Small Traditional Human Communities Sustain Genomic Diversity over Microgeographic Scales despite Linguistic Isolation

Murray P. Cox,*[,†,1] Georgi Hudjashov,[†,1] Andre Sim,[1] Olga Savina,[2] Tatiana M. Karafet,[2] Herawati Sudoyo,[3] and J. Stephen Lansing[4]

[1]Statistics and Bioinformatics Group, Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand

[2]ARL Division of Biotechnology, University of Arizona

[3]Eijkman Institute for Molecular Biology, Jakarta, Indonesia

[4]Complexity Institute, Nanyang Technological University, Singapore

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: m.p.cox@massey.ac.nz.

Associate editor: Connie Mulligan

Abstract

At least since the Neolithic, humans have largely lived in networks of small, traditional communities. Often socially isolated, these groups evolved distinct languages and cultures over microgeographic scales of just tens of kilometers. Population genetic theory tells us that genetic drift should act quickly in such isolated groups, thus raising the question: do networks of small human communities maintain levels of genetic diversity over microgeographic scales? This question can no longer be asked in most parts of the world, which have been heavily impacted by historical events that make traditional society structures the exception. However, such studies remain possible in parts of Island Southeast Asia and Oceania, where traditional ways of life are still practiced. We captured genome-wide genetic data, together with linguistic records, for a case–study system—eight villages distributed across Sumba, a small, remote island in eastern Indonesia. More than 4,000 years after these communities were established during the Neolithic period, most speak different languages and can be distinguished genetically. Yet their nuclear diversity is not reduced, instead being comparable to other, even much larger, regional groups. Modeling reveals a separation of time scales: while languages and culture can evolve quickly, creating social barriers, sporadic migration averaged over many generations is sufficient to keep villages linked genetically. This loosely-connected network structure, once the global norm and still extant on Sumba today, provides a living proxy to explore fine-scale genome dynamics in the sort of small traditional communities within which the most recent episodes of human evolution occurred.

Key words: **genetic diversity, linguistic diversity, gene flow, population structure.**

Article

Introduction

At least since the Neolithic, our ancestors have mostly lived in small groups—networks of settled communities that were often widely dispersed across the landscape, in striking contrast to the large, urbanized and high inter-connected societies that have emerged within the last few thousand years (Bellwood 2013). It was within these kinds of very small groups that many recent episodes of human evolution occurred, including well-known instances of migration, admixture, and selection (1000 Genomes Project Consortium 2015; Gurdasani et al. 2015; Karmin et al. 2015; Mathieson et al. 2015; Pääbo 2015; Sudmant et al. 2015). Such groups tend to be culturally diverse, with different languages commonly spoken over scales as small as tens of kilometers. Low levels of intermarriage between different linguistic communities impose what superficially appear to be high levels of inter-group isolation (Friedlaender 1975). Population genetic theory tells us that genetic drift should reduce diversity quickly within such small groups (Ewens 2004). But is this what actually happens? Or is occasional migration between neighboring communities sufficient to keep them connected, thus maintaining genome diversity even where there are extensive linguistic boundaries?

This question can no longer be asked for humans in most parts of the world. Small traditional societies that resemble prehistoric settings have been replaced in most regions, including large parts of Europe, Africa, Asia, and the Americas, by major post-Neolithic movements. Further population restructuring has frequently been driven by the actions of modern states during the historic era (for instance, see Leslie et al. 2015 for a history of these processes in the British Isles). Many of these heavily restructured populations have since become dominant players in the modern world and hence are frequently the subject of genetic studies (McVean et al. 2005; 1000 Genomes Project Consortium

2012). However, to understand genome dynamics in the kinds of societies where humans once evolved, we instead need to intensely sample regions where small traditional community networks still exist, notably in more remote regions of Island Southeast Asia and Oceania.

Sumba, a small island in eastern Indonesia, is an ideal case–study system. The first modern humans to settle this region reached Sumba ~50,000 years ago, but the island's cultural, linguistic, and genetic landscapes were heavily restructured ~4,000 years ago with the arrival of Asian farming cultures. Despite its small size (comparable to Jamaica or Cyprus), the Neolithic population of Sumba—a complex mixture of incoming Asian and preexisting Melanesian ancestries—has evolved since the mid-Holocene into an interlocking network of small modern villages that speak nine different languages (Lansing et al. 2007). In contrast to neighboring islands such as Timor (Tumonggor et al. 2014), most individuals on Sumba are monolingual. The island's current ~685,000 inhabitants (Badan Pusat Statistik 2010) live in villages of only a few hundred to a few thousand people (Lansing et al. 2008), with the largest city hosting fewer than 12,000. Many of these communities speak different languages than their closest neighbors, some just 10-km away. The island has proved amenable to microevolutionary questions, with previous studies highlighting the co-divergence of languages and Y chromosome diversity over small geographical scales (Lansing et al. 2007), a limited historical role for dominant males (Lansing et al. 2008) and the importance of complex marriage rules in structuring social connections (Guillot et al. 2015). The island therefore provides a natural 'living laboratory,' where communities from a common source, but with different historical trajectories and languages, can be used to explore the relative effects of isolation and contact for humans living in small traditional groups.

Here, we report genome-wide genetic diversity in 235 individuals drawn from a network of eight villages distributed across Sumba. By combining deep datasets of genetic and linguistic data, we can discern patterns of isolation and contact on the island with unprecedented detail. The most striking finding is that although most of these small communities speak different languages and are sufficiently isolated so as to be genetically distinguished, they still maintain levels of genetic diversity comparable to even much larger regional groups. Small human populations therefore seem able to sustain long periods of apparent isolation with both external regions and neighboring villages, enough to evolve extensive linguistic diversity, without necessarily being adversely affected by genetic drift. This pattern—once globally common, but now erased in most regions by extensive post-Neolithic movements (Bellwood 2013)—has implications for how cultural versus genetic evolution has occurred in recent human history. Small communities, like those observed on Sumba today, therefore provide one of the few remaining modern proxies for understanding the fine-scale evolutionary processes associated with traditional society structures, which no longer exist in most larger and better-studied regions of the world.

Results

Data

Genome-scale single nucleotide polymorphisms (SNPs) were screened using an Illumina OmniExpress SNP array in 235 individuals from eight communities spanning the eastern Indonesian island of Sumba: Anakalang, Kodi, Lamboya, Loli, Mamboro, Rindi, Wanokaka, and Wunga (fig. 1A). Following genotyping quality checks and removal of cryptic close relatives, the final dataset contained 204 individuals, with 22–28 individuals per population (supplementary table S1, Supplementary Material online).

Descriptive Measures of Population Structure

A principal components analysis (PCA) was performed on the eight Sumba groups (fig. 1B and C). Although the first two principal components explain only a small proportion of the total variance (2.38%), the communities broadly form distinguishable clusters, consistent with an ADMIXTURE analysis (supplementary fig. S1, Supplementary Material online). These clusters, with some notable exceptions, have limited overlap. Some of the biggest genetic differences are observed between populations that are geographically close, such as Anakalang and Wunga, even though these communities speak related languages. Indeed, the most geographically distant population (Rindi) falls near the center of the PCA plot and Wunga displays the biggest within-population variance. Lamboya, Loli, and Wanokaka—a trio of nearby villages—emphasize these variable effects of geography. Although they are the closest villages studied here (only 10–12-km apart), Lamboya and Loli are genetically very similar. However, they cluster separately from their equally close neighbor Wanokaka, and they instead show a close genetic connection with Kodi, which lies ~50 km to the west (fig. 1a and $K = 3$ sub-plot in supplementary fig. S1, Supplementary Material online). With expected exceptions, these clusters are statistically robust (supplementary table S2, Supplementary Material online).

Pairwise F_{ST} values (supplementary table S3, Supplementary Material online) between populations show no significant association with geography within Sumba (Mantel $r = -0.14$, $P = 0.66$), although a significant association is found for long-range Identity-by-Descent (IBD) regions (Mantel $r = -0.85$, $P < 1.0 \times 10^{-6}$) (supplementary fig. S2, Supplementary Material online). Particularly small F_{ST} distances between Rindi and many other Sumba communities perhaps reflect the historical role of Rindi as a political center, which may have attracted migrants from elsewhere on the island (Guillot et al. 2013).

Language does not seem to be a strong driver of population structure. Lansing et al. (2007) identified five overarching language groups on Sumba, of which four are sampled here: group A, Loli; group B, Kodi and Lamboya; group C, Anakalang, Mamboro, Wanokaka, and Wunga; and group E, Rindi (fig. 1). However, population structure is not obviously associated with language. Even communities in the same language group (such as Anakalang, Mamboro, Wanokaka, and Wunga) have distinctive ancestry profiles (supplementary fig.

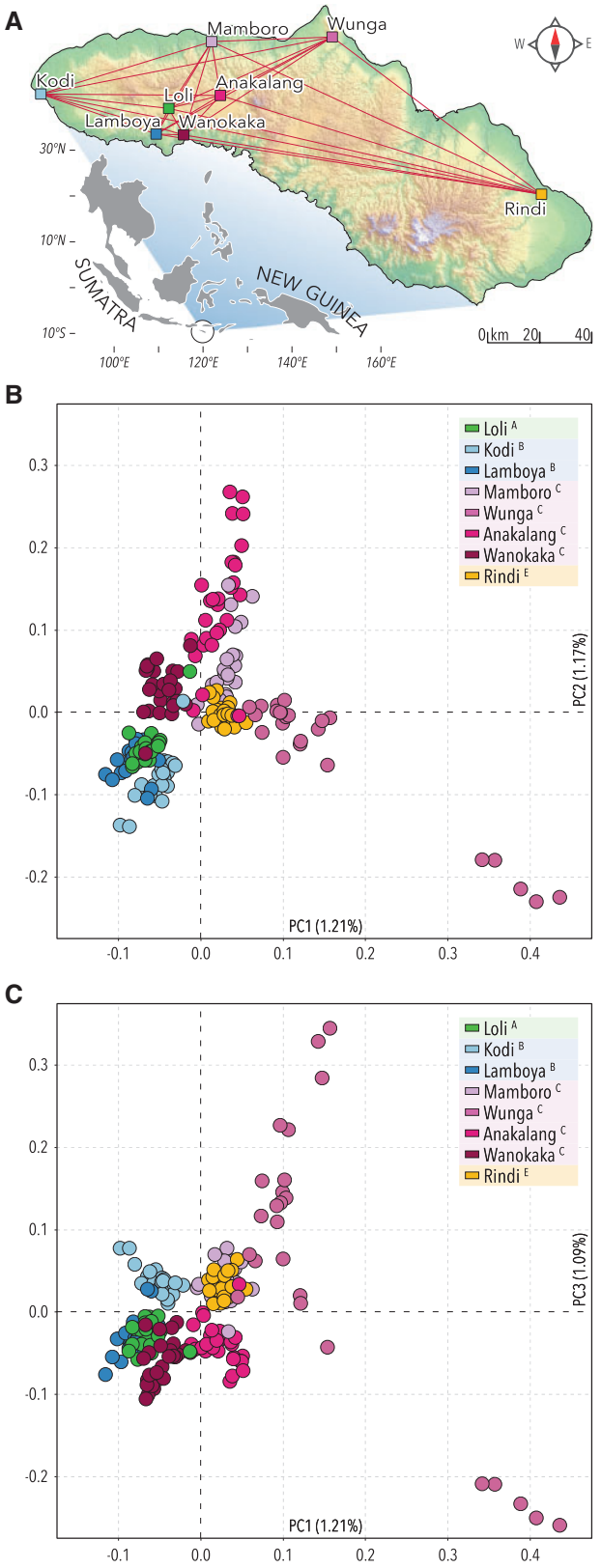

FIG. 1. (*A*) Locations of the eight communities sampled on Sumba in eastern Indonesia. Lines indicate the 28 pairwise community comparisons used throughout the study. (*B*, *C*) PCA of genome-wide SNP diversity in 204 individuals from the eight communities. Axes are scaled by the proportion of variance described by the corresponding principal component (PC): PC1 versus PC2 in (*B*), and PC1 versus PC3 in (*C*). As noted in the legend, colors indicate related language communities, as per Lansing et al. (2007): group A, Loli (green); group B, Kodi and Lamboya (blue shades); group C, Anakalang, Mamboro, Wanokaka and Wunga (red shades); and group E, Rindi (yellow). Note genetic discrimination at the level of individuals and villages, as well as recent migrants between communities, such as the Wanokaka individual (maroon) with recent ancestry from Lamboya or Loli (dark blue and green, respectively) in (*B*).

S1, Supplementary Material online, $K = 3$ and higher) and form their own separate clusters in the PCA plot; an Analysis of Molecular Variance (AMOVA) shows no significant genetic variance associated with the language groupings (supplementary table S4, Supplementary Material online); and F_{ST} values between populations in the same language group are not significantly smaller than those between populations in different language groups (Monte Carlo permutation, $P = 0.63$), although there is significantly more sharing of long-range haplotypes (an average total length per individual of 67 vs. 55 cM; Monte Carlo permutation, $P = 0.020$). All four analyses suggest that language played only a weak role in determining autosomal genetic structure.

Some recent individual movements appear on the PCA plot (fig. 1B and C). At least three individuals cluster with their ancestral community, but carry the color code of their new home: recent migrants from Anakalang or Wanokaka to Loli (green) and Kodi (light blue), and Lamboya or Loli to Wanokaka (maroon). Perhaps unsurprisingly, these migrants often appear to have moved between nearby villages. The causes of these movements—perhaps community-driven to promote trade linkages or for prosaic personal reasons—are not known.

The population structure of the island was placed within a broader regional context by exploring the ancestral Asian and Melanesian genomic components present in modern Sumba communities (fig. 2). Previous studies, specially designed to address the question of admixture proportions, estimated that individuals on Sumba carry an average of 74% Asian alleles on the autosomes, with one of the strongest regional biases towards Asian ancestry on the X chromosome (86%) (Cox et al. 2010). The ancestry profile for the eight Sumba populations as determined by ADMIXTURE seems broadly consistent, and also matches expectations from a range of other regional groups (Li et al. 2008; 1000 Genomes Project Consortium 2012). Average cross-validation errors for multiple ADMIXTURE runs representing the most frequent modal solution were minimized at $K = 4$ (fig. 2). For Sumba populations, the proportion of two main ancestries found on the island, Asian and Melanesian, remain similar for all values of K ($K = 2$–7) (supplementary fig. S3, Supplementary Material online), and more importantly, vary little between villages.

Statistical Inference of Population Structure

The extent of past migration between Sumba communities was determined by sequential coalescent modeling and statistical inference against observed population pairwise F_{ST} values (supplementary table S3, Supplementary Material online). Sumba populations radiated rapidly to establish

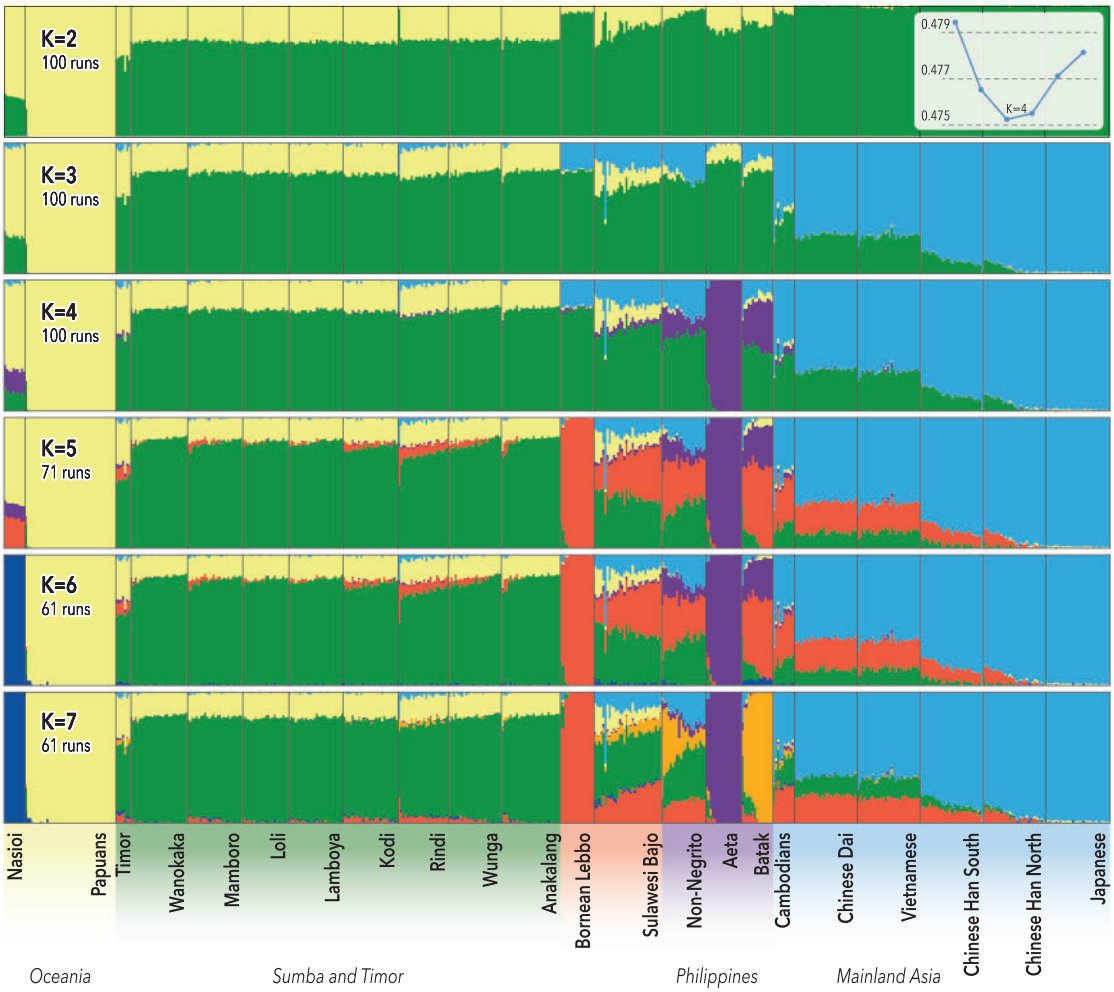

Fig. 2. Ancestral genomic components in Sumba relative to other regional populations. For every *K*, the modal solution with the highest number of ADMIXTURE runs is shown; individual ancestry proportions were averaged across all runs from the same mode and the number of runs (out of 100) assigned to the presented solution is shown. Note that ancestry components are largely shared across all eight Sumba communities, but differ from regional neighbors. Average cross validation statistics were calculated across all runs from the same mode and are minimized for four ancestry components in this dataset (inset).

communities across the island, and although contact between villages on Sumba appears to have been limited, there was apparently even less interaction with neighboring islands (Lansing et al. 2007). This premise is confirmed by analysis of long-range IBD regions, which occur most commonly within Sumba communities, then between Sumba communities, and finally show only limited sharing with neighboring populations from Island Southeast Asia and Melanesia (fig. 3). This is consistent with expectations observed at a much higher geographical scale: most genetic diversity is found globally within populations with little sharing between them (Gravel et al. 2011).

Sanderson et al. (2015) developed a generalized model of Asian–Melanesian admixture based on demographic information drawn from a detailed body of literature on human populations across Island Southeast Asia (Cox et al. 2008, 2010; Xu et al. 2012). In brief, Asian and Melanesian ancestral groups merged during the Neolithic expansion to create an admixed population, with relative proportions as inferred previously for Sumba (Cox et al. 2010). Evidence from genetics, linguistics, and oral history all support a rapid radiation of this ancestral group to establish the eight communities studied here (albeit not always in their current locations). This history was formalized into a demographic model for Sumba (fig. 4A), which was run with a wide range of migration values *m* and compared to the observed F_{ST} distances using Bayesian inference. Cross-validation returns a very low prediction error ($E_{pred} = 0.044$), suggesting that the migration rate can be inferred with high statistical accuracy (supplementary fig. S4, Supplementary Material online). The migration rate *m* on Sumba was estimated at 1.5% of the population per generation (95% confidence interval 1.3–1.9%) (fig. 4B). Alternative methods, including rejection and neural networks, and alternative tolerance values (0.001–0.05), produced quantitatively similar results. This migration estimate implies that communities exchanged on average only a small number of individuals per generation (although likely not distributed uniformly through time). Importantly, however, they still developed observable population structure with clear linguistic diversification.

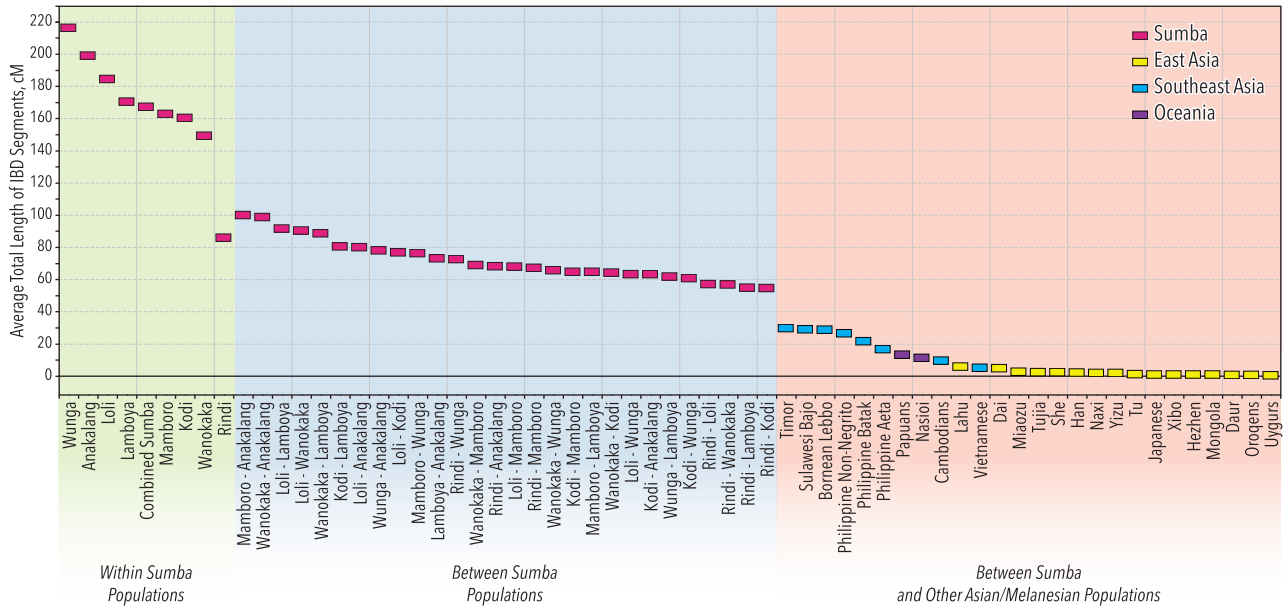

Fig. 3. Extent of long-range IBD regions (>1 cM) between individuals within each Sumba community (green background), between communities on Sumba (blue background), and between Sumba and other regional populations (red background). Data points for individual populations are color-coded according to the legend.

Levels of Genetic Diversity

Given the relative isolation of populations on Sumba since they were founded around 4,000 years ago, we explored the downstream effects on levels of genetic diversity. Two different approaches were adopted, genotype-based mean pairwise nucleotide diversity per site (π) and haplotype-based gene diversity (H). Both statistics were calculated for the eight Sumba communities, together with a global range of reference populations. Contrary to expectations, diversity within individual Sumba populations ($\pi = 0.281–0.289$ and $H = 0.690–0.699$) is not significantly different from global values (0.288 and 0.707, respectively, estimated excluding Sumba) (fig. 5 and supplementary table S5, Supplementary Material online). Indeed, Sumba populations lie close to the mean of worldwide distributions for both indices. Furthermore, pairwise nucleotide diversity was also estimated for whole genome sequences from five East and Southeast Asian populations in the 1000 Genomes Project dataset and compared with values obtained using the set of 360,452 SNPs from the main study dataset. As expected, whole genome sequences show lower pairwise nucleotide diversity than ascertained SNP array data, which preferentially exclude rare variants. However, the extent of this bias varies little between groups (supplementary fig. S5, Supplementary Material online), regardless of whether they formed part of the ascertainment panel or not (i.e., a population from which markers were chosen to design the genotyping array). This suggests that genetic diversity indexes calculated from SNP array data are a fair reflection of relative diversity levels between populations within this broader region.

Finally, the coefficient of inbreeding (F_{IS}) and Runs of Homozygosity (ROH)—long stretches of homozygous alleles associated with higher levels of inbreeding—were calculated on the same global dataset with a minimum length of 50 homozygous SNPs. Notably, the coefficient of inbreeding (supplementary fig. S6, Supplementary Material online), the number of ROH and the total amount of the genome located in homozygous stretches (fig. 6) also do not differ markedly between Sumba and other large regional groups. (Qualitatively similar results were obtained for alternative ROH analyses with minimum run lengths ranging from 20 to 95 SNPs and different linkage disequilibrium pruning parameters.) Nor was any community-level structure in homozygous runs observed among the populations sampled on Sumba (fig. 6B). The villages therefore have similar levels of genomic diversity, comparable to even much larger regional and global populations, and show no clear genetic signs of inbreeding.

Discussion

Genome-wide autosomal diversity was screened in 235 individuals from eight communities distributed across the eastern Indonesian island of Sumba. With an area of only 11,000 km^2, comparable in size with Jamaica or Cyprus, Sumba is relatively small in global terms and is dwarfed by some of its neighboring islands in the Indonesian archipelago. Yet despite its small size, the island hosts nine languages in the Central-Eastern Malayo-Polynesian group of the widespread Austronesian language family. While studies of genetic–linguistic coevolution are common at much larger geographical scales (Hunley 2015), we instead collected a broad suite of data for eight communities located across Sumba with the aim of undertaking a microgeographic survey of genetics and language in one of the most diverse regions on Earth.

The most distant community pairs (Kodi and Rindi) lie only 192-km apart, while the closest villages (Lamboya and Wanokaka) are only 10-km apart, with a mean distance of just

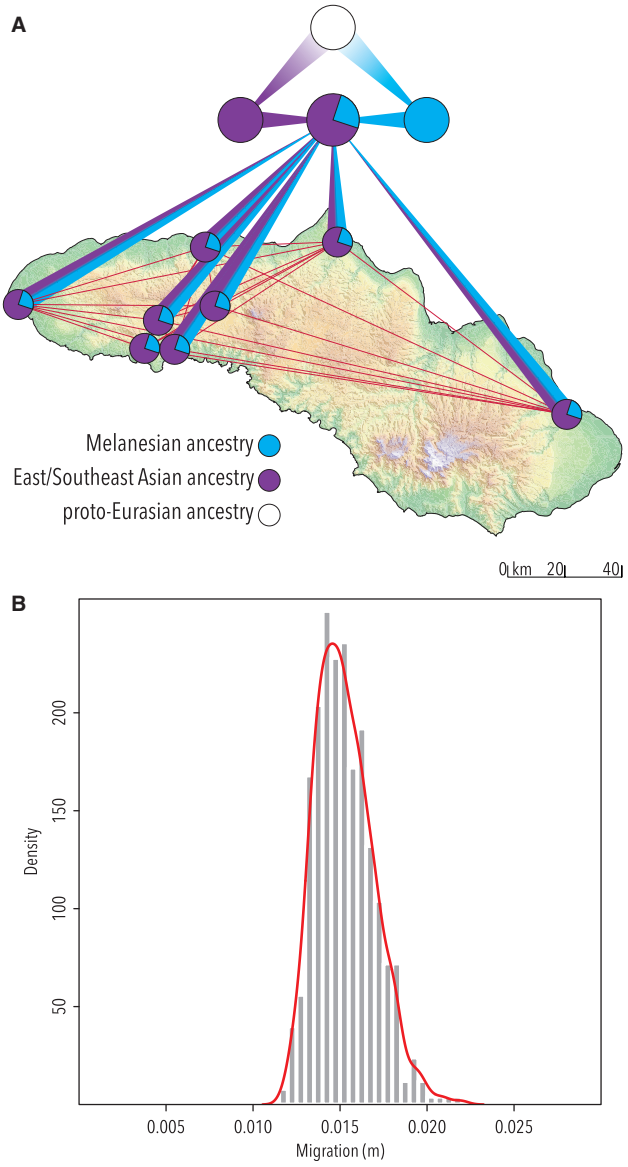

FIG. 4. (*A*) Genome-scale structured coalescent model for the demographic history of Sumba. An ancestral Eurasian population (white) diverges into Asian (purple) and Melanesian (blue) ancestral groups, which subsequently merge to form an admixed population on Sumba during the Neolithic farming expansion. This founding population radiates to establish the eight sampled communities, which continue to exchange migrants to the present (red lines). (*B*) Posterior distribution of the mean migration rate *m* inferred across all 28 community pairs on Sumba with a final uniform prior from 0 to 0.03. The red line shows a density curve overlaid on the histogram using a Gaussian smoothing kernel.

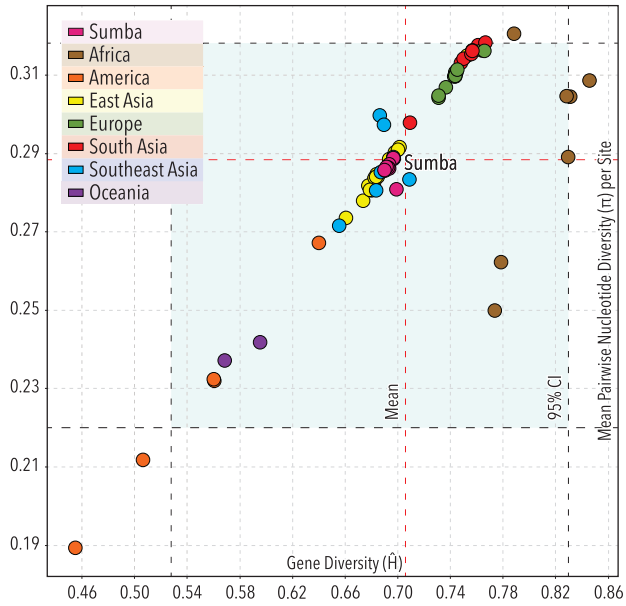

FIG. 5. Average pairwise nucleotide diversity (π) and gene diversity ($\hat{H}$) for Sumba (magenta) and a global range of reference populations. The mean values (dashed red lines) and 95% confidence intervals (shaded area) are calculated from all populations except Sumba.

68 km. The three closest communities (Lamboya, Loli, and Wanokaka) each speak different languages. However, a traditional comparative linguistic analysis of lexical and phonetic data collected from these villages, as made by an experienced historical linguist, shows that all Sumba languages derive from a single ancestral language, proto-Sumbanese, and are therefore more closely related to each other than to any language outside the island (Lansing et al. 2007) (also see figs. 2, 3 and supplementary fig. S3, Supplementary Material online, for corroborating genetic evidence). Further, the language data are treelike, but deep branching, suggestive of rapid early radiations. This is consistent with limited existing genetic evidence, almost entirely from haploid loci, which suggests that the two most distant villages diverged no more than 4,875 years ago (Lansing et al. 2007), and that incoming farming groups, with ultimate Asian ancestry (Soares et al. 2016), began mixing with preexisting local Melanesian populations on Sumba ~4,085 years ago (95% confidence interval 3,716–4,484; Xu et al. 2012). This view is consistent with oral history, which holds that the modern communities of Sumba all derive from a single ancestral village on the island's northern coast, close to the modern village of Wunga (Hoskins 1993).

Multiple lines of evidence therefore point to a relatively simple model that appears to capture much of the history of Sumba. Around 4,000 years ago, incoming farming groups, ultimately with genetic ties back to the Asian mainland, reached Sumba and intermarried with local hunter-gatherer Melanesian populations that had lived on the island since its initial settlement nearly 50,000 years ago. This admixed group, which spoke proto-Sumbanese, then split and radiated to establish the first settled farming communities across the island. Over time, the languages and genes of these villages evolved and drifted, ultimately leading to the extraordinary human diversity observed on this small island today. A key feature of the island's history is its relative isolation, in striking contrast to the large number of post-Neolithic influences that have impacted places like Europe, or as a geographically closer example, western Indonesia, with its extensive historic-era contact with India and the Middle East (Kusuma et al. 2016). While such a simple demographic model undoubtedly hides a great deal of additional complexity (for instance, see

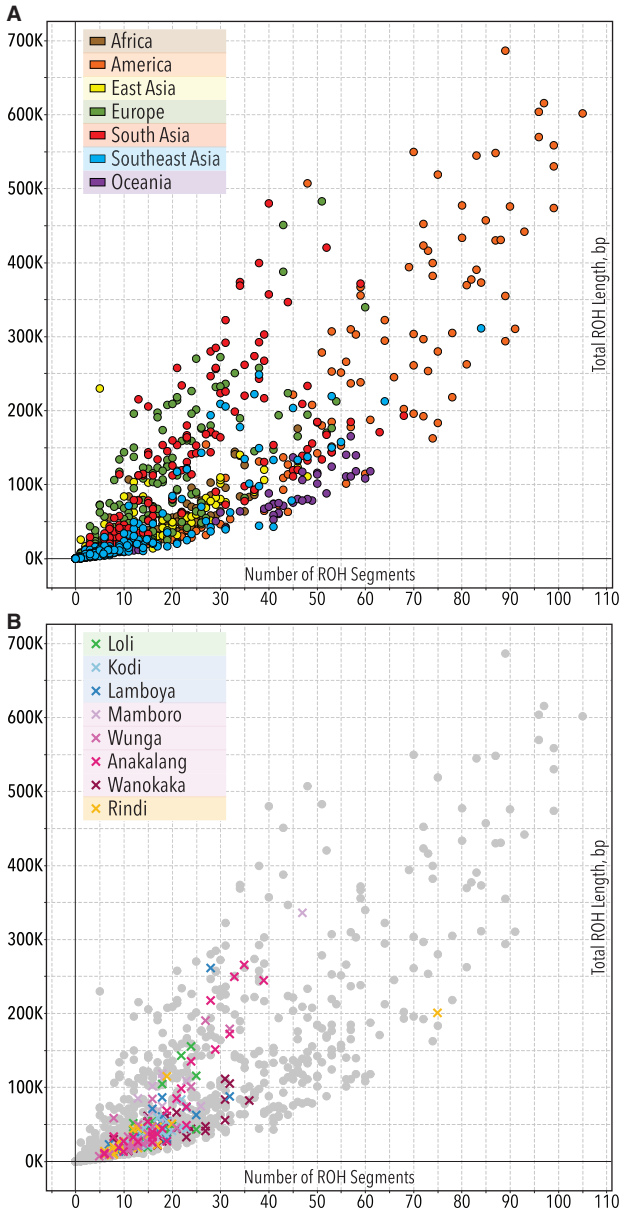

Fig. 6. (*A*) ROH within individuals for a global range of reference populations. (*B*) ROH within individuals from Sumba. The number of homozygous runs is shown on the *x*-axis; the total length of the genome contained in homozygous runs is shown on the *y*-axis. Note that both measures observed for individuals on Sumba overlap with regional and global populations (grey circles) and show no evidence of community-level structuring within Sumba.

Guillot et al. 2015 for the role of complex marriage rules), it nevertheless provides a rare opportunity to explore the genomic effects of long-term isolation at the community level over the extremely small microgeographic scales at which human evolution has typically acted.

The first key finding is that villages on Sumba have been sufficiently isolated that many of them can now be genetically distinguished (fig. 1B and C and supplementary fig. S1). While some villages are clearly functioning as a joint population unit (such as Lamboya and Loli), others form their own genetic clusters. This is surprisingly true even of the two closest villages, Lamboya and Wanokaka, which lie a mere 10-km apart. Conversely, the most distant village, Rindi, falls right in the center of the plot. However, there is also ample evidence of recent inter-community movements (fig. 1B and C). As with Y chromosome diversity (Lansing et al. 2007), autosomal F_{ST} distances are not significantly correlated with geographical distances. A cline of one specific Austronesian-associated haplogroup O lineage has been attributed to the original spread of Asian farming populations across the island (Lansing et al. 2007), but no similar association is observed on the autosomes, perhaps because their larger effective size means that such processes are reflected in autosomal patterns of variation more slowly. Mitochondrial DNA is surprisingly more similar between distant populations (Mantel $r = -0.48$, $P = 0.048$) (Tumonggor et al. 2013), suggesting that those women who did move perhaps traveled long distances, not necessarily simply relocating to the nearest village. However, no clear association is found between autosomal markers and language groupings: Lamboya (group B) is genetically more similar to Loli (group A) than to Kodi (group B); and Rindi (group E) falls genetically in the middle of the group C speakers (Anakalang, Mamboro, Wanokaka, and Wunga) (fig. 1B and C and supplementary table S4, Supplementary Material online). Other historical drivers of population structure must therefore have been more important than shared geography or language.

The second key finding is that past mobility between villages occurred, but was limited. Genome-wide coalescent modeling infers that only a small proportion of the population has typically moved between any two communities per generation (a through-time average of 1.5% of the population per generation) (fig. 4B). This is similar to an earlier estimate of $<3\%$ mobility per generation between villages that speak different languages on the island of Karkar, Papua New Guinea, as estimated from blood group diversity in the late 1970s (Boyce et al. 1978). Few other genetic studies at this geographical scale exist, although an ethnographic study for one population on Sumba does suggest that marriage tends to occur within linguistic communities (Forth 1981). The strong ascertainment bias of SNP arrays makes them poor sources of data for effective population size calculations, and consequently, community sizes on Sumba are not known with any accuracy. However, estimates from haploid loci tend to be relatively high; maternally inherited mitochondrial DNA suggests a mean effective size of 6,575 for women (range 1,700–29,000; Guillot et al. 2013), while paternally inherited Y chromosome data indicate a smaller mean effective size of only 250 for men (range 116–669; Lansing et al. 2008). While the magnitude of this difference can in part be attributed to sex-specific cultural processes such as sex-biased migration (Tumonggor et al. 2013), estimates of the number of separate households in each Sumba community (mean 248; range 93–450) favor the middle of this spectrum (mean census size 1,240; range 465–2,250, conservatively assuming a nuclear family of two parents and three dependents) (Lansing et al. 2008). Archeological estimates of past population sizes are not available for this region, but those inferred for more well-studied regions, such as Neolithic Europe, seem broadly

comparable (Shennan et al. 2013). Looking across these population size estimates for men and women, perhaps 4–100 individuals (with more weight on the lower end of this scale) moved between any two communities on average in each generation in the past. Some recent cases can be observed directly on the PCA plots (fig. 1B and C). Although most people lived their lives within a single community, mobility therefore did occur at low rates between groups. Population genetic theory shows that exchanging just one individual per generation is sufficient to prevent neutral alleles from fixing by genetic drift (i.e., reaching frequencies of 0% or 100%; reviewed in Slatkin 1987) and the observed migration rates inferred here appear sufficient to have maintained genetic connections between communities even as they linguistically diverged. Importantly, though, the villages do still show some signs of genetic differentiation. While ongoing gene flow limits large frequency differences at individual loci, community-level differences like those observed in the PCA plot (fig. 1B and C) can still readily emerge from small allelic frequency differences across a very large number of genomic markers, emphasizing that the villages of Sumba are not simply acting as a single homogeneous group.

The third key finding is that any apparent cultural and linguistic isolation between communities has had little effect on genome-wide levels of genetic diversity within communities. Despite sufficient past isolation that neighboring villages now often speak different languages and can frequently be genetically distinguished, measures of genetic diversity like π and H approach global norms. Indeed, even effective population size estimates are broadly similar to global values (11,600–13,000 for non-African populations; 1000 Genomes Project Consortium 2010). Diversity values for Sumba are only marginally lower than global averages and fall well within the range observed for many other populations, even much larger ones, across Mainland and Island Southeast Asia (fig. 5 and supplementary table S5, Supplementary Material online). This holds true for both genotype- and haplotype-based measures (e.g., pairwise nucleotide diversity π vs. gene diversity H). While genotype-based measures can sometimes be sensitive to the ascertainment bias of SNP arrays (although the effect here appears to be minor; supplementary fig. S5, Supplementary Material online), haplotype-based measures are calculated on large numbers of congruent markers and are consequently more robust. Importantly, for all of the summary statistics calculated here, relative estimates of genetic diversity are higher on Sumba than in the two closest populations in the comparative dataset, Papuans from New Guinea and the Nasioi from Bougainville in the Solomon Islands. The Papuan and Nasioi groups speak non-Austronesian languages and carry relatively few Austronesian genetic lineages, thus hinting that they may have been isolated since well before the Neolithic farming expansion (Friedlaender et al. 2008). It therefore remains possible for genetic diversity to be reduced over extremely long time depths, but at least on Sumba, where we have a much clearer understanding of the island's history, over 4,000 years of relative isolation from surrounding islands, and to a lesser extent between neighboring villages on Sumba itself, does not appear to have affected levels of genetic diversity within communities to any appreciable extent.

This should not be taken to imply that genetic isolation has no consequences. Founder events out of Africa clearly reduced genetic diversity in descendent Eurasian populations, slightly increasing numbers of deleterious mutations (Henn et al. 2016). Within Island Southeast Asia, strong pressure to marry within a community (endogamy) on Bali is likely the cause of a common recessive genetic defect that produces deafness in a small, but significant, proportion of the community (Winata et al. 1995) (the number is large enough that the community has invented an indigenous sign language to facilitate communication; Marsaja 2008). In contrast, compliance with complex marriage rules in one of our communities (Rindi), which if followed strictly would lead to a reduction in genetic diversity, appears to be more relaxed in practice (Guillot et al. 2015).

The bigger point is that human communities are capable of sustaining high levels of genetic diversity even after thousands of years of relative isolation, both with external regions (surrounding islands in the case of Sumba) and neighboring communities. Genetic diversity can even be maintained in the face of extensive linguistic boundaries. In part, this probably reflects a separation of time scales. Populations can remain connected genetically as long as they share just a few migrants over a large number of generations (which, in humans, can rapidly sum to hundreds of years). In contrast, languages and culture can change much more quickly, and cultural divergence more closely mirrors the extent of social contact between communities as opposed to occasional intermarriage. The different time scales at which these two processes act—biological evolution slowly, cultural evolution more quickly—helps explain how even small islands like Sumba can become so linguistically diverse, while having relatively little impact on genetic structure and levels of genetic diversity.

Importantly, ways of life similar to those observed on Sumba today were once common following the Neolithic, including in places like Europe, which are now heavily impacted by later waves of population movements and restructuring caused by the emergence of large states and other post-Neolithic processes. Nevertheless, it is within community structures broadly of this type that most recent human evolution actually occurred. Our analysis of Sumba shows that loss of genetic diversity is not necessarily an outcome of long-term community isolation within loosely connected population networks structured on a microgeographic scale (i.e., tens to hundreds of kilometers). This is important because communities with similar population structures were once a key feature of most global human groups.

Materials and Methods

Ethics

Biological samples were collected by J.S.L., H.S., and a team from the Eijkman Institute for Molecular Biology, with the assistance of Indonesian Public Health clinic staff, following protocols for the protection of human subjects established by the Eijkman Institute, Nanyang Technological University and

the University of Arizona institutional review boards. Permission to conduct research in Indonesia was granted by the State Ministry of Research and Technology.

Sampling and Genetic Screening

Genetic markers were screened in 235 consenting and apparently healthy individuals from eight communities on the eastern Indonesian island of Sumba: Anakalang, Kodi, Lamboya, Loli, Mamboro, Rindi, Wanokaka, and Wunga (supplementary table S1, Supplementary Material online). Seven new samples from west Timor (Kamanasa, $n = 4$; and Umanen Lawalu, $n = 3$) are also reported. Apart from excluding known close relatives, individuals were approached randomly during the course of community-based medical visits. Participant interviews confirmed ethnic, linguistic, and geographic affiliations with local communities for at least two generations into the past. A set of 716,503 SNPs was screened in all individuals using the Illumina Human OMNI Express-24 BeadChip (genotyping by GeneByGene, Houston, TX) and 695,789 autosomal SNPs were extracted for further analysis. Two samples with >5% missing genotypes were excluded. Inference of cryptic relationships between samples was performed using KING v. 1.4 (Manichaikul et al. 2010) and first-degree relatives with a kinship coefficient >0.354 (following the software guidelines) were removed from the dataset by randomly deleting one individual of each related pair. Our final dataset included 204 Sumbanese (supplementary table S1, Supplementary Material online) and seven Timorese samples that passed all filtering criteria. Data for all successfully genotyped samples, including those previously used for an in-depth study of marriage patterns in the community of Rindi (Guillot et al. 2015), are available on the NCBI GEO repository (project accession number: GSE76645).

Comparative Datasets

The Sumba dataset was merged with autosomal data from the following genotyping datasets: the global Human Genome Diversity Project (HGDP) (Li et al. 2008), Papuan and Philippine samples from Migliano et al. (2013), Borneo and Sulawesi samples from Pierron et al. (2014), and Philippine Aeta samples from Rasmussen et al. (2011). The following samples were each combined into a single group: all non-Negrito samples from the Philippines, and all samples from New Guinea. In addition, the positions of genotyped autosomal SNPs were extracted from Phase 3 of the 1000 Genomes Project data (2 May 2013 release) (1000 Genomes Project Consortium 2012) and merged into the combined dataset. To minimize merging errors between different datasets, all A/T and C/G polymorphisms and multiallelic polymorphisms were removed. The final combined dataset included 3,885 samples and 360,452 SNPs after removing SNPs with >5% missing data (supplementary table S6, Supplementary Material online). The coefficient of inbreeding, within-island F_{ST}, PCA, and ADMIXTURE analyses were performed after additionally removing highly linked SNPs with $R^2 \geq 0.2$ (–indep-pairwise 50 5 0.2 in Plink v. 1.90 beta; Chang et al. 2015).

A core set of samples from Sumba, Timor, and neighboring regions in Southeast Asia (Borneo, Cambodia, the Philippines, Sulawesi and Vietnam to the west, and New Guinea and Bougainville to the east) was used in all analyses. Additional samples from a wider set of comparative data were added to individual analyses as follows. First, to compare levels of inbreeding and to place Sumba populations within a broader global framework of human genetic diversity, we calculated the coefficient of inbreeding, the number and length of ROH, genotype-based mean pairwise nucleotide diversity, and haplotype-based gene diversity for the core dataset together with a worldwide dataset comprising the complete HGDP panel. Second, to infer individual ancestries and the proportion of the genome shared between Sumba and its geographical neighbors, we used the core dataset and regional data: for ADMIXTURE analysis, we added four mainland East Asian proxy populations (CDX, CHB, CHS, and JPT from the 1000 Genomes Project); and for the geographically fine-scale IBD analysis, we added the diverse range of East Asian samples from the HGDP panel. Additional details on the sample sets used for each analysis and the detailed description of the comparative dataset are given in supplementary table S6, Supplementary Material online.

Complete genome sequences were obtained for five East Asian populations (Chinese CDX, CHB and CHS, Japanese JPT and Vietnamese KHV) from the 1000 Genomes Project (1000 Genomes Project Consortium 2012). This dataset was used to assess the effects of the ascertainment bias of the genotyping array on genetic diversity and to confirm interpretations about pairwise genetic diversity estimates inferred from the combined genotyping dataset described above. Calculations were performed with (1) the entire dataset and (2) SNPs with minor allele frequency <5% removed, the latter mimicking one aspect of the ascertainment bias of SNP arrays. For comparability with the SNP array data, diversity values were estimated only from polymorphic sites (i.e., invariant sites were excluded).

Population Genetic Analyses

The coefficient of inbreeding (F_{IS}) was calculated using GENEPOP v. 4.4 (Rousset 2008). F_{ST}, a measure of population differentiation, was calculated for newly reported Sumba samples with filtFst v. 1.0 (http://mpcox.github.io/filtFst) using the method of Hudson et al. (1992) with a correction for unequal sample sizes, as described by Plagnol and Wall (2006).

PCA of Sumba samples was performed with the smartpca function of EIGENSOFT v. 3.0 (Patterson et al. 2006). To determine the robustness of village clusters, we used a jack-knife procedure of randomly excluding either one sample or 10% of samples from every Sumba population and performing PCA on a new sample set. One thousand iterations of the jack-knife were performed and no outlier removal step was applied in EIGENSOFT to keep the sample set consistent among all runs. Only the first two principal components were considered for further calculations. For each of the eight Sumba villages, we counted how many times the average pairwise Euclidian distance between samples within the test village was smaller than the average pairwise Euclidian distance

between samples of the test village and each of the remaining seven Sumba villages. The proportions of these runs were used to assess robustness.

Mean pairwise nucleotide diversity π was calculated using the method of Nei and Li (1979) with rare alleles (minor allele frequency <5%) removed. In addition, gene diversity (H) (Nei 1987), the probability that two randomly chosen haplotypes in the sample are different, was calculated to minimize the effect of the genotyping array ascertainment bias, which could affect π estimates, using the modified procedure of Verdu et al. (2014). Genotypes were first phased with SHAPEIT v. 2 (Delaneau et al. 2014) using the HapMap phase II b37 recombination map (International HapMap et al. 2007). Next, haplotype blocks with low recombination (<0.5 cM/Mb between each consecutive SNP pair) and ranging in size between 5 and 15 SNPs were extracted from the phased data. If an interval between a pair of adjacent SNPs in the genotyping dataset included additional positions reported in the HapMap phase II data, then the average recombination rate across all HapMap SNPs that mapped to this genomic region was taken. To avoid overlap between windows and thus sampling the same SNP multiple times, the combined minimum rate between two adjacent windows was set to exceed 0.5 cM/Mb. Out of 360,452 SNPs, the proportion assigned to haplotype blocks with low overall recombination rate varied between 16% (11,484 haplotype blocks with a block size of 5 SNPs) and 3% (677 haplotype blocks with a block size of 15 SNPs). The mean number of blocks per chromosome varied from 522 $\pm$ SD 279 for 5-SNP haplotype blocks to 31 $\pm$ SD 17 for 15-SNP haplotype blocks. Gene diversity (H) was calculated using the defined haplotype blocks, and estimates were averaged across all chromosomes and the whole range of window sizes.

ROH were calculated within individuals using Plink v. 1.90 beta (Chang et al. 2015). Following Howrigan et al. (2011), the minimum number of homozygous SNPs to call a run was varied from 20 to 95 with increments of 15 SNPs and three different linkage disequilibrium pruning parameters—'light', 'moderate', and 'heavy' (variance inflation factor >10, 2 and 1.1, respectively).

IBD regions longer than 1 cM were determined using the Refined IBD algorithm implemented in Beagle v. 4.1 after excluding SNPs with a minor allele frequency <1% (Browning and Browning 2013).

Maximum likelihood estimation of individual ancestries was performed with ADMIXTURE v. 1.30 (Alexander et al. 2009). One hundred randomly seeded runs were performed for each number of ancestral populations ($K = 2$–5 for Sumba only and $K = 2$–7 for Sumba and other populations), and the results within each K were summarized with CLUMPP v. 1.1.2 (Jakobsson and Rosenberg 2007). Runs with symmetric similarity coefficient >0.9 were assigned to the same modal solution, following Verdu et al. (2014). Individual ancestry proportions were averaged across runs that belong to the same mode.

Great circle geographical distances were obtained using the Geographic Distance Matrix Generator v. 1.2.3 (http://biodiversityinformatics.amnh.org/open_source/gdmg).

Associations between genetic and geographic distances were determined using the Mantel test with all possible permutations as implemented in the R package ade4 v. 1.7-3 (Dray and Dufour 2007). The natural log of geographic distances and scaled F_{ST} distances ($F_{ST}/1 - F_{ST}$) were used as per Rousset (1997). Associations between genetic distances and community pairs speaking languages in the same or different language groups, as defined by Lansing et al. (2007), were determined by Monte Carlo permutation. Holding the genetic distance matrix constant, shared and nonshared language states were permuted 10^6 times. A one-tailed probability was returned showing when the difference in mean F_{ST} values within versus between language groups equaled or exceeded the observed value. An AMOVA was calculated using Arlequin v. 3.5 (Excoffier and Lischer 2010).

Modeling and Inference of Migration Rates

Genome-wide SNP data were simulated using the sequential coalescent software MaCS (6 July 2015 version) (Chen et al. 2009). The model for Sumba implemented here directly expands on a more general model of Asian–Melanesian admixture developed by Sanderson et al. (2015), which is in turn based on demographic information previously inferred for populations across the Indo-Pacific region (Cox et al. 2008, 2010; Xu et al. 2012). In brief, an ancestral Eurasian population diverged 50,000 years ago (25-year generation interval) to form two daughter groups, which subsequently evolved into the Asian and Melanesian parental populations. These groups merged on Sumba 4,085 years ago (Xu et al. 2012) to form an admixed population with 74% Asian ancestry and 26% Melanesian ancestry (Cox et al. 2010). Linguistic, genetic, and cultural evidence show that this ancestral Sumba population radiated rapidly (Lansing et al. 2007). We model eight daughter communities that exchanged individuals to the present with some migration rate m. To fit the simulated and observed data, 30 individuals were sampled per community, and SNP variants were obtained by simulating 100 Mb of genomic sequence and randomly selecting polymorphisms to match the observed site frequency spectrum of the SNP array data, as in Sanderson et al. (2015). The mean migration rate m was inferred within an approximate Bayesian computation (ABC) setting. Initially, 1×10^4 simulations were run with migration rates drawn from the full uniform prior $U(0, 1)$. As accepted values were only observed in a small part of this range, the prior was subsequently reduced to $U(0, 0.03)$ and a longer ABC was run with 1×10^5 simulations taking $\gg$1 CPU year. F_{ST} was calculated for all 28 community pairs and the mean value compared with the observed data. ABC, including cross validation with median inferred values, was performed using local linear regression and a tolerance value of 0.05 with the R package abc v. 2.1 (Csilléry et al. 2012).

Supplementary Material

Supplementary figures S1–S6 and tables S1–S6 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This research was supported by Nanyang Technological University (2013, "Transition Pathways: Islands of Order") and the Singapore Ministry of Education (2015, "Co-Evolution of Humans and Infectious Diseases in Island Southeast Asia") to J.S.L., and by the Royal Society of New Zealand through a Rutherford Fellowship (RDF-10-MAU-001) to M.P.C. Computational resources were provided by Massey University and the High Performance Computing Center, University of Tartu, Estonia. G.H. was supported by a NEFREX grant funded by the European Union (People Marie Curie Actions, International Research Staff Exchange Scheme, call FP7-PEOPLE-2012-IRSES-number 318979).

References

1000 Genomes Project Consortium 2010. A map of human genome variation from population-scale sequencing. *Nature* 467:1061–1073.

1000 Genomes Project Consortium 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56–65.

1000 Genomes Project Consortium 2015. A global reference for human genetic variation. *Nature* 526:68–74.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 19:1655–1664.

Badan Pusat Statistik. 2010. Sensus Penduduk 2010. [cited 2016 May 20]. Available from: http://sp2010.bps.go.id

Bellwood P. 2013. First migrants: ancient migration in global perspective. Oxford, UK: Wiley Blackwell.

Boyce AJ, Harrison GA, Platt CM, Hornabrook RW, Serjeantson S, Kirk RL, Booth PB. 1978. Migration and genetic diversity in an island population: Karkar, Papua New Guinea. *Proc R Soc B* 202:269–295.

Browning BL, Browning SR. 2013. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* 194:459–471.

Chang C, Chow C, Tellier L, Vattikuti S, Purcell S, Lee J. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* 4:7.

Chen GK, Marjoram P, Wall JD. 2009. Fast and flexible simulation of DNA sequence data. *Genome Res*. 19:136–142.

Cox MP, Karafet TM, Lansing JS, Sudoyo H, Hammer MF. 2010. Autosomal and X-linked single nucleotide polymorphisms reveal a steep Asian-Melanesian ancestry cline in eastern Indonesia and a sex bias in admixture rates. *Proc R Soc B* 277:1589–1596.

Cox MP, Woerner AE, Wall JD, Hammer MF. 2008. Intergenic DNA sequences from the human X chromosome reveal high rates of global gene flow. *BMC Genet*. 9:e76.

Csilléry K, François O, Blum MGB. 2012. abc: an R package for approximate Bayesian computation (ABC). *Methods Ecol Evol*. 3:475–479.

Delaneau O, Marchini J, 1000 Genomes Project Consortium, Genomes Project Consortium. 2014. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat Commun*. 5:3934.

Dray S, Dufour AB. 2007. The ade4 package: implementing the duality diagram for ecologists. *J Stat Softw*. 22:1–20.

Ewens WJ. 2004. Mathematical population genetics. I. Theoretical introduction. New York: Springer.

Excoffier L, Lischer HEL. 2010. Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Res*. 10:564–567.

Forth GL. 1981. Rindi. The Hague: Martinus Nijhoff.

Friedlaender JS. 1975. The demography, genetics, and phenetics of Bougainville islanders. Cambridge: Harvard University Press.

Friedlaender JS, Friedlaender FR, Reed FA, Kidd KK, Kidd JR, Chambers GK, Lea RA, Loo J-H, Koki G, Hodgson JA, et al. 2008. The genetic structure of Pacific Islanders. *PLoS Genet*. 4:e19.

Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, The 1000 Genomes Project, Bustamante CD. 2011. Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci U S A*. 108:11983–11988.

Guillot EG, Hazelton ML, Karafet TM, Lansing JS, Sudoyo H, Cox MP. 2015. Relaxed observance of traditional marriage rules allows social connectivity without loss of genetic diversity. *Mol Biol Evol*. 32:2254–2262.

Guillot EG, Tumonggor MK, Lansing JS, Sudoyo H, Cox MP. 2013. Climate change influenced female population sizes through time across the Indonesian archipelago. *Hum Biol*. 85:135–152.

Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, Karthikeyan S, Iles L, Pollard MO, Choudhury A, et al. 2015. The African Genome Variation Project shapes medical genetics in Africa. *Nature* 517:327–332.

Henn BM, Botigué LR, Peischl S, Dupanloup I, Lipatov M, Maples BK, Martin AR, Musharoff S, Cann H, Snyder MP, et al. 2016. Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. *Proc Natl Acad Sci U. S. A*. 113:E440–E449.

Hoskins J. 1993. The play of time: Kodi perspectives on calendars, history and exchange. Berkeley, CA: University of California Press.

Howrigan D, Simonson M, Keller M. 2011. Detecting autozygosity through runs of homozygosity: a comparison of three autozygosity detection algorithms. *BMC Genomics* 12:460.

Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from DNA sequence data. *Genetics* 132:583–589.

Hunley K. 2015. Reassessment of global gene-language coevolution. *Proc Natl Acad Sci U S A*. 112:1919–1920.

International HapMap C, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449:851–861.

Jakobsson M, Rosenberg NA. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 23:1801–1806.

Karmin M, Saag L, Vicente M, Wilson Sayres MA, Jarve M, Talas UG, Rootsi S, Ilumae AM, Magi R, Mitt M, et al. 2015. A recent bottleneck of Y chromosome diversity coincides with a global change in culture. *Genome Res*. 25:459–466.

Kusuma P, Cox MP, Brucato N, Sudoyo H, Letellier T, Ricaut FX. Forthcoming 2016. Western Eurasian genetic influences in the Indonesian Archipelago. *Quat Int*.

Lansing JS, Cox MP, Downey SS, Gabler B, Hallmark B, Karafet TM, Norquest P, Schoenfelder JW, Sudoyo H, Watkins JC, et al. 2007. Coevolution of languages and genes on the island of Sumba, eastern Indonesia. *Proc Natl Acad Sci U S A*. 104:16022–16026.

Lansing JS, Watkins JC, Hallmark B, Cox MP, Karafet TM, Sudoyo H, Hammer MF. 2008. Male dominance rarely skews the frequency distribution of Y chromosome haplotypes in human populations. *Proc Natl Acad Sci U S A*. 105:11645–11650.

Leslie S, Winney B, Hellenthal G, Davison D, Boumertit A, Day T, Hutnik K, Royrvik EC, Cunliffe B, Wellcome Trust Case Control Consortium, et al. 2015. The fine-scale genetic structure of the British population. *Nature* 519:309–314.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman MW, Cavalli-Sforza LL, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319:1100–1104.

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. 2010. Robust relationship inference in genome-wide association studies. *Bioinformatics* 26:2867–2873.

Marsaja IG. 2008. Desa Kolok—a deaf village and its sign language in Bali, Indonesia. Nijmegen: Ishara Press.

Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, Harney E, Stewardson K, Fernandes D, Novak M, et al. 2015. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* 528:499–503.

McVean G, Spencer CC, Chaix R. 2005. Perspectives on human genetic variation from the HapMap Project. *PLoS Genet*. 1:e54.

Migliano AB, Romero IG, Metspalu M, Leavesley M, Pagani L, Antao T, Huang DW, Sherman BT, Siddle K, Scholes C, et al. 2013. Evolution of the pygmy phenotype: evidence of positive selection from genome-wide scans in African, Asian, and Melanesian pygmies. *Hum Biol*. 85:251–284.

Nei M. 1987. Molecular evolutionary genetics. New York: Columbia University Press.

Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A*. 76:5269–5273.

Pääbo S. 2015. The diverse origins of the human gene pool. *Nat Rev Genet*. 16:313–314.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet*. 2:e190.

Pierron D, Razafindrazaka H, Pagani L, Ricaut FX, Antao T, Capredon M, Sambo C, Radimilahy C, Rakotoarisoa JA, Blench RM, et al. 2014. Genome-wide evidence of Austronesian-Bantu admixture and cultural reversion in a hunter-gatherer group of Madagascar. *Proc Natl Acad Sci U S A*. 111:936–941.

Plagnol V, Wall JD. 2006. Possible ancestral structure in human populations. *PLoS Genet*. 2:e105.

Rasmussen M, Guo X, Wang Y, Lohmueller KE, Rasmussen S, Albrechtsen A, Skotte L, Lindgreen S, Metspalu M, Jombart T, et al. 2011. An aboriginal Australian genome reveals separate human dispersals into Asia. *Science* 334:94–98.

Rousset F. 1997. Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics* 145:1219–1228.

Rousset F. 2008. genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Mol Ecol Resour*. 8:103–106.

Sanderson J, Sudoyo H, Karafet TM, Hammer MF, Cox MP. 2015. Reconstructing past admixture processes from local genomic ancestry using wavelet transformation. *Genetics* 200:469–481.

Shennan S, Downey SS, Timpson A, Edinborough K, Colledge S, Kerig T, Manning K, Thomas MG. 2013. Regional population collapse followed initial agriculture booms in mid-Holocene Europe. *Nat Commun*. 4:2486.

Slatkin M. 1987. Gene flow and the geographic structure of natural populations. *Science* 236:787–792.

Soares PA, Trejaut JA, Rito T, Cavadas B, Hill C, Khong Eng K, Mormina M, Brandão A, Fraser RM, Wang T-Y, et al. 2016. Resolving the ancestry of Austronesian-speaking populations. *Hum Genet*. 135:309–326.

Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, Coe BP, Baker C, Nordenfelt S, Bamshad M, et al. 2015. Global diversity, population stratification, and selection of human copy-number variation. *Science* 349:aab3761.

Tumonggor MK, Karafet TM, Downey S, Lansing JS, Norquest P, Sudoyo H, Hammer MF, Cox MP. 2014. Isolation, contact and social behavior shaped genetic diversity in West Timor. *J Hum Genet*. 59:494–503.

Tumonggor MK, Karafet TM, Hallmark B, Lansing JS, Sudoyo H, Hammer MF, Cox MP. 2013. The Indonesian archipelago: an ancient genetic highway linking Asia and the Pacific. *J Hum Genet*. 58:165–173.

Verdu P, Pemberton TJ, Laurent R, Kemp BM, Gonzalez-Oliver A, Gorodezky C, Hughes CE, Shattuck MR, Petzelt B, Mitchell J, et al. 2014. Patterns of admixture and population structure in native populations of Northwest North America. *PLoS Genet*. 10:e1004530.

Winata S, Arhya IN, Moeljopawiro S, Hinnant JT, Liang Y, Friedman TB, Asher JH. Jr. 1995. Congenital non-syndromal autosomal recessive deafness in Bengkala, an isolated Balinese village. *J Med Genet*. 32:336–343.

Xu S, Pugach I, Stoneking M, Kayser M, Jin L, The Hugo Pan-Asian SNP Consortium. 2012. Genetic dating indicates that the Asian-Papuan admixture through Eastern Indonesia corresponds to the Austronesian expansion. *Proc Natl Acad Sci U S A*. 109:4574–4579.