Complex Patterns of Admixture across the Indonesian Archipelago

Georgi Hudjashov,*[,1,2] Tatiana M. Karafet,[3] Daniel J. Lawson,[4] Sean Downey,[5] Olga Savina,[3] Herawati Sudoyo,[6,7,8] J. Stephen Lansing,[9] Michael F. Hammer,[2] and Murray P. Cox*[,1]

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

[2]Estonian Biocentre, 51010 Tartu, Estonia

[3]ARL Division of Biotechnology, University of Arizona, Tucson, AZ

[4]School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom

[5]Department of Anthropology, University of Maryland, College Park, MD

[6]Genome Diversity and Diseases Laboratory, Eijkman Institute for Molecular Biology, Jakarta, Indonesia

[7]Department of Medical Biology, Faculty of Medicine, University of Indonesia, Jakarta, Indonesia

[8]Sydney Medical School, University of Sydney, Sydney, NSW, Australia

[9]Complexity Institute, Nanyang Technological University, Singapore

*Corresponding authors: E-mails: g.hudjashov@massey.ac.nz; m.p.cox@massey.ac.nz.

Associate editor: John Novembre

Abstract

Indonesia, an island nation as large as continental Europe, hosts a sizeable proportion of global human diversity, yet remains surprisingly undercharacterized genetically. Here, we substantially expand on existing studies by reporting genome-scale data for nearly 500 individuals from 25 populations in Island Southeast Asia, New Guinea, and Oceania, notably including previously unsampled islands across the Indonesian archipelago. We use high-resolution analyses of haplotype diversity to reveal fine detail of regional admixture patterns, with a particular focus on the Holocene. We find that recent population history within Indonesia is complex, and that populations from the Philippines made important genetic contributions in the early phases of the Austronesian expansion. Different, but interrelated processes, acted in the east and west. The Austronesian migration took several centuries to spread across the eastern part of the archipelago, where genetic admixture postdates the archeological signal. As with the Neolithic expansion further east in Oceania and in Europe, genetic mixing with local inhabitants in eastern Indonesia lagged behind the arrival of farming populations. In contrast, western Indonesia has a more complicated admixture history shaped by interactions with mainland Asian and Austronesian newcomers, which for some populations occurred more than once. Another layer of complexity in the west was introduced by genetic contact with South Asia and strong demographic events in isolated local groups.

Key words: **genetic diversity, population structure, human migration, genetic admixture.**

Introduction

Indonesia, the world's fourth largest country by population, comprises an archipelago of about 900 permanently inhabited islands in tropical Asia, and hosts an astonishing array of human diversity that remains largely underrepresented in modern biological surveys (Horton 2016). Of the >700 languages still spoken in Indonesia, most belong to the Austronesian (AN) language family, with some Papuan languages present in the east, making Indonesia, together with Papua New Guinea, the most linguistically diverse region on earth (Lewis et al. 2016). The Indonesian archipelago, the Philippines and Taiwan form Island Southeast Asia (ISEA), a maritime region unique for its key role in both the early and recent evolution of *Homo sapiens* outside of Africa. It has one of the first traces of anatomically modern humans in Eurasia, possibly dating as early as 67 ka (Barker et al. 2007; Mijares et al. 2010); archaic *H. floresiensis* likely coexisted with modern humans here (Sutikna et al. 2016); and eastern Indonesian populations are among the few living groups showing substantial traces of archaic introgression from Denisovans (Reich et al. 2011). More recently, Indonesia, and ISEA more generally, was ground zero for the spread of Neolithic culture by AN speakers. Advancing maritime technologies allowed farming populations to treat this region as a springboard to reach oceanic islands as remote as Madagascar, Hawaii, Easter Island, and New Zealand.

Linguistic, archeological and genetic evidence all point to Taiwan as the most likely origin of expanding AN speakers, whose demic spread began 2500–2000 BCE (Gray et al. 2009; Bellwood 2014; Ko et al. 2014). Whether these people were strict agriculturalists or practiced a more complex range of subsistence strategies remains unclear (Blench 2012), but the

Open Access

Article

Fast Track

Neolithic items that appeared at this time include specific red-slipped pottery, stone barkcloth beaters, new types of stone adzes, and widely traded tools and ornaments made from eastern Taiwanese nephrite. The Austronesian expansion spread rapidly across ISEA, reaching the Philippines by 2000–1500 BCE, and Borneo and Sulawesi by 1500–1000 BCE. Present in western Melanesia by 1350–750 BCE, it was followed by the settlement of the remote and previously uninhabited islands of the Pacific Ocean (Bellwood 2014).

Although this broad history is now well known, multiple lines of evidence suggest that the Neolithic transition in ISEA was more complex than a simple movement of genes, languages, and technology solely out of Taiwan. PreAustronesian linguistic substrates in Indonesia show influences from mainland Southeast Asia (MSEA) (Blench 2010), and domesticated pigs also likely spread from the mainland to the islands (Larson et al. 2007). New Guinea, an independent domestication center focused on fruits and tubers, was itself an important hub for innovation, with new cultivars such as bananas spreading from east to west (Denham and Donohue 2009; Spriggs 2012).

Debate has traditionally revolved around whether the AN dispersal was primarily a movement of people, accompanied by admixture with local populations, or instead driven by transfers of language, culture and technology. Haploid loci (mtDNA and Y chromosome) reinforce the presence of preAustronesian contact between mainland and island Southeast Asian populations, but hint at only minor contributions from Neolithic newcomers (Karafet et al. 2010; Jinam et al. 2012; Tumonggor et al. 2013; Gomes et al. 2015; Soares et al. 2016). However, the limits of uniparental markers (Pugach and Stoneking 2015) are increasingly being circumvented by genome-wide autosomal data. This includes ancient DNA from the first settlers in Vanuatu and Tonga, where the genomes of individuals dated to 1100–300 BCE suggest that the first Austronesian migrants arriving in Remote Oceania had little to no admixture with Papuan groups (Skoglund et al. 2016).

Unexpectedly, the autosomal genetic variation in ISEA, and Indonesia in particular, is poorly characterized, especially given the extraordinary extent of its human diversity. Current studies on the history of ISEA are all based on a small number of genome-wide polymorphisms (*ca* 55k) screened by the HUGO Pan-Asian SNP Consortium or are geographically restricted (HUGO Pan-Asian SNP Consortium 2009; Xu et al. 2012; Lipson et al. 2014; Sedghifar et al. 2015; Kusuma, Brucato, et al. 2016; Mörseburg et al. 2016; Soares et al. 2016). Estimated dates of admixture between incoming AN and local populations span a wide timeframe, ranging from 3800–1500 BCE (Xu et al. 2012; Sanderson et al. 2015; Sedghifar et al. 2015) to 550 CE (Lipson et al. 2014), likely reflecting unresolved dynamics in the admixture process. This is reinforced by a substantial proportion of Taiwan-related ancestry noted in both eastern and western Indonesia, coupled with evidence of genetic contact in western Indonesia between incoming AN populations and groups from Vietnam or peninsular Malaysia (Lipson et al. 2014). Later historical contacts—Indianization followed by Islamization (Ooi 2004)—add additional layers of genetic complexity (Kusuma, Cox, et al. 2016; Mörseburg et al. 2016).

Here, we present new genomic data for an extensive set of populations across ISEA and beyond, with a special focus on Indonesia. Spanning the whole of the archipelago from Sumatra in the west to New Guinea in the east, this high-resolution data set allows us to reconstruct fine-scale population structure within ISEA. We test whether the expansions from Taiwan and mainland Asia had different effects on eastern and western Indonesia, and address both spatial and temporal aspects of these migrations.

Results

Data

Here, we present a new genomic data set sampled from multiple locations in MSEA, ISEA, New Guinea, and Oceania. Genome-wide genotyping data were generated for 25 new communities, primarily focusing on the previously poorly represented Indonesian archipelago, including the islands of Sumatra, Nias, Mentawai, Java, Bali, Sulawesi, Sumba, Flores, Lembata, Alor, Pantar, and Timor. Genotyping quality control filters were passed for 498 new unrelated samples, for each of which ~510k SNPs are reported. Together with publicly available genomes from the Philippines, Borneo, and Sulawesi (Pagani et al. 2016), this data set represents the most comprehensive population-based collection of autosomal human genetic variation in ISEA (fig. 1*A* and supplementary table S1, Supplementary Material online).

Population Structure

To place these newly generated data points within broader regional and global contexts, we performed Principal Component Analysis (PCA) together with a selection of East and Southeast Asian genomes from our comparative data set (1000 Genomes Project Consortium 2012; Pagani et al. 2016). As with other studies (Xu et al. 2012; Kusuma, Brucato, et al. 2016), the first two principal components explain a large proportion of the total variance (53%, fig. 1*B*). Variation on the first axis can broadly be ascribed to the geographical location of populations from west to east, with populations from mainland Asia and Melanesia (Papuans from PNG, Baining, Nasioi, and Vanuatu) occupying the extremes. The second principal component stretches from two western Indonesian outliers, Nias and Mentawai, to the Baining, and represents a second axis of separation within the region. A measure of population substructure, F_{ST}, further supports the very high levels of genetic differentiation in our sample, between populations from MSEA, ISEA, and Melanesia in particular (supplementary table S2, Supplementary Material online). For example, the F_{ST} between Baining and MSEA/western ISEA (0.17 ± 0.02, mean $\pm$ SD) is comparable to that between African Yoruba and non-African groups (0.19 ± 0.03).

The ADMIXTURE analysis, a model-based approach that splits the ancestry of individual genomes into predefined

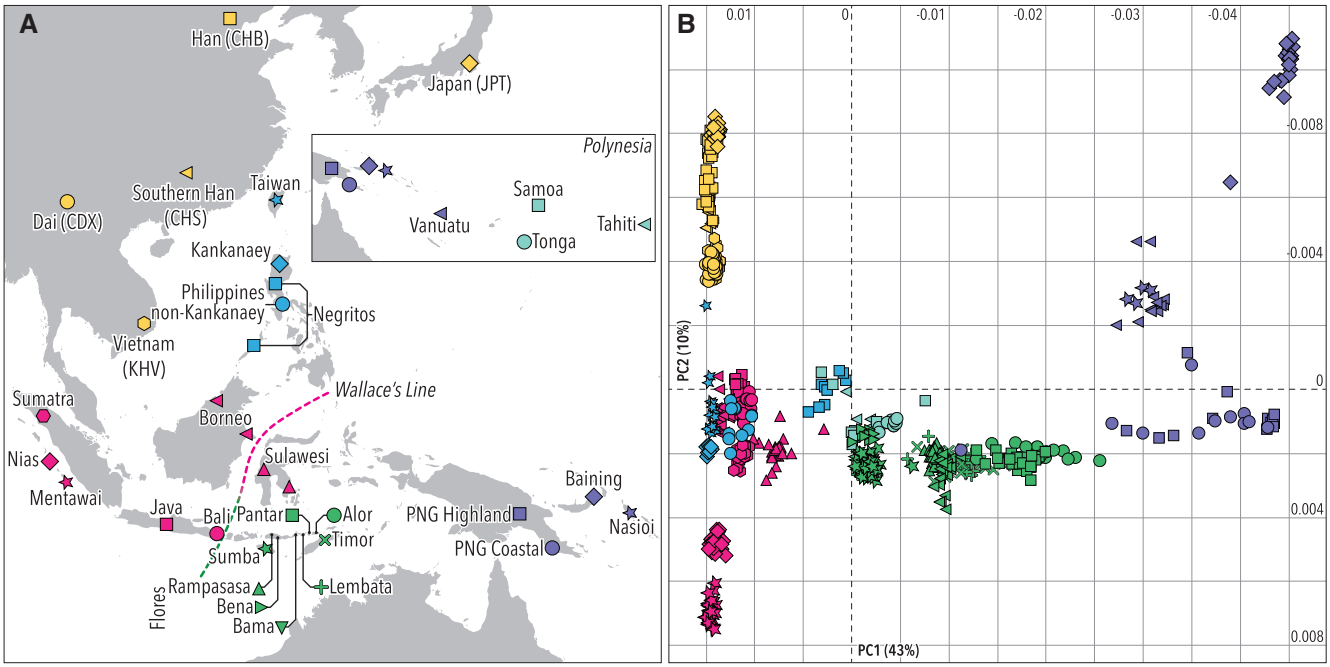

Fig. 1. Sampling locations and overview of genomic diversity. (*A*) Locations of new and published data from Mainland and Island Southeast Asia, Melanesia, and Polynesia used in the present study. Colors indicate regional affiliation of populations: yellow: mainland Asia, blue: Taiwan and the Philippines, magenta: western Indonesia, green: eastern Indonesia, purple: Melanesia, and turquoise: Polynesia. Detailed sample information is given in supplementary table S1, Supplementary Material online. (*B*) Principal Components Analysis of genome-wide SNP diversity in 703 individuals from the 32 populations shown in panel *A*. Note the different percentage variances explained by the two principal components.

ancestral components (K), further corroborates the PCA, and broadly matches the results of previous studies (Xu et al. 2012; Cox et al. 2016; Kusuma, Cox, et al. 2016; Mörseburg et al. 2016; Soares et al. 2016). Two main components can be observed within the region: mainland Asian (light yellow), and Papuan (light purple) (fig. 2). A third major component can be defined, and is commonly classified as AN (light blue) based on its high frequency in Taiwanese, the Kankanaey of the Philippines ($K = 12$) (Mörseburg et al. 2016), and Polynesians ($K = 5$–7) (Wollstein et al. 2010). There is a strong gradient between Papuan and mainland Asian/AN ancestries with a transition occurring slightly to the east of Wallace's line: eastern Indonesian islands have little to no mainland component (depending on the value of K) compared with Sumatra, Nias, Mentawai, Java, Bali, Borneo, and Sulawesi, which we broadly classify here as the western part of archipelago. Conversely, no Papuan ancestry is observed in western Indonesia. Except for Sumatra, and to a lesser extent Java, Bali and Negrito samples from the Philippines, which show evidence of historic gene flow from South Asia (red component), no other population has ancestry from outside MSEA, ISEA, or Papua at $K = 9$, which is the modal solution with the minimum average cross-validation score among 11 runs of ADMIXTURE ($K = 2$–12). At $K = 12$, a new component (gold) shared between mainland Southeast Asians (Chinese Dai and Vietnamese) and predominantly western Indonesian populations appears. The peak frequency of this component is found in Java and Bali. Most remaining ADMIXTURE components can be ascribed to populations that show effects of genetic drift (Baining at $K = 6$, Polynesians at $K = 8$, Mentawai at $K = 9$, Nias at $K = 10$; see evidence later). As with a previous study (Cox et al. 2016), we detect low levels of within island diversity, and adjacent islands (e.g., Alor and Pantar) are sometimes indistinguishable at the maximum K analyzed ($K = 12$).

Admixture Analysis

FINESTRUCTURE and GLOBETROTTER

To obtain deeper insight into fine-scale population structure within the new data set, we used the fineSTRUCTURE (FS) framework to capture information held by patterns of haplotype similarity. This method uses chromosome "painting" to assign nonrecombining portions (chunks) of an individual's genome (the recipient) to a set of different donors. Each recipient can therefore be represented as a mixture, or matrix of chunks, received or copied from every other donor. Similarities between the matrices are then used to cluster individuals into genetic groups.

The population structure inferred by FS reaffirms the ADMIXTURE results and has high individual branch support. We observe almost perfect clustering of samples from within the same population and/or island label (fig. 3 and supplementary figs. S1 and S2, Supplementary Material online). Multiple clusters can be broadly defined on the population dendrogram: 1) Taiwan, the Philippines, Borneo and Sulawesi, 2) western Indonesia (excluding Borneo and Sulawesi), 3) eastern Indonesia, 4) Melanesia, and 5) Polynesia (fig. 3 and supplementary fig. S1, Supplementary Material online). Individuals receive most of their genome from donors within the same cluster. As expected, self-copying, a process whereby

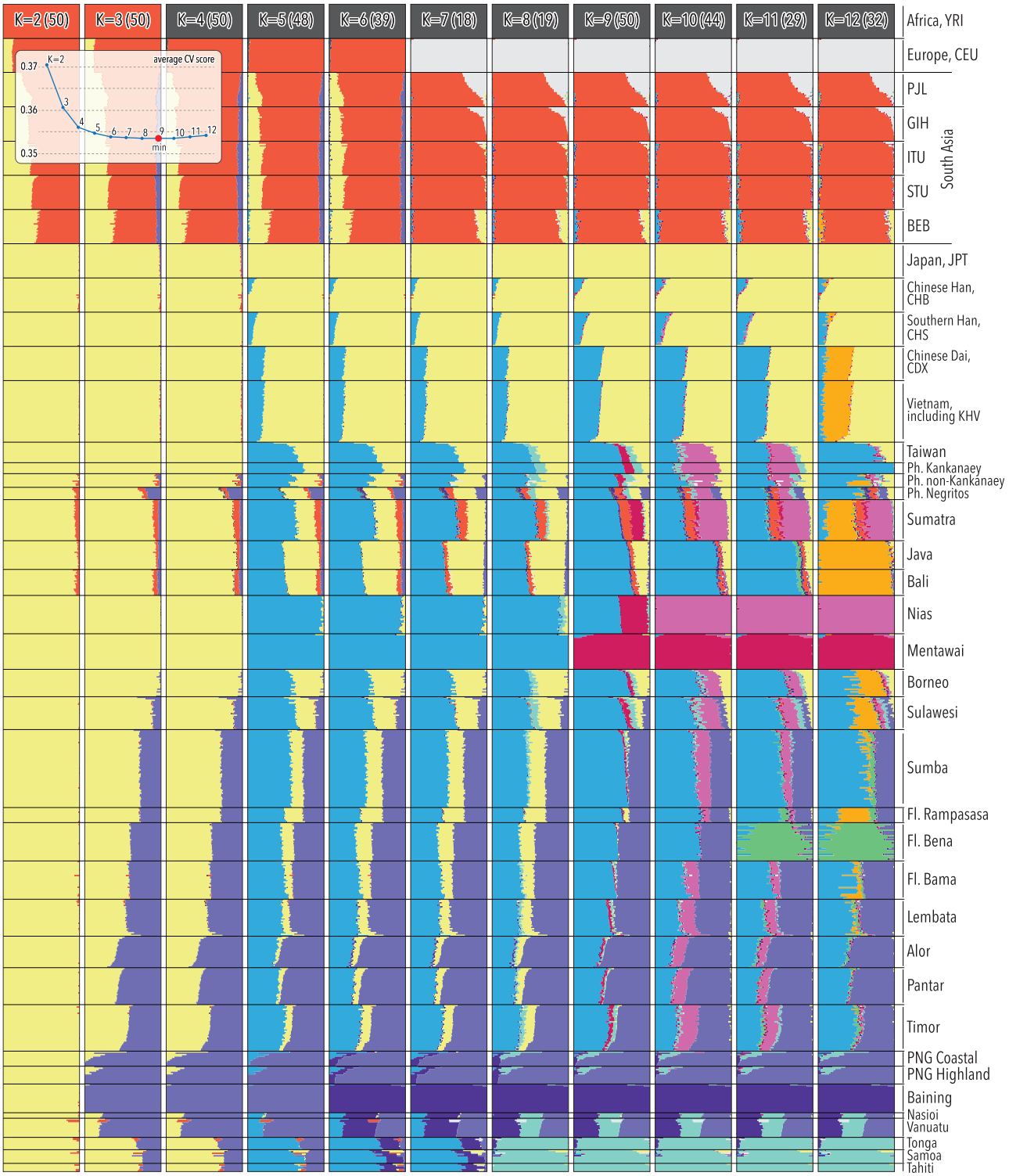

Fig. 2. Ancestral genomic components in regional populations. For every *K*, the modal solution with the highest number of ADMIXTURE runs is shown; individual ancestry proportions were averaged across all runs from the same mode and the number of runs (out of 50) assigned to the presented solution is shown in parentheses. Average cross validation statistics were calculated across all runs from the same mode (insert). The minimum cross-validation score is observed at $K = 9$. Note major ancestry components in Indonesia and ISEA—Papuan (light purple), mainland Asian (light yellow), and AN (light blue)—as well as major differences in the distribution of these three ancestries between eastern and western Indonesia. Populations from the Philippines and Flores are abbreviated as "Ph." and "Fl.," respectively.

recipients copy from their own population, prevails (Lawson et al. 2012). Additional mainland Asian and European donors cluster together into their respective continental groups.

After defining genetic clusters using FS (fig. 3 and supplementary fig. S1 and table S1, Supplementary Material online), we performed further analysis with GLOBETROTTER (GT) to determine which populations became admixed and to place

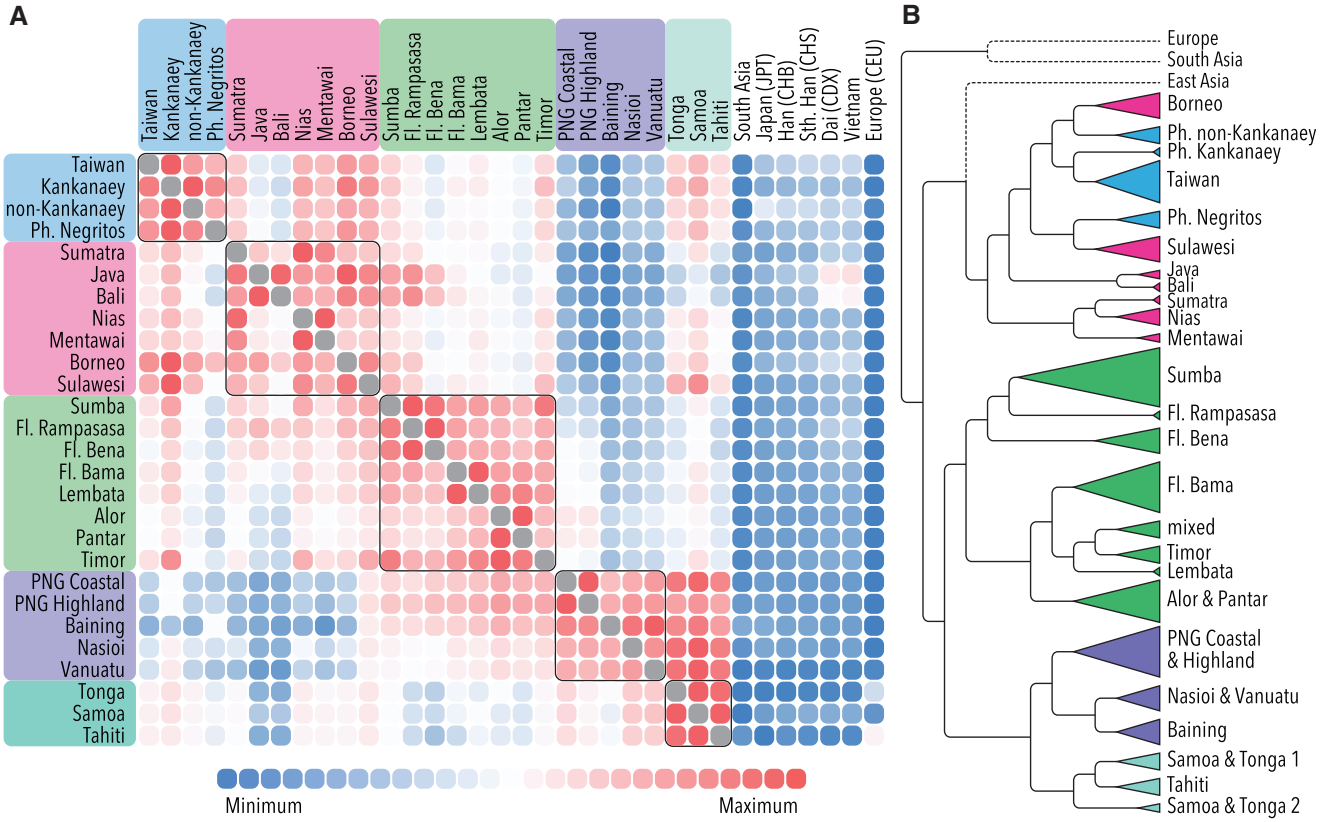

Fig. 3. Inference of population structure from haplotype sharing. (*A*) Simplified coancestry heatmap generated by chromosome painting with fineSTRUCTURE. Donor populations are shown in columns, recipients are in rows. Each cell depicts the average proportion of the genome copied by samples in the recipient population from the respective donor group, weighted by the average number of chunks received from this donor. Each row is scaled between 0 (blue, the minimum proportion of the genome copied for the given recipient group) to 1 (red, the maximum proportion of the genome copied). As expected, the largest proportion of the genome is copied from the population's own label. The diagonal (gray), representing self-copying, was excluded from the scaling procedure to emphasize copying patterns from other donor populations. Recipient groups from outside ISEA, Melanesia, and Polynesia are not shown. A detailed matrix of copying vectors is given in supplementary figure S2, Supplementary Material online. (*B*) Simplified dendrogram showing the clustering of individuals with similar copying vectors into genetic groups. These groups were subsequently used in the main GLOBETROTTER analysis. Clusters are shaded according to the color scheme from figure 1A. A detailed dendrogram is given in supplementary figure S1, Supplementary Material online; individual sample affiliations are given in supplementary table S1, Supplementary Material online.

estimated times on these admixture events. Unlike other methods (Loh et al. 2013), this allows the structure of unsampled source populations to be assessed (Hellenthal et al. 2014). First, a "full" approach was taken, whereby every recipient population could copy from any other donor group. As expected, in many inferred admixture events, sources of admixture were dominated by the geographical neighbors of the target population. This was especially apparent in eastern Indonesia, where one of the sources was a mixture of western Indonesian surrogates, and the other always included multiple eastern Indonesian islands (fig. 4A and supplementary table S3, Supplementary Material online). Although this could reflect a real admixture process between western and eastern parts of the archipelago, another highly likely explanation is that nearby groups simply have a similar population history, which would result in excess copying from neighboring islands (Hellenthal et al. 2014). To control for this effect, we performed a "regional" analysis (fig. 4B), where no self-copying was allowed within the five regions defined by geography: the Philippines, western Indonesia, eastern Indonesia, Melanesia, and Polynesia (supplementary fig. S1 and tables S1 and S4, Supplementary Material online).

Admixture events are inferred with high quality scores, with only a few events having goodness-of-fit (R^2) values <0.9 (supplementary table S3, Supplementary Material online). Several of the events fit well with historical information (fig. 4A, supplementary table S3, Supplementary Material online). First, South Asian admixture was detected in ISEA—around 1200 CE in Sumatra, 760 CE in Bali and 1060 CE in Philippine Negritos. Second, the Philippine non-Kankanaey group shows evidence of a complex admixture event with a European source around 1710 CE. This is consistent with the arrival of Western explorers in the archipelago starting from the 16th century (Ooi 2004). Third, the first European contact in Tonga, Samoa, and Tahiti occurred in the mid to late 18th century (Oliver 1974), which coincides with inferred dates for European admixture in Tahiti (1705 CE, 95% CI 1636–1772 CE) and the Tongan–Samoan cluster (1820 CE, 95% CI 1782–1865 CE).

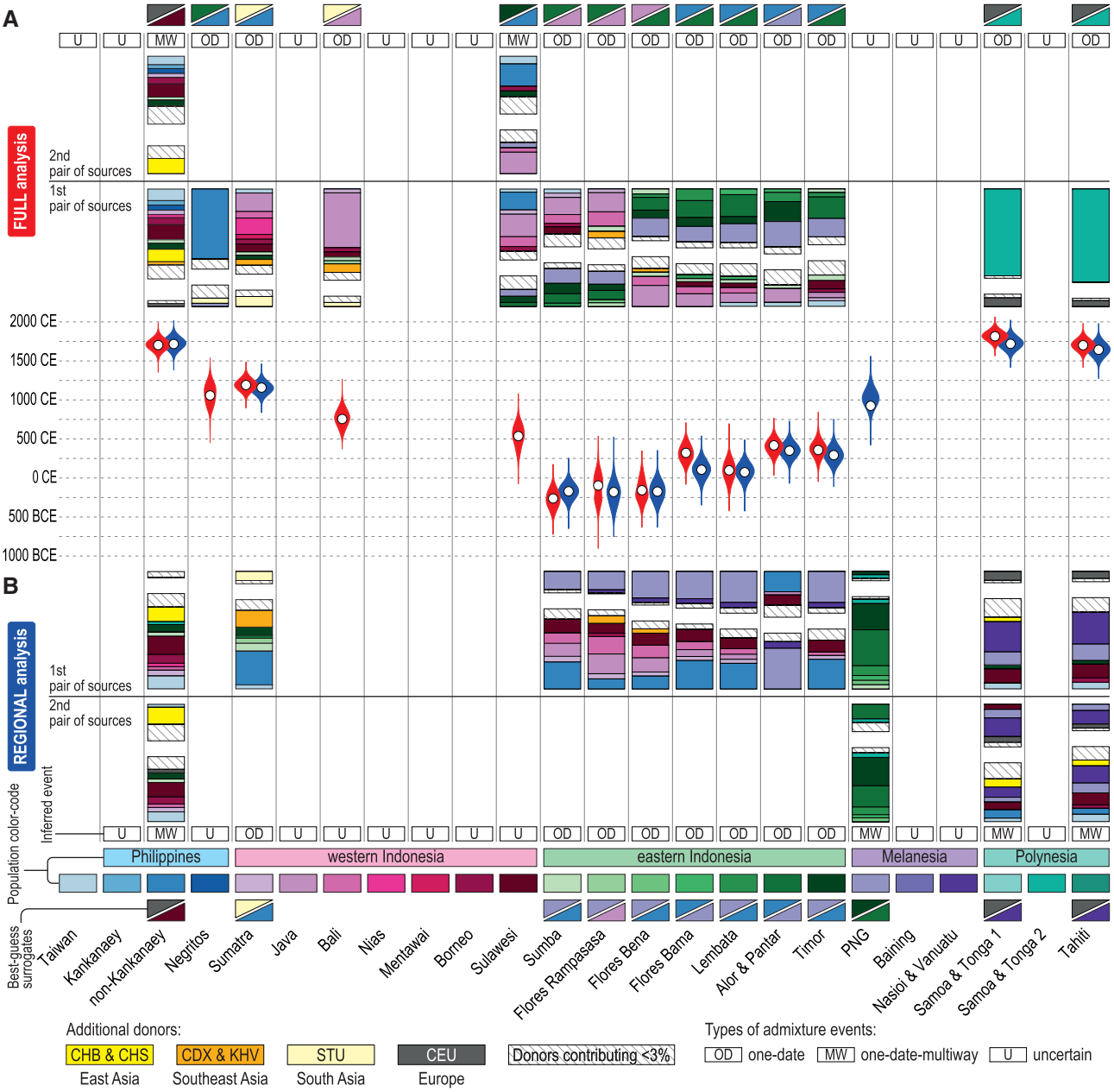

Fig. 4. Population admixture events and inferred contact dates as inferred by the main GT analysis. Individual raw FS populations were grouped into larger population clusters (see fineSTRUCTURE results, fig. 3*B* and supplementary fig. S1 and table S1, Supplementary Material online) for the GLOBETROTTER analysis. Inferred admixture events were classified into three types: one-date admixture (OD, involving a pair of mixing sources); one-date-multiway admixture (MW, involving two pairs of mixing sources); and uncertain (U). The inferred composition of mixing sources is shown as barplots (one pair for OD and two pairs for MW). A pair of "best-guess" surrogate populations that matches the inferred contributing sources for the primary admixture event are shown as shaded triangles. The bean plots in the center depict the distributions of admixture dates as estimated by 100 bootstrap replicates. Dates inferred by "full" and "regional" analysis are colored in red and blue, respectively. White circles represent point estimates of the date of admixture. The color scheme assigned to the additional mainland East Asian (CHB and CHS), Southeast Asian (CDX and KHV), South Asian (STU), and European (CEU) surrogates is shown at the bottom of the plot. Surrogates contributing <3% to mixing sources are hatched in grey. Detailed information about the inferred admixture events and composition of mixing sources is given in supplementary table S3, Supplementary Material online. (*A*) Results of the "full" analysis. Each target population was allowed to copy from all donor populations, excluding self-copying from its own group. (*B*) Results of the "regional" analysis. Populations were divided into five geographical regions: the Philippines, western Indonesia, eastern Indonesia, Melanesia, and Polynesia (fig. 1*A* and supplementary table S1, Supplementary Material online). Each target group was allowed to copy from all donor populations, excluding populations with the same regional label (supplementary table S4, Supplementary Material online).

Except for shared South Asian influence in some islands, the admixture patterns inferred by GT varied widely across western Indonesia (fig. 4). In contrast, all eastern Indonesian groups show a similar signal of genetic contact between Papuan-like and western Indonesian/Philippines-like sources, as reflected by the "regional" analysis (fig. 4*B* and supplementary table S3, Supplementary Material online). Estimated dates of admixture are similar across the eastern part of the archipelago, with point estimates ranging from around 185 BCE to 360 CE (95% CIs for different islands span the period from *ca* 510 BCE to 475 CE). All events infer simple one-date admixture between two source populations. The composition of mixing sources is also strikingly similar between different groups: one source always includes western Indonesian and Philippine non-Kankanaey surrogates, whereas the other is dominated by PNG, with no Taiwanese surrogate present in the sources of admixture. Previous analysis suggests that some Philippine non-Negrito groups are close proxies for the original AN ancestry (Mörseburg et al. 2016). Therefore, the detected events likely reflect interactions between local Papuan groups and AN speakers, who arrived in eastern Indonesia via the Philippines. Notably, the inferred dates of genetic contact are late relative to archeological signals associated with the AN expansion in this region.

Unlike other studies (Busby et al. 2015), we decided to describe admixture by clustering individual raw FS populations into larger topologically robust groups with high bootstrap support (supplementary fig. S1, Supplementary Material online), which closely match individual island groupings in our data set. To detect admixture shared between multiple raw populations, GT requires greater differences between groups than within them. These population groupings should therefore allow us to detect more ancient events, and be less subject to noise than using the raw FS populations, many of which are small, comprising only a few samples. To test whether admixture inferences are affected by the way our groups are formed, we also performed a second "regional" GT inference on groups defined by the tips of the FS tree. Unsurprisingly, given the homogeneity observed within studied islands, the results of this approach (supplementary fig. S3, Supplementary Material online) follow the main "regional" run closely (fig. 4*B*). For example, all 32 eastern Indonesian clusters display one-date admixture events between Papuan- and western Indonesian/Philippines-like sources around 60 CE $\pm$ 260 years (mean $\pm$ SD) (supplementary fig. S3 and table S5, Supplementary Material online). Furthermore, we also applied ancestry-based (fig. 3*B* and supplementary table S4, Supplementary Material online) rather than geography-based clustering for western Indonesians, as in the main "regional" run, and again observed similar results.

To gain additional insights into mainland Asian and AN-related population history in Indonesia, we performed an analysis of FS chunks received from Chinese, Vietnamese, and Philippine Kankanaey and non-Kankanaey donors, where—as noted above—non-Negrito populations from the Philippines appear to act as an AN proxy in a previous study (Mörseburg et al. 2016), as well as in our study (supplementary fig. S4, Supplementary Material online). Among Western Indonesian populations, Borneo and Sulawesi stand out as having much longer average chunk lengths and a larger proportion of their genome received from the Philippines, suggesting additional genetic contact between these regions. In contrast, individuals from Java and Bali, while having the highest proportion of their genome and one of the longest chunks accredited to mainland Asian donors, receive a smaller proportion and the shortest chunks from the Philippines. The former indicates that Java and Bali experienced additional contact with people from MSEA compared with other western Indonesians, and the latter may suggest more ancient admixture between these two populations and incoming AN-speaking groups than GT can detect (Pagani et al. 2016). Nias and Mentawai individuals copy only a small proportion of their genomes from mainland or AN donors, but have an unusually high average chunk length—an effect produced by low effective population size.

MALDER

To test whether the mainland Asian component was introduced into Indonesia as part of the AN expansion, or as part of a separate migration event, we turned to MALDER. This approach assesses the exponential decay of admixture-induced linkage disequilibrium (LD) in a target group, and requires a priori definition of two mixing sources. We performed this test for every target population in Indonesia by representing it as a mixture between PNG (a proxy for the ancestral Papuan population) and either mainland Asian (CHB, CHS, CDX, and KHV) or proxy AN sources (Philippine Kankanaey and non-Kankanaey samples). If introgression from mainland Asia predates the AN migration, we would expect admixture dates inferred using the PNG-mainland Asian pair to be older than that of the PNG-AN pair.

Figure 5 summarizes the statistically significant results obtained from all tests performed on the Indonesian target populations, as specified using self-reported population labels (see also supplementary table S6*A*, Supplementary Material online). Five eastern Indonesian groups and Sulawesi show results similar to, and corroborate, the GT findings. Importantly, inferred dates for the admixture between PNG-mainland Asia and PNG-AN references in each of the target populations are close with overlapping confidence intervals. Analysis of FS-based groups (as in fig. 4) produces similar results (supplementary table S6*B*, Supplementary Material online). One parsimonious explanation for this pattern is that both mainland Asian and AN ancestries were introduced into eastern Indonesia during a single migration event, and that these newcomers already had a mainland Asian component in their genomes at that time. The composition of western Indonesian/Philippines-like sources of AN-associated admixture inferred by the GT analysis for two Flores populations is also consistent with the MALDER results and indicates that Neolithic newcomers already had MSEA ancestry by around 200 BCE (fig. 4 and supplementary table S3, Supplementary Material online). Further, our mainland Asian and AN references have significantly different

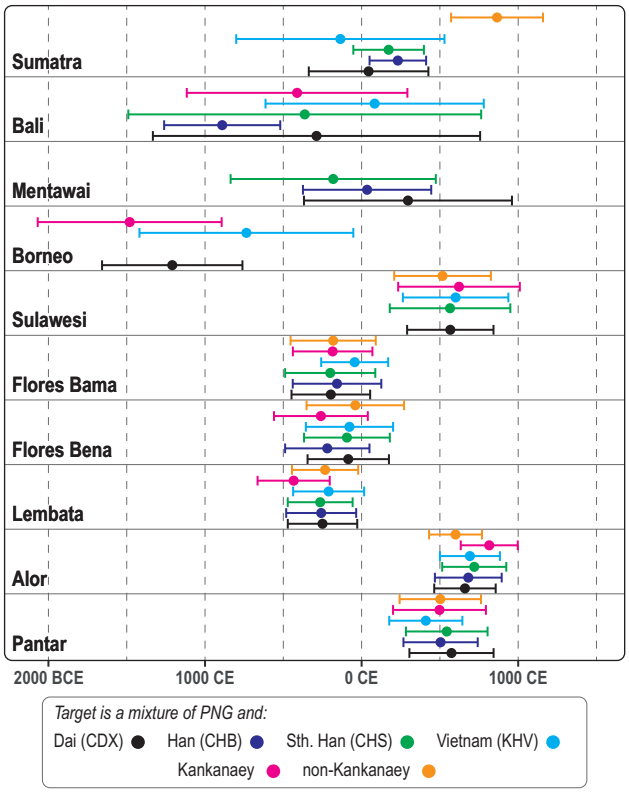

Fig. 5. Alternative genomic dating of admixture events with MALDER. Each target Indonesian population (vertical axis) was represented as a mixture of Papuan (PNG) sources, and either mainland Asian (CHB, CHS, CDX, and KHV) or AN (Philippine Kankanaey and non-Kankanaey) sources. Point estimates and the standard error of dates for statistically significant admixture events are shown on the horizontal axis.

patterns of LD (supplementary figs. S1 and S2, Supplementary Material online), thus ruling out the possibility that the observed close correspondence between the results of the PNG-mainland Asian and PNG-AN references in the MALDER analysis is due to similarity in their patterns of LD or ascertainment bias of the genotyping data (Cox et al. 2016 and evidence below). In contrast, however, the western islands do not exhibit any clear pattern of admixture in the MALDER analysis.

Outgroup f3

To further determine whether the mainland Asian component was introduced as part of the AN expansion, the outgroup *f3* test, a measure of common branch length between a pair of target populations and an outgroup (here, Yoruba from Africa, YRI), was performed (supplementary table S7, Supplementary Material online). Each Indonesian population was paired with the same set of mainland Asian and AN references as for the MALDER analysis, and in addition, with four ancient Lapita genomes (Skoglund et al. 2016). We observe almost perfect correlation of *f3* values estimated with mainland Asian and AN references (supplementary fig. S5, Supplementary Material online). If mainland Asian and AN components were introduced into Indonesia at substantially different times, it might be expected that the shared branch leading to the African outgroup in ((Indonesian target, mainland reference), YRI) and the ((Indonesian target, AN reference), YRI) tests would be significantly different. This is not observed, again suggesting that both mainland Asian and AN components were introduced across Indonesia during a single event, in line with the MALDER results for eastern Indonesia. This, however, does not rule out the possibility that there was some contact with mainland Asia before the AN expansion, the traces of which were either erased by incoming Neolithic populations or are too weak to be detected by the suite of methods applied here.

In addition, examination of outgroup *f3* test results reveals that two groups, Java and Bali, are pulled toward the mainland reference, suggesting additional genetic contact with MSEA compared with other Indonesian populations.

Measures of Genetic Drift and Inbreeding

To assess the effects of genetic drift and inbreeding in Indonesian populations, we estimated the excess of runs of homozygosity (ROH), mean pairwise nucleotide diversity (π), and gene diversity ($\hat{H}$) compared with larger continental groups (supplementary figs. S6 and S7 and table S8, Supplementary Material online). Although genotype-based measures (e.g., π) can potentially be sensitive to the ascertainment bias of SNP arrays, haplotype-based gene diversity is calculated from combinations of linked markers, and is consequently more robust. The observed high correlation between nucleotide and gene diversity in ISEA (supplementary fig. S7, Supplementary Material online) therefore suggests that ascertainment bias in the genotyping data is unlikely to significantly underestimate these measures of genetic diversity, in line with previous results (Cox et al. 2016).

Setting aside Nias, Mentawai, Sulawesi, and Flores Bena, the total amount of the genome located in homozygous stretches is about five times higher in Indonesians (33 ± 14 Mb, mean $\pm$ SD) compared with East Asians (6 ± 2 Mb), and is within the range of South Asian populations (41 ± 28 Mb) (supplementary fig. S6A, Supplementary Material online, ROHs defined as a minimum of 50 homozygous SNPs). Apart from Nias and Mentawai, nucleotide and gene diversity estimates in Indonesia—including Sulawesi and Flores Bena—overlap with East Asia (supplementary fig. S7 and table S8, Supplementary Material online). Two western Indonesian islands, Nias and Mentawai, show strongly reduced levels of genetic diversity, as well as an excess of ROH. Mentawai individuals have the lowest diversity among all the sampled Indonesian populations, coupled with the largest proportion of their genome located in ROH (120 Mb in total). Together with an excess of self-copying in the FS analysis (supplementary figs. S2 and S4, Supplementary Material online), this indicates that the population history of Nias and Mentawai was strongly influenced by genetic drift.

Correlation between Linguistic and Genetic Diversity

It is currently thought that languages, culture and genes spread in lockstep across Island Southeast Asia during the Neolithic (Gray et al. 2009). To test whether these correlations

can still be observed, we performed a Mantel test between genetic distances, as estimated by F_{ST} (supplementary table S2, Supplementary Material online), and phonetic distances, as estimated by ALINE (Downey et al. 2008) (supplementary table S9, Supplementary Material online). Genetic distances are strongly correlated with language distances within Indonesia ($r = 0.56$, $P = 0.002$), and a significant correlation remains even when controlling for the geographical covariate ($r = 0.38$, $P = 0.02$) (supplementary table S9, Supplementary Material online).

Discussion

Indonesia, linguistically and culturally one of the most diverse regions on earth, was affected by multiple massive human movements. Here, we report a new high-density autosomal genotyping data set from multiple Indonesian islands that significantly advances previous data. We apply the latest admixture inference approaches to obtain insight into the demographic processes that gave rise to modern population diversity, with a primary focus on the Holocene.

A key overarching feature of this region is the distribution of two major genetic ancestries, Papuan and Asian, which vary considerably between Indonesian islands, with a particular distinction between west and east (figs. 1–3). A cline in the proportion of the two ancestries is observed against longitude—the Papuan component, which can be traced back to the first anatomically modern humans in Sunda and Sahul, reaches its highest frequency in eastern Indonesia, but is essentially absent west of Wallace's line (fig. 2). The second most frequent genetic component in Indonesia appears to have originated from near Taiwan around 3500–3000 BCE (Bellwood 2014), and is thought to be associated with the Neolithic expansion of AN speakers. This dispersal likely accounts for the observed correlation between genetic and language diversity (supplementary table S9, Supplementary Material online). As with previous analyses of small traditional island communities in Indonesia (Cox et al. 2016), most studied populations do not show strong signs of inbreeding or a substantial decrease in genetic diversity, pointing to effective population sizes similar to those of much larger continental groups (supplementary figs. S6 and S7 and table S8, Supplementary Material online).

Most eastern Indonesian populations show traces of admixture that appear to reflect an expansion of AN speakers (fig. 4B and supplementary fig. S3, Supplementary Material online). There is a striking similarity between inferred events—each admixed population includes both a Philippine non-Kankanaey and western Indonesian-like source likely representing Holocene movements of Asian farming groups, as well as a Papuan-like source representing local indigenous ancestry. One reason for the lack of clear Taiwanese sources may be because the aboriginal populations of Taiwan were heavily affected by post-AN movements from mainland East Asia, most recently sinicization by Han Chinese, and thus no longer depict the ancestral AN gene pool (Mörseburg et al. 2016). However, this notable pattern could equally be explained by the dominance of language and culture transfers during early phases of the Neolithic expansion from Taiwan into the Philippines, followed by people with predominantly Philippine ancestry driving later demic diffusion into the Indonesian archipelago. Interestingly, Mörseburg et al. (2016), by using a different sample set and genotype-based analytical toolkit, indicated that the Kankanaey ethnic group from the Philippines is likely the closest living proxy of the source population that gave rise to the AN expansion. We did not detect this population among sources of admixture in eastern Indonesia, and therefore suggest that the place of individual Philippine groups in the AN expansion needs to be further addressed by better sampling in the Philippine archipelago.

Sumba and Flores, the two westernmost islands to the east of Wallace's line, display a high proportion of Java and Bali surrogates in their AN admixing source. This suggests that the AN movement into eastern Indonesia, especially for Sumba and Flores, had earlier experienced some degree of genetic contact with western Indonesian groups. In contrast, the sources of AN admixture in Lembata, Alor, Pantar, and Timor are dominated by Sulawesi (fig. 4B and supplementary fig. S3 and tables S3 and S5, Supplementary Material online). This generally agrees with expectations from the geography of the region, whereby AN groups exiting the southern Philippines were likely funneled into at least two streams, including a western path through Borneo and a central path through Sulawesi (Blust 2014).

Point estimates of genetic admixture times in eastern Indonesia lie within a narrow timeframe ranging between *ca* 185 BCE to 360 CE or 75 to 56 generations ago (95% CI 510 BCE—475 CE or 87–52 generations) (fig. 4B supplementary table S3, Supplementary Material online). These inferred dates are younger than some previous estimates (120–200 generations ago) (Xu et al. 2012; Sanderson et al. 2015; Sedghifar et al. 2015). A major analysis of admixture in Indonesia estimated the date of AN contact in the eastern part of archipelago to be around 500–600 CE (*ca* 50 generations, CI estimates between 58 and 42 generations ago) (Lipson et al. 2014), surprisingly young given the archaeological evidence. However, the study pooled a very small sample of genetically heterogeneous eastern Indonesian islands including, for example, Flores and Alor. As we show here (figs. 2, 4, and 5 and supplementary fig. S3 and tables S3, S5, and S6, Supplementary Material online), whereas the wave of AN speakers left a common genetic trace across the whole of eastern Indonesia, the details and dates of this contact vary considerably not only between islands (e.g., Flores and Alor), but also within individual islands (e.g., Flores Rampasasa vs. Flores Bama). The genetic dates, which were obtained here by denser geographical sampling of eight eastern islands, a much larger number of individuals (28 per island on average) and a greater number of SNPs, are up to 30 generations older, predating the Common Era in many cases. It therefore took migrants at least half a millennium to proceed from islands around Wallace's line to the easternmost sampled part of eastern Indonesia.

Nevertheless, observed dates for AN contact in eastern Indonesia are still approximately a millennium younger

than the earliest Neolithic archaeological evidence in the region, and two explanations seem most likely here. First, the AN migration may have involved several waves of people leaving Taiwan, spanning multiple generations, which would bias date estimates later than the first arrival of the Neolithic archeological assemblage (Sedghifar et al. 2015). Second, there may have been a substantial time gap between the spread of culture and technological traditions, and the beginning of extensive genetic contact between incoming farming groups and native inhabitants in Indonesia (Lansing et al. 2011). The lack of considerable admixture with Papuan groups was recently noted in ancient Lapita individuals from Remote Oceania, whose genomes are mostly Asian and carry little to no Papuan ancestry, suggesting limited contact as they moved through Melanesia to previously uninhabited islands in the Pacific (Skoglund et al. 2016). A lag in admixture between local and incoming Neolithic groups has also been observed in Europe, where hunter-gatherer and farming populations initially coexisted for nearly a thousand years without substantial genetic interaction (Malmström et al. 2015).

Unlike eastern parts of the Indonesian archipelago, the western islands show more complex traces of the AN expansion—this ancestry is seen in the ADMIXTURE plot, but is not inferred by GT and MALDER (figs. 2, 4, and 5). A combination of multiple factors is potentially responsible for this observation. From archaeological evidence, the timeframe of AN admixture in western Indonesia is older than in eastern islands, which severely diminishes the power of LD-based admixture inference (Hellenthal et al. 2014). Only Sulawesi, lying between the prehistoric continents of Sahul and Sunda, shows traces of AN gene flow (figs. 4A and 5 and supplementary fig. S3, Supplementary Material online). However, additional insights are given by FS chunk analysis—Borneo and Sulawesi may have experienced the most recent AN admixture, with Java and Bali experiencing the most ancient AN contact among all western Indonesian populations (supplementary fig. S4, Supplementary Material online). In addition, historically well-documented gene flow from South Asia from the first century CE (Ooi 2004), the timing and extent of which have been clarified here, further decrease the statistical power to detect older AN-associated admixture in western Indonesia (figs. 2 and 4) (Hellenthal et al. 2014).

Although the genetic effects of the AN expansion are now well established (figs. 2 and 4), the presence of a mainland Southeast Asian genetic substrate in Indonesia, as confirmed here (fig. 2), still needs to be addressed (Karafet et al. 2010; Jinam et al. 2012; Lipson et al. 2014). The timeframe for this contact is open to debate. Analyses of haploid loci suggest a pre-Austronesian origin during the late Pleistocene–early Holocene (Karafet et al. 2010; Jinam et al. 2012; Vallee et al. 2016). However, autosomal data indicates that some expanding AN speakers interacted with groups from MSEA by at least 350 CE (or 55 generations ago), and introduced mainland ancestry into western Indonesia (Lipson et al. 2014). Although we did not detect any major admixture events with MSEA in our GT analysis (fig. 4), the panel of outgroup *f3* tests (supplementary fig. S5 and table S7, Supplementary Material online) suggests that the mainland component was likely introduced into the Indonesian archipelago in parallel with the AN migration. This is more clearly the case in eastern Indonesia, which shows similar dates of admixture with both mainland Asian and AN references (fig. 5). Western Indonesians display a more complex pattern of contacts with MSEA. Some populations, particularly Nias and Mentawai, seem to have remained remained isolated and are affected by strong genetic drift (fig. 2 and supplementary figs. S4, S6, and S7 and table S8, Supplementary Material online) (van Oven et al. 2011; Kennerknecht et al. 2012). In contrast, Java and Bali have experienced additional gene flow from MSEA compared with other western Indonesian islands (fig. 2 and supplementary figs. S4 and S5, Supplementary Material online).

In summary, the Indonesian archipelago, and ISEA in general, is a region of genetic contrasts, shaped by quite different prehistoric and historic human movements that cumulatively drove the immense cultural and linguistic diversity observed today. Growing evidence argues for a complex transition from hunter-gatherer to farming populations, with multiple centers of neolithization, albeit still with a strong driver of human population movements (Spriggs 2012). Our results provide new insights into the genetic diversity of the region. We highlight key differences in the population history of western versus eastern Indonesia, and establish the details of both the AN expansion and contact with mainland Asia, thus proposing directions for future work in the era of complete genome sequences.

Materials and Methods

Ethics

Biological samples were collected by JSL, HS, and a team from the Eijkman Institute for Molecular Biology, with the assistance of Indonesian Public Health clinic staff. All collections followed protocols for the protection of human subjects established by institutional review boards at the Eijkman Institute, Nanyang Technological University, and the University of Arizona. Permission to conduct research in Indonesia was granted by the State Ministry of Research and Technology. Genotyping and analyses of newly reported non-Indonesian samples were approved by the institutional review board at the University of Arizona.

Data Availability

Genotype data for all new samples are available from the NCBI GEO repository (project accession number: GSE80534).

Sampling and Genetic Screening

Genetic markers were screened in 498 healthy and unrelated individuals from Vietnam, Taiwan, Philippines, Indonesia (Sumatra, Java, Bali, Nias, Mentawai, Sulawesi, Sumba, Flores Rampasasa, Flores Bena, Flores Bama, Lembata, Alor, Pantar, and Timor), Melanesia (Papuans from PNG, Baining from New Britain, Nasioi from Bougainville, and Vanuatu), and Polynesia (Tonga, Samoa, and Tahiti) representing 25 populations in total. Detailed information on individual populations used in the current study is given in supplementary

table S1, Supplementary Material online. A set of 567,096 SNPs was screened using the Affymetrix Axiom Genome-Wide Human Array. All samples yielded high genotyping success rates (<5% missing genotypes), and 538,139 autosomal SNPs with <5% missing data were kept for further analyses. Inference of cryptic relationships between samples was performed using KING v. 1.4 (Manichaikul et al. 2010) and no first degree relatives were detected (kinship coefficient > 0.354, following the software guidelines).

Comparative Data Sets

Newly genotyped data were merged with autosomal data from the following complete genome sequencing data sets: African (YRI: Yoruba), European (CEU), South Asian (BEB: Bengali, GIH: Gujarati Indian, ITU: Indian Telugu, PJL: Punjabi, STU: Sri Lankan Tamil), mainland East Asian (JPT: Japanese, CHB: Chinese Han, CHS: Chinese Southern Han) and Southeast Asian (CDX: Chinese Dai, KHV: Vietnamese) populations from Phase 3 of the 1000 Genomes Project data (May 2, 2013 release) (1000 Genomes Project Consortium 2012), and populations from Vietnam, the Philippines, Borneo and Sulawesi from the Estonian Biocentre Human Genome Diversity Panel (EGDP) (Pagani et al. 2016) (supplementary table S1, Supplementary Material online). Twenty-five random samples from each population of the 1000 Genomes Project were used to save computation time and balance sample size with Indonesian data. Philippine samples were broadly classified into three groups: Kankanaey, Negritos, and a general non-Kankanaey group, where the latter included samples that did not belong to either of the two former groups. Philippine Kankanaey and non-Kankanaey samples are together referred to as non-Negritos throughout the text. We purposely avoided merging our data with public genotyping data sets to maximize the number of common SNPs available for further analysis. The comparative data set thus includes 853 samples with 508,572 autosomal SNPs, and <5% missing data. For some of the analyses, including PCA, ADMIXTURE, and F_{ST}, highly linked SNPs with $R^2 > 0.2$ were removed (-indep-pairwise 50 5 0.2 in Plink v. 1.90 beta; Chang et al. 2015); 167,855 SNPs passed this criterion.

In addition, four ancient human genomes dated to the Lapita and immediately post-Lapita period, 1100–300 BCE, from Tonga (CP30) and Vanuatu (I1368, I1369, and I1370) were added to the comparative data set (Skoglund et al. 2016). The total number of unlinked SNPs in common between the comparative data set and all four aDNA samples was limited to 1,977 polymorphisms, whereas the number of SNPs shared with individual ancient genomes, as analyzed separately, varied between 22k and 38k. As this number of SNPs is too low for the haplotype-based analyses performed here, only outgroup *f3* tests were undertaken using individual aDNA data points.

Population Genetic Analysis

F_{ST}, a measure of population differentiation, was calculated using GENEPOP v. 4.4 (Rousset 2008). PCA was performed with the smartpca function of EIGENSOFT v. 3.0 (Patterson et al. 2006) with no outlier removal step. Mean pairwise nucleotide diversity π was calculated using the method of Nei and Li (1979) with rare alleles removed (minor allele frequency <5%). Gene diversity H (Nei 1987) was estimated using the same procedure as in Cox et al. (2016). Runs of Homozygosity (ROH) were calculated within individuals in Plink v. 1.90 beta. Following Howrigan et al. (2011), the minimum number of homozygous SNPs to define a run was varied from 20 to 95 in increments of 15 SNPs, and two different linkage disequilibrium pruning parameters were used—"moderate" and "heavy" ($R^2 > 0.2$ and 0.09, respectively).

Maximum likelihood estimates of the ancestry of individuals were performed with ADMIXTURE v. 1.30 (Alexander et al. 2009). Fifty randomly seeded runs were performed for each number of ancestral populations ($K = 2–12$), 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, and individual ancestry proportions were averaged across runs belonging to the same mode. The most frequent modal solution is reported.

Outgroup *f3* statistics, a measure of the shared branch length between two population samples, was calculated for each Indonesian population with ADMIXTOOOLS v. 4.1 (Patterson et al. 2012) using the Yoruba (YRI) population as the outgroup. The test took the form ((Indonesian target, reference), YRI), and used reference populations from mainland (South-)East Asia (CHB, CHS, CDX, and KHV), Philippine Kankanaey and non-Kankanaey samples, and ancient Lapita genomes.

Inference of Genetic Clustering and Admixture Events

Two different methods were used to infer admixture. First, we employed the fineSTRUCTURE (FS), CHROMOPAINTER, and GLOBETROTTER (GT) framework. This approach uses the haplotypic structure of the data to reconstruct the original admixing source populations and date(s) of admixture given a set of surrogate populations. Unlike other approaches (e.g., ALDER/MALDER), this framework does not require the specification of source populations of admixture, but instead reconstructs it from the given set of sampled surrogates. This multistep algorithm can be described as follows:

(1) Genotypes were first phased with SHAPEIT v. 2 (Delaneau et al. 2014) using the HapMap phase II b37 recombination map (International HapMap Consortium 2007).
(2) Population assignments to genetic groups were performed using fineSTRUCTURE v. 2 (Lawson et al. 2012). A quick test iteration of the algorithm was performed using the complete comparative data set, without ancient genomes. For the main run, the Yoruba, which donated only a negligible number of chunks to the newly genotyped MSEA, ISEA, Melanesian and Polynesian recipients (<0.1% per recipient genome on average) was excluded to speed up computation. The inferred genetic clustering of the remaining

samples was essentially invariant between the test run and the main runs. Initial N_e and θ parameters were estimated using 10% of the samples and ten Expectation-Maximization steps of the algorithm. Next, each individual recipient chromosome was described as a mixture of genetic chunks from the set of all other individuals, or donors. This generated a matrix of copying vectors, which was further used to cluster the individuals with the Bayesian algorithm. Two parallel runs were performed and convergence between them was assessed using Gelman–Rubin statistics, as implemented in the software. Six and 24 million MCMC iterations were performed for the test and main runs, respectively, with the first 3 million iterations discarded as burn-in. The main runs were performed until the convergence diagnostic statistic between two MCMC chains reached the 1.3 threshold, as defined in the manual. Finally, the tree building step was performed using default settings, and the run with the highest observed posterior likelihood was used to cluster the samples into genetic groups.

(3) The population dendrogram produced in the previous step was manually inspected and samples were assigned to individual groups (supplementary fig. S1 and table S1, Supplementary Material online), including 15 from Indonesia (7 western and 8 eastern groups). Group-based admixture inference (also referred to as the main GT analysis, fig. 4) was performed in target populations from ISEA, Melanesia and Polynesia. A group of nine eastern Indonesian individuals with mixed geographical origin was excluded. Additional selection of samples from two mainland (South-)East Asian (CHS combined with CHB, and CDX combined with KHV), South Asian (STU), and European (CEU) clusters were used as surrogates (supplementary table S1, Supplementary Material online).

(4) We used two different approaches for the main GT analysis (Hellenthal et al. 2014). First, a "full" analysis was performed (fig. 4*A*), where each recipient genome could copy chunks from the genomes of all other donor populations. However, this approach does not account for the possibility that geographically close samples assigned to two different genetic clusters can share a similar admixture history. These would preferentially copy from each other, competing with more distant donors and potentially masking the true admixture signal. To minimize this effect of shared population history, we performed a "regional" analysis (fig. 4*B*). Populations were clustered into five geographical regions: the Philippines, western Indonesia, eastern Indonesia, Melanesia, and Polynesia. Individual recipient genomes were not allowed to copy from donors with the same regional label. For details of regional clusters, refer to supplementary table S4, Supplementary Material online.

(5) In addition to the main GT analysis, we also assessed the effect of sample clustering on our admixture inference by performing a GT analysis on the raw FS populations (i.e., terminal tips of the FS dendrogram, supplementary fig. S3, Supplementary Material online), implemented using the "regional" approach. To save computation costs, analysis was performed only on Indonesian tips (17 western and 32 eastern) as targets of admixture, but including 43 non-Indonesian clades as surrogates (supplementary fig. S1 and table S1, Supplementary Material online; 1.6 CPU years to complete). As with the approach described above, eastern Indonesian tips, including the group with mixed sample origin, were allowed to copy from outside eastern Indonesia only (supplementary table S4, Supplementary Material online). However, for western Indonesia, unlike the geography-driven clustering in the main "regional" analysis (fig. 4*B*), individual tips were now assigned into four monophyletic regions using the FS results (fig. 3*B* and supplementary fig. S1, Supplementary Material online): a) Sumatra, Mentawai and Nias, b) Java and Bali, c) Sulawesi (with Philippine Negritos), and d) Borneo (with Taiwan and Philippine non-Negrito samples) (supplementary table S4, Supplementary Material online). Tips comprising a single individual were grouped with the closest neighbor.

Following the GT guidelines, a copying vector for each donor group and a set of painted chromosomes for each recipient group were generated with CHROMOPAINTER v. 2 (Hellenthal et al. 2014) using the "full" and "regional" algorithms described above. Cluster-specific values of N_e and θ were estimated using a "leave-one-out" approach, where each individual was allowed to copy from every other individual with the same cluster label k, but only n_j—1 random individuals from other donor clusters $k \neq j$, where n is the sample size of cluster j. For the "regional" analysis, N_e and θ estimates were averaged over all clusters with the same regional label to avoid any potential bias.

For the main GT analysis of grouped samples (fig. 4), two parallel runs were performed with and without a "NULL" individual (see Hellenthal et al. [2014] for details), and 100 bootstraps were used to assess the statistical significance of the admixture event and uncertainty of the inferred date(s). Additionally, for all multiple date events, 100 bootstraps were performed for the two-date admixture event. If the lower bound of the more recent admixture bootstrap was $\leq$3 generations and the fit quality of a single admixture event was $>$0.975, the final result was classified as a one-date event. Otherwise, it was classified as a multiway event. Convergence between two runs was checked manually, and only results showing compatible events, sources of admixture and overlapping admixture time confidence intervals were considered further. For the GT analysis of raw FS populations (supplementary fig. S3, Supplementary Material online), only runs without a "NULL" individual and bootstrapping were performed.

Second, an alternative approach, implemented in MALDER (Pickrell 2015)—a version of ALDER (Loh et al. 2013) that has been modified to allow multiple admixture events—was employed to detect and date past admixture. This requires the definition of two sources of admixture, as

well as the admixed target population. We used this test to assess the timeframe of mainland Asian versus AN admixture across our Indonesian populations. The test was designed to describe each target population as a mixture between a combined sample of Papua New Guinea highlanders and lowlanders (referred to as PNG) and each of the following Asian populations: Chinese Dai (CDX), Chinese Han (CHB), Chinese Southern Han (CHS), Vietnamese (KHV), and Kankanaey and non-Kankanaey from the Philippines. Two different sample sets were examined: individuals were clustered using either the original self-reported population labels, or the same set of 14 natural Indonesian groups as used for the main GT analysis (supplementary fig. S1 and tables S1 and S6, Supplementary Material online). The same recombination map was used as for GT, and the minimum genetic distance at which to start LD decay curve fitting was determined automatically. For both MALDER and GT, admixture dates were converted to years using the formula $(1950 - (x + 1) * 28)$, where x is the number of generations since the admixture event and the generation interval is 28 years (Fenner 2005).

Linguistic Variation

Language word lists of 136 Swadesh terms were collected by JSL, analyzed by a trained linguist, and combined with other sources, including the Austronesian Basic Vocabulary Database (Greenhill et al. 2008). Linguistic and genetic data sets were manually cross-referenced based on geographic location, and only samples from AN-speaking populations were used. Genetic distances, in the form of F_{ST}, were calculated as described above. Geographic pairwise distances were calculated using the R package "geosphere" v. 1.5-5 (Hijmans et al. 2016). Pairwise linguistic distances were calculated as the mean ALINE distance (Downey et al. 2008) for all lexical items that were shared between each pair of languages using the R package "alineR" v. 1.3.3 (Downey and Guowei 2016). Complete and partial Mantel tests were conducted with the mantel function in the R package "vegan" v. 2.4-2 (Oksanen et al. 2017).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank colleagues for their helpful contributions: Alex Mentzer, John Clegg, Adrian Hill, and David Weatherall (University of Oxford, UK); Alena Kushniarevich (Estonian Biocentre, Estonia); Peter Norquest (University of Arizona); Satoshi Horai (National Institute of Genetics, Japan); Hie Lim Kim (Nanyang Technological University, Singapore); Nuualofa Tuuau (Ministry of Health, Western Samoa); Jean Roux (Institut Territorial de Recherches Medicales Louis Malardé, Tahiti); Sunia Foliaki (Ministry of Health, Tonga); Don Bowden (British Medical Service, Vanuatu), and George Bule (Ministry of Health, Vanuatu). This research was supported by the University of Arizona via funding to M.F.H., 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.

References

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

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

Barker G, Barton H, Bird M, Daly P, Datan I, Dykes A, Farr L, Gilbertson D, Harrisson B, Hunt C, et al. 2007. The 'human revolution' in lowland tropical Southeast Asia: The antiquity and behavior of anatomically modern humans at Niah cave (Sarawak, Borneo). *J Hum Evol.* 52:243–261.

Bellwood P. 2014. Southeast Asian islands: archaeology. In: Bellwood P, editor. The global prehistory of human migration. Hoboken: John Wiley and Sons Inc.

Blench RM. 2012. Almost everything you believed about the Austronesians isn't true. In: Tjoa-Bonatz ML, Reinecke A, Bonatz D, editors. Selected papers from the 13th international conference of the European Association of Southeast Asian archaeologists. Singapore: NUS Press.

Blench RM. 2010. Was there an Austroasiatic presence in Island Southeast Asia prior to the Austronesian expansion? *Bull Indo Pac Pre Hi* 30:133–144.

Blust R. 2014. Southeast Asian islands and Oceania: Austronesian linguistic history. In: Bellwood P, editor. The global prehistory of human migration. Hoboken: John Wiley and Sons Inc.

Busby GB, Hellenthal G, Montinaro F, Tofanelli S, Bulayeva K, Rudan I, Zemunik T, Hayward C, Toncheva D, Karachanak-Yankova S, et al. 2015. The role of recent admixture in forming the contemporary west Eurasian genomic landscape. *Curr Biol.* 25:2518–2526.

Chang C, Chow C, Tellier L et al. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* 4:7.

Cox MP, Hudjashov G, Sim A, Savina O, Karafet TM, Sudoyo H, Lansing JS. 2016. Small traditional human communities sustain genomic diversity over microgeographic scales despite linguistic isolation. *Mol Biol Evol.* 33:2273–2284.

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

Denham T, Donohue M. 2009. Pre-Austronesian dispersal of banana cultivars west from New Guinea: linguistic relics from eastern Indonesia. *Archaeol Ocean* 44:18–28.

Downey SS and Guowei S. 2016. alineR: Alignment of phonetic sequences using the 'ALINE' algorithm [Internet]. Available from: https://cran.r-project.org/package=alineR.

Downey SS, Hallmark B, Cox MP, Norquest P, Lansing JS. 2008. Computational feature-sensitive reconstruction of language relationships: Developing the ALINE distance for comparative historical linguistic reconstruction. *J Quant Linguist.* 15:340–369.

Fenner JN. 2005. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am J Phys Anthropol.* 128:415–423.

Gomes SM, Bodner M, Souto L, Zimmermann B, Huber G, Strobl C, Rock AW, Achilli A, Olivieri A, Torroni A, et al. 2015. Human settlement history between Sunda and Sahul: a focus on East Timor (Timor-Leste) and the pleistocenic mtDNA diversity. *BMC Genomics* 16:70.

Gray RD, Drummond AJ, Greenhill SJ. 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323:479–483.

Greenhill SJ, Blust R, Gray RD. 2008. The Austronesian basic vocabulary database: From bioinformatics to lexomics. *Evol Bioinform Online* 4:271–283.

Hellenthal G, Busby GB, Band G, Wilson JF, Capelli C, Falush D, Myers S. 2014. A genetic atlas of human admixture history. *Science* 343:747–751.

Hijmans RJ, et al. 2016. Geosphere: Spherical trigonometry [Internet]. Available from: https://cran.r-project.org/package=geosphere.

Horton R. 2016. Indonesia—unravelling the mystery of a nation. *The Lancet* 387:830.
Howrigan DP, Simonson MA, Keller MC. 2011. Detecting autozygosity through runs of homozygosity: a comparison of three autozygosity detection algorithms. *BMC Genomics* 12:460.
HUGO Pan-Asian SNP Consortium. 2009. Mapping human genetic diversity in Asia. *Science* 326:1541–1545.
International HapMap Consortium. 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.
Jinam TA, Hong LC, Phipps ME, Stoneking M, Ameen M, Edo J, HUGO Pan-Asian SNP, Consortium, Saitou N. 2012. Evolutionary history of continental Southeast Asians: "Early train" hypothesis based on genetic analysis of mitochondrial and autosomal DNA data. *Mol Biol Evol*. 29:3513–3527.
Karafet TM, Hallmark B, Cox MP, Sudoyo H, Downey SS, Lansing JS, Hammer MF. 2010. Major east-west division underlies Y chromosome stratification across Indonesia. *Mol Biol Evol*. 27:1833–1844.
Kennerknecht I, Hämmerle JM, Blench RM. 2012. The peopling of Nias, from the perspective of oral literature and molecular genetic data. In: Tjoa-Bonatz ML, Reinecke A, Bonatz D, editors. Selected papers from the 13th international conference of the European Association of Southeast Asian archaeologists. Singapore: NUS Press.
Ko AM, Chen CY, Fu Q, Delfin F, Li M, Chiu HL, Stoneking M, Ko YC. 2014. Early Austronesians: into and out of Taiwan. *Am J Hum Genet*. 94:426–436.
Kusuma P, Brucato N, Cox MP, Pierron D, Razafindrazaka H, Adelaar A, Sudoyo H, Letellier T, Ricaut FX. 2016. Contrasting linguistic and genetic origins of the Asian source populations of Malagasy. *Sci Rep*. 6:26066.
Kusuma P, Cox MP, Brucato N, Sudoyo H, Letellier T, Ricaut FX. 2016. Western Eurasian genetic influences in the Indonesian archipelago. *Quatern Int*. 416:243–248.
Lansing JS, Cox MP, de Vet TA, Downey SS, Hallmark B, Sudoyo H. 2011. An ongoing Austronesian expansion in Island Southeast Asia. *J Anthropol Archaeol*. 30:262–272.
Larson G, Cucchi T, Fujita M, Matisoo-Smith E, Robins J, Anderson A, Rolett B, Spriggs M, Dolman G, Kim TH, et al. 2007. Phylogeny and ancient DNA of *Sus* provides insights into Neolithic expansion in Island Southeast Asia and Oceania. *Proc Natl Acad Sci U S A*. 104:4834–4839.
Lawson DJ, Hellenthal G, Myers S, Falush D. 2012. Inference of population structure using dense haplotype data. *PLoS Genet*. 8:e1002453.
Lewis MP, Simons GF, Fennig CD. 2016. Ethnologue: languages of the world, 19th ed. Dallas, Texas: SIL International.
Lipson M, Loh PR, Patterson N, Moorjani P, Ko YC, Stoneking M, Berger B, Reich D. 2014. Reconstructing Austronesian population history in Island Southeast Asia. *Nat Commun*. 5:4689.
Loh PR, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, Berger B. 2013. Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* 193:1233–1254.
Malmström H, Linderholm A, Skoglund P, Storå J, Sjödin P, Gilbert MT, Holmlund G, Willerslev E, Jakobsson M, Lidén K, et al. 2015. Ancient mitochondrial DNA from the northern fringe of the Neolithic farming expansion in Europe sheds light on the dispersion process. *Philos Trans R Soc Lond B Biol Sci*. 370:20130373.
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.
Mijares AS, Detroit F, Piper P, Grun R, Bellwood P, Aubert M, Champion G, Cuevas N, De Leon A, Dizon E. 2010. New evidence for a 67,000-year-old human presence at Callao cave, Luzon, Philippines. *J Hum Evol*. 59:123–132.
Mörseburg A, Pagani L, Ricaut FX, Yngvadottir B, Harney E, Castillo C, Hoogervorst T, Antao T, Kusuma P, Brucato N, et al. 2016. Multi-layered population structure in Island Southeast Asians. *Eur J Hum Genet*. 24:1605–1611.
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.
Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, et al. 2017. Vegan: Community ecology package [Internet]. Available from: https://cran.r-project.org/package=vegan.
Oliver DL. 1974. Ancient Tahitian society. Honolulu: University Press of Hawaii.
Ooi KG. 2004. Southeast Asia: A historical encyclopedia, from Angkor Wat to East Timor. Santa Barbara, Calif.: ABC-CLIO.
Pagani L, Lawson DJ, Jagoda E, Mörseburg A, Eriksson A, Mitt M, Clemente F, Hudjashov G, DeGiorgio M, Saag L, et al. 2016. Genomic analyses inform on migration events during the peopling of Eurasia. *Nature* 538:238–242.
Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012. Ancient admixture in human history. *Genetics* 192:1065–1093.
Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet*. 2:e190.
Pickrell J. 2015. MALDER [Internet]. Available from: https://github.com/joepickrell/malder.
Pugach I, Stoneking M. 2015. Genome-wide insights into the genetic history of human populations. *Investig Genet*. 6:6.
Reich D, Patterson N, Kircher M, Delfin F, Nandineni MR, Pugach I, Ko AM, Ko YC, Jinam TA, Phipps ME, et al. 2011. Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *Am J Hum Genet*. 89:516–528.
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.
Sedghifar A, Brandvain Y, Ralph P, Coop G. 2015. The spatial mixing of genomes in secondary contact zones. *Genetics* 201:243–261.
Skoglund P, Posth C, Sirak K, Spriggs M, Valentin F, Bedford S, Clark GR, Reepmeyer C, Petchey F, Fernandes D, et al. 2016. Genomic insights into the peopling of the southwest Pacific. *Nature* 538:510–513.
Soares PA, Trejaut JA, Rito T, Cavadas B, Hill C, Eng KK, Mormina M, Brandao A, Fraser RM, Wang TY, et al. 2016. Resolving the ancestry of Austronesian-speaking populations. *Hum Genet*. 135:309–326.
Spriggs M. 2012. Is the Neolithic spread in Island Southeast Asia really as confusing as the archaeologists (and some linguists) make it seem? In: Tjoa-Bonatz ML, Reinecke A, Bonatz D, editors. Selected papers from the 13th international conference of the European Association of Southeast Asian archaeologists. Singapore: NUS Press.
Sutikna T, Tocheri MW, Morwood MJ, Saptomo EW, Jatmiko ARD, Wasisto S, Westaway KE, Aubert M, Li B, et al. 2016. Revised stratigraphy and chronology for *Homo floresiensis* at Liang Bua in Indonesia. *Nature* 532:366–369.
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.
Vallee F, Luciani A, Cox MP. 2016. Reconstructing demography and social behavior during the Neolithic expansion from genomic diversity across Island Southeast Asia. *Genetics* 204:1495–1506.
van Oven M, Hämmerle JM, van Schoor M, Kushnick G, Pennekamp P, Zega I, Lao O, Brown L, Kennerknecht I, Kayser M. 2011. Unexpected island effects at an extreme: Reduced Y chromosome and mitochondrial DNA diversity in Nias. *Mol Biol Evol*. 28:1349–1361.
Wollstein A, Lao O, Becker C, Brauer S, Trent RJ, Nurnberg P, Stoneking M, Kayser M. 2010. Demographic history of Oceania inferred from genome-wide data. *Curr Biol*. 20:1983–1992.
Xu S, Pugach I, Stoneking M, Kayser M, Jin L, 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.