobal$ value of the original data) is away from the mean of $\{ParaFitGlobal^{perm}\}$ is calculated. The Z score is taken over 2^5 randomizations r, each with 999 permutations p of $ParaFitGlobal^{perm}$.

$$Z = \frac{ParaFitGlobal^{data} - \frac{\sum_{r,p} ParaFitGlobal^{perm}_{r,p}}{r+p}}{s(\{ParaFitGlobal^{perm}_{r,p}\})} \qquad [8]$$

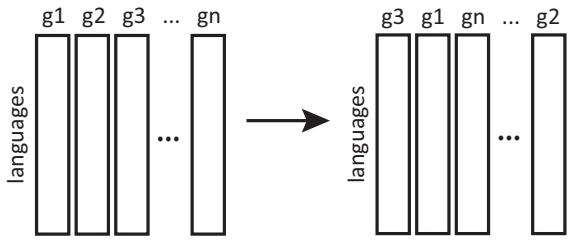

Fig. S11. Permutation model. Columns of the **LG** matrix are shuffled to remove associations, while preserving the number of multilingual individuals and the number of languages they speak.

Language switching on genetic trees. To estimate host switching probabilities, we assume that both gene and language trees are accurately reconstructed, and seek to generate a one-to-one mapping of branching points in the gene and language trees, thus describing how the languages were transmitted and ultimately leading to the current language(s) spoken at each leaf of the gene tree today. A plausible model shown in Fig. S12 predicts the languages spoken at each branch of the gene tree at generation t. This stochastic simulation is run forward in time over the trees using two rules. First, where the language tree branches into two daughter languages, all genetic clades that speak the ancestral language are randomly assigned to one of the daughter languages. Second, in each generation, there is a probability α that a given clade switches to a new language chosen from among the languages that exist in the population at that time. That is, a proportion of lineages α switches to a new language at each generation. These two rules provide a simple model of language diversification and host switching (Fig. S12) that is sufficient to reconstruct patterns of language sharing observed in the present.

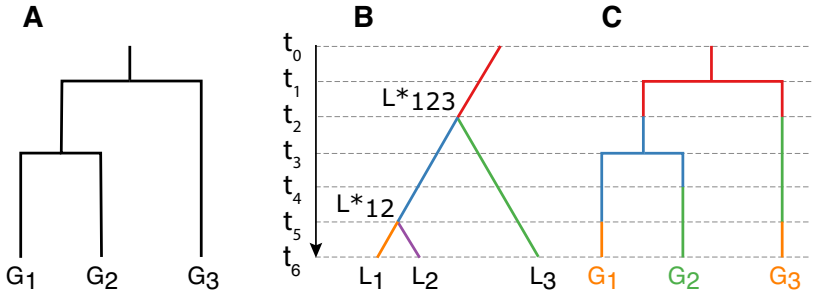

Fig. S12. Language switching in a gene tree conditioned on a language tree. Shown is (A) an unannotated gene tree, (B) a language tree and (C) an example of a simulated language-annotated gene tree. Language diversifies i) during language branching (L^*_{123} splits into L^*_{12} and L_3 at t_2), and ii) during host switching, as in the branches leading to G_2 (language switch at t_4) and G_3 (language switch at t_5).

This plausible stochastic model of language transmission along the branches of the gene tree was run forward in time, starting from the gene tree root (t_0 in Fig. S12C). The following process was performed at each generation of 25 years: i) determine the genetic clades and languages at generation t (e.g., two genetic clades and languages L^*_{12} and L_3 at t_2 in Fig. S12B-C); ii) if a language branches at generation t, each genetic branch that carried the ancestral language at $t-1$ is randomly assigned to one of the two new languages (e.g., at t_2 in Fig. S12C, L^*_{12} is assigned to the first branch and L_3 to the second branch); iii) each genetic branch then switches to a different language within the current pool of languages with probability α (e.g., genetic branch G_2 switches to language L_3 at t_4).

This simulation process was repeated for many gene-language simulations across the range $\alpha = [0, 100]$ where $\alpha = 0$ indicates no language switching and $\alpha = 100$ means all lineages switch their language every generation. The average Z of each α was then compared with the Z of the observed data to find the degree of language switching that best fits the data. Variants of this gene-language co-evolution model yield qualitatively similar results.

Multilinguality in the model of language switching between genetic clades. The model of language switching between genetic clades as shown in Fig. S12 was repeated with simulations across a range of values of the language switching rate α. From these, we obtained a distribution of Z scores and identified the range of α values that yield Z scores similar to those seen in the observed data. This simulation model is sufficient to generate plausible patterns of language sharing in Sumba, where individuals are monolingual. However, multilinguality is common in Timor, and we therefore modified the model such that individuals have a list of languages. When a new language is introduced (i.e., an ancestral language branches into daughter languages), or language switching occurs (as determined by α), there is now a fixed probability β that the new language is appended to the language list rather than replacing the language list. We are primarily interested in fitting the language switch rate α, and so only a limited range of β values were explored. In main text Fig. 5, we take $\beta = \alpha$. Other β values yield qualitatively similar results, with all values leading to Z score convergence at $\alpha < 5\%$.

Alternative model of language switching between genetic clades. A plausible alternative stochastic model of language transmission along the branches of the gene tree run forward in time (starting from the root of the gene tree) involves the following process, which is performed at each generation t (with an interval of 25 years):

1. Set parameters – determine the genetic clades and languages at generation t;
2. Cospeciation – if a language branches at generation t, each genetic branch that carried the ancestral language at $t-1$ is randomly assigned one of the two new daughter languages. Genetic branches associated with a language that does not branch inherit the language(s) spoken by that branch at time $t-1$; and
3. Language switch – each genetic branch then switches to a different language within the current pool of languages with probability α.

In the case of multilinguality (as on Timor), every individual has a list of languages and the set of languages spoken by an individual is determined by these additional rules:

4. Parent language replaced by daughter language during cospeciation – whenever language branching occurs, for every genetic branch that spoke the ancestral language at $t-1$, one of the daughter languages replaces the ancestral language in the current list of languages of the genetic branch; and

5. Additional language during switch – when a genetic branch switches to a language not in its language list at $t-1$, it retains the other languages it already spoke at $t-1$.

For this alternative model, language switching rates α converge to 1–5% per generation for all trees examined, as shown in Fig. S13. Under this alternative model, the 'half life' of a language on a lineage is 325–1,700 years.

The difference between this alternative model and the model used in main text Fig. 5 lies in steps 4 and 5. In the model used in the main text, the new language is appended with a probability β; in this alternative model, i) the chosen new daughter language always replaces the ancestral language in step 4; and ii) the language chosen during switching in step 5 is always appended to the language list of the genetic branch. That is, in the main model, β is allowed to vary, while in this alternative model, β is always 0 during language tree branching, and β is always 1 during language switching.

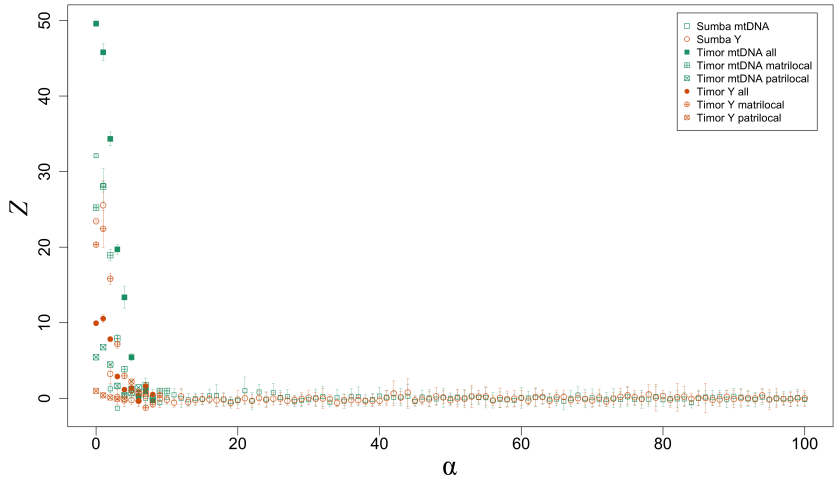

Fig. S13. Language switching rates under the alternative model. The Z score measure of association between gene and language phylogenies on Sumba and Timor is shown for different language switching rates α under the alternative model. All cases independently converge and abruptly lose gene-language associations, behaving similar to randomized cases, when $\alpha \approx 5\%$.

Probability of shared gene-language heritage. The probability that a pair of individuals (i, j) share a common language l given that they belong to the same genetic clade g at generation t is given by

$$P(i_l, j_l | i_{g(t)}, j_{g(t)}) = \frac{P(i_l, j_l \cap i_{g(t)}, j_{g(t)})}{P(i_{g(t)}, j_{g(t)})} = \frac{\sum_{g=1}^{c_t} \left(\sum_{i,j \neq i} 1_{cos(\mathbf{l}_{i_g}, \mathbf{l}_{j_g})>0} \right)}{\sum_{g=1}^{c_t} n_g(n_g-1)/2}, \quad [9]$$

where c_t is the number of genetic clades in generation t each lasting 25 years, $\mathbf{l}_{i_g}$ and $\mathbf{l}_{j_g}$ are the language vectors of individuals i and j belonging to g, $1_{cos(\mathbf{l}_{i_g}, \mathbf{l}_{j_g})>0}$ is 1 if i and j share a language and 0 otherwise, and n_g are the number of individuals in g.

Probability of shared gene-language heritage in idealized scenarios. In Fig. 2 of the main text, we show $P(i_l, j_l | i_{g(t)}, j_{g(t)})$ for the randomly permuted case. Here, we present alternative hypothesis modeling by estimating language distributions using the language switching model (shown in Fig. S12) for two idealized scenarios:

1. The language-switching rate is very high ($\alpha = 1$), such that every lineage has the opportunity to switch its language every generation; and
2. No language-switching occurs ($\alpha = 0$), such that language sharing is purely generated by divergence in the language tree.

Considering mtDNA, the $\alpha = 0$ scenario is consistent with rigid matrilocality, while the $\alpha = 1$ scenario is consistent with (very) frequent female movement, as might accompany patrilocality. The converse is true for the Y chromosome, with $\alpha = 0$ corresponding to rigid patrilocality. Running these scenarios for our genetic and language trees (both mtDNA and the Y chromosome) allows us to make language sharing predictions in the case of extremely sex-biased migrations.

For patrilocal Sumba, the observed mtDNA pattern is relatively close to the patrilocal model prediction throughout the time period explored, and clearly departs from the matrilocal model. The observed Y chromosome pattern is very distant from the matrilocal prediction and generally approaches the idealized patrilocal model, until about 20 kya. These results support

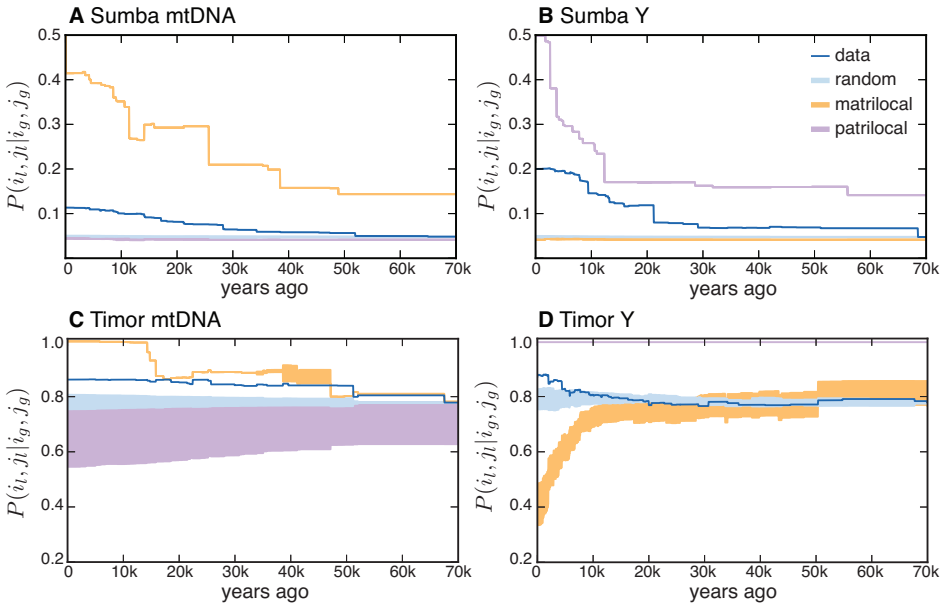

Fig. S14. Language sharing in the mtDNA and Y phylogenies of (A-B) Sumba and (C-D) Timor. The probability is shown of sharing a language l given that each pair of individuals are in the same genetic clade g at a given time in the past. Solid blue lines represent the observed metric, with shaded blue bands indicating the result of random permutations of the linguistic data. Idealized scenarios are shown for strict matrilocality (yellow) and patrilocality (violet). Probabilities in the Sumba dataset approach the strict patrilocal case for mtDNA (all times) and the Y chromosome (until ∼20 kya). In Timor, matrilocal probabilities are higher than chance for mtDNA, and mostly never greater than chance for the Y chromosome.

our primary conclusions; not only is $P(i_l, j_l|i_g, j_g)$ greater overall for the Y chromosome in the patrilocal Sumba data, but this even begins to approach the predictions of an idealized, very strict patrilocal model.

The situation is more complicated for Timor because most individuals are multilingual. The higher prevalence of multilinguality in central Timor that we observed probably reflects recent history (notably forced migrations over the last couple of generations). In central Timor, there have been population movements caused by warfare over the last century, potentially contributing to the mixture of languages.

To capture the idealized matrilocal and patrilocal patterns, we now need to additionally consider the implications of such rigid mating systems on multilinguality. To illustrate one method of calculating $P(i_l, j_l|i_g, j_g)$ while taking multilinguality into account, we ran the same model used for Sumba, but with an additional probability parameter β that the new language is appended to the language list rather than replacing it (see the 'Multilinguality in the model of language switching between genetic clades' section below). For both the strict matrilocality and patrilocality cases, we used $\beta = 0.2$ when language switching occurs. When a new language is introduced (i.e., language tree branching), we used $\beta = 0.7$ for mtDNA and $\beta = 0.96$ for the Y chromosome. These β values were chosen as they represent the transition points below which the resulting language distribution and shared probabilities closely fit the Timor data.

Note that we are in no way claiming that this method is necessarily the right way to handle multilinguality, and instead present this analysis as an exploratory study only. Multilinguality is an extremely complex question and largely beyond the scope of this manuscript. Nevertheless, these analyses clarify how the $P(i_l, j_l|i_g, j_g)$ findings for Timor relate to alternative idealized scenarios. Here, we find that for mtDNA, the pattern on Timor is greater than random chance, analogous to rigid matrilocality. For the Y chromosome, the pattern on Timor is mostly never greater than random chance, similar to strict matrilocality, which is also always below random chance.

1. Van Oven M, Kayser M (2009) Updated comprehensive phylogenetic tree of global human mitochondrial dna variation. *Human Mutation* 30(2).
2. ISOGG (2017) International society of genetic genealogy (http://www.isogg.org/tree, accessed 10-24-2016).
3. Karafet TM et al. (2008) New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Research* pp. 830–838.
4. Bruvo R, Michiel NK, D'Souza TG, Schulenburg H (2004) A simple method for the calculation of microsatellite genotype distances irrespective of ploidy level. *Molecular Ecology* 13(7):2101–2106.
5. Fu Q et al. (2013) A revised timescale for human evolution based on ancient mitochondrial genomes. *Current Biology* 23(7):553–559.
6. Paradis E, Claude J, Strimmer K (2004) Ape: Analyses of phylogenetics and evolution in r language. *Bioinformatics* 20(2):289.
7. Lewis PM, Simons GF, Fennig CD (2013) Ethnologue: Languages of the world.
8. Paradis E (2013) Molecular dating of phylogenies by likelihood methods: a comparison of models and a new information criterion. *Molecular Phylogenetics and Evolution* 67(2):436–444.
9. Xu S et al. (2012) Genetic dating indicates that the asian–papuan admixture through eastern indonesia corresponds to the austronesian expansion. *Proceedings of the National Academy of Sciences* 109(12):4574–4579.
10. Wilkinson-Herbots HM (2008) The distribution of the coalescence time and the number of pairwise nucleotide differences in the "Isolation with Migration" model. *Theoretical Population Biology* 73:277–288.
11. Slatkin M, Hudson RR (1991) Pairwise comparisons of mitochondrial dna sequences in stable and exponentially growing populations. *Genetics* 129(2):555–562.
12. Ewing G, Hermisson J (2010) MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics* 26(16):2064–2065.
13. Therik T (2004) *Wehali – The Female Land: Traditions of a Timorese Ritual Centre.* (Pandanus Books: ANU Research School of Pacific and Asian Studies, Canberra, Australia).
14. Legendre P, Desdevises Y, Bazin E, Page RDM (2002) A statistical test for host–parasite coevolution. *Systematic Biology* 51:217–234.
15. Dray S, Legendre P (2008) Testing the species traits–environment relationships: The fourth-corner problem revisited. *Ecology* 89:3400–3412.
16. Tehrani JJ, Collard M, Shennan SJ (2010) The cophylogeny of populations and cultures: reconstructing the evolution of iranian tribal craft traditions using trees and jungles. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365(1559):3865–3874.