Supplementary Information for: Ancient Genomics Reveals Four Prehistoric Migration Waves into Southeast Asia

Authors: Hugh McColl[1,*,^], Fernando Racimo[1,*], Lasse Vinner[1,*], Fabrice Demeter[1,2,*], J. Víctor Moreno Mayar[1], Uffe Gram Wilken[1], Andaine Seguin-Orlando[1], Constanza de la Fuente Castro[1], Sally Wasef[3], Ana Prohaska[4], Ashot Margarayan[1,5], Peter de Barros Damgaard[1], Rasmi Shoocongdej[6], Viengkeo Souksavatdy[7], Thongsa Sayavongkhamdy[7], Mohd Mokhtar Saidin[8], Supannee Kaewsutthi[9], Patcharee Lertrit[9], Huong Mai Nguyen[10], Hsiao-chun Hung[11], Thi Minh Tran[10], Huu Nghia Truong[10], Shaiful Shahidan[8], Ketut Wiradnyana[12], Anne-Marie Bacon[13], Philippe Duringer[14], Jean-Luc Ponche[14,15], Laura Shackelford[16], Elise Patole-Edoumba[17], Anh Tuan Nguyen[10], Bérénice Bellina-Pryce[18], Jean-Christophe Galipaud[19], Rebecca Kinaston[20], Hallie Buckley[20], Christophe Pottier[21], Simon Rasmussen[22], Tom Higham[23], Robert A. Foley[24], Marta Mirazón Lahr[24], Ludovic Orlando[1,25], Martin Sikora[1], Charles Higham[20], David M. Lambert[3], Eske Willerslev[1,4,26,^]

Affiliations:

[1]Centre for GeoGenetics, Natural History Museum of Denmark, Copenhagen, Denmark.

[2]National Museum of Natural History, Ecoanthropology and Ethnobiology, Musée de l'Homme, Paris, France.

[3]Australian Research Centre for Human Evolution, Griffith University, Nathan, Australia.

[4]Department of Zoology, University of Cambridge, UK.

[5]Institute of Molecular Biology, National Academy of Sciences, Armenia.

[6]Department of Archaeology, Silpakorn University, Bangkok, Thailand.

[7]Department of Museums, Historic Buildings and Archaeology, Ministry of Information, Culture and Tourism, Lao Democratic People's Republic.

[8]Centre for Global Archaeological Research, Universiti Sains Malaysia, Malaysia.

[9]Department of Biochemistry, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand.

[10]Anthropological and Palaeoenvironmental Department, Institute of Archaeology, Hanoi, Vietnam.

[11]Department of Archaeology and Natural History, Australian National University, Canberra, Australia.
[12]Balai Archeology, Medan, Indonesia.
[13]Laboratoire AMIS, Université Paris Descartes, Faculté de chirurgie dentaire, Montrouge, Paris, France.
[14]École et Observatoire des Sciences de la Terre, Université de Strasbourg, Strasbourg, France.
[15]Institut de Chimie et Procédés pour l'Energie, l'Environnement et la Santé, Université de Strasbourg, France.
[16]Department of Anthropology, University of Illinois at Urbana-Champaign, USA.
[17]Natural History Museum of La Rochelle, La Rochelle, France.
[18]Prehistory and Technology, Maison Archéologie et Ethnologie, Nanterre, France.
[19]Research Institute for Development, National Museum of Natural History, UMR Paloc, France.
[20]Department of Anatomy, Otago School of Medical Sciences, University of Otago, Dunedin, New Zealand.
[21]École française d'Extrême-Orient, Paris, France.
[22]Department of Bio and Health Informatics, Technical University of Denmark, Denmark.
[23]Radiocarbon Accelerator Unit, University of Oxford, UK.
[24]Leverhulme Centre for Human Evolutionary, Studies, Department of Archaeology, University of Cambridge, UK.
[25]Laboratoire AMIS, Université Paul Sabatier (UPS), Toulouse, France
[26]Wellcome Trust Sanger Institute, UK.
* Joint first authors
^ Corresponding authors (HM - hmccoll@snm.ku.dk and EW - ewillerslev@snm.ku.dk)

SOM1. Assessment of target enrichment methods

Given that ancient human samples from tropical regions (e.g. Southeast Asia, SEA) are generally quite poorly preserved, their endogenous DNA is expected to be highly degraded. In preparation for the present study, we benchmarked three commercially available capture

systems developed for human genomes, using ancient DNA libraries from samples with low endogenous content. For each kit, we initially compared enrichment at both high and low stringency reaction conditions, followed by a second evaluation on poorly preserved samples.

Experimental design and methods

A pool of sequencing libraries from a single sample (R23) was split in identical aliquots for target enrichment. We included the SeqCap EZ Human Exome Kit v3.0 cat no. 6740294001 (Roche Nimblegen, CA, USA), the SureSelect Human All Exon V5+UTRs cat. no. 5190-6213 (Agilent Technologies) and the Custom MYbaits Whole Genome Enrichment (WGE) Kit version 2.0 (MYcroarray, MI, USA). Each kit was used for the hybridization and recovery of captured libraries. DNA library solutions were evaporated in a vacuum centrifuge at 65°C and reconstituted in water, matching the specific protocol volume. DNA was denaturated at 95°C in reaction buffer(s) and reannealed in the presence of blocking oligonucleotides specific to adapter regions and/or low-complexity DNA specified in the kit protocols. Washing and recovery of captured libraries were performed using the reagents recommended by the manufacturer in each case. The Kapa U^+ PCR enzyme (KapaBiosystems) was used for PCR amplification post capture, according to the manufacturer's instructions. Below, we describe the specific modifications introduced to each of the protocols for reducing capture stringency and for allowing adaptation to one common protocol for library preparation.

a) Roche Nimblegen SeqCap EZ Human Exome Kit v3.0: For the low-stringency reactions conditions, the formamide concentration was decreased to 10% (*85*). Blocking oligonucleotides described in the protocol were replaced with 1 nmol of oligos matching dual-indexed Illumina adapters, including a universal index-binding hexamer inosine motif. Wash and recovery were performed using SeqCap EZ Hybridization and Wash Kit (Roche Nimblegen, # 05634261001).
b) The Agilent Technologies SureSelect Human All Exon V5+UTRs protocol was adapted for use on libraries with full-length adapters. Additional blockers as described above were added to the hybridization reaction (2x1 nmol). Low-stringency conditions in the initial experiments were achieved by lowering the hybridization temperature to 45°C

instead of 65°C. For all hybridization reactions, the lid of the thermocycler was adjusted to +10°C higher than the hybridization reaction temperature.

c) MYcroarray MYbaits Whole Genome Enrichment kit ver. 2.0. For low- or high-stringency reactions conditions, the hybridization temperature was adjusted to 45°C or 65°C, respectively. The block #3 reagent of the kit was replaced with 1 nmol of each of the blockers as described above.

To permit direct comparison of results across the different capture conditions, we randomly sampled 1,000,000 sequencing reads passing trimming quality filters after running *AdapterRemoval2 (75)*. This was achieved using *seqtk* (https://github.com /lh3/seqtk) with default parameters. The sequence data were then aligned against the the human mitochondrial (rCRS) and nuclear genomes (hg19) with BWA 0.5.9-r26-dev (*86*) through PALEOMIX (*87*) and following the procedure described previously (*76*, *88*).

Length distributions and damage of DNA were assessed for each experimental condition using *mapDamage* v2.0 (*80*), disregarding bases showing quality scores (Phred) strictly inferior to 30.

Determining the capture condition stringency

Shotgun sequencing revealed that the endogenous content of the test sample (R23) was 0.39%. Following exome capture, high stringency conditions were found to result in higher fold-enrichment (9.0 - 14.8x) than low stringency conditions (1.2 - 3.2x). The reverse was true for the whole genome capture procedure (Table S1).

We next considered the length distribution and the misincorporation patterns of the captured mapping reads. For both exome enrichment kits, high stringency conditions resulted in an increased median fragment length, compared to low stringency conditions. For whole genome capture, the size distribution at high or low stringency resulted in less different median lengths (Figures S1 and S2).

Using high stringency conditions, the threshold of the lower 25 percentile length of the shortest reads was increased by 9 - 19 bp for all three capture systems, indicating discrimination against the shortest reads. The misincorporation patterns recapitulated those of shotgun sequencing in all cases except that the proportion of damaged reads differed after Agilent exome capture, suggesting that this kit may be particularly sensitive to damaged DNA. Based on these results subsequent samples were captured using high stringency conditions for the exome kits, and low stringency conditions for the whole genome kit, with the aim of maximizing the enrichments of on-target reads.

Analyses of target-enrichment of a panel of poorly preserved human samples

After this initial experiment, we tested a total of eight samples from different geographic locations with ages ranging from medieval to over ten thousand years before present. For each sample, multiple libraries were prepared and pooled to enable direct comparison of the three kits tested. The number of extracts and libraries included in each pool is listed in Table S1.

In order to obtain baseline values for each sample, we estimated the endogenous DNA content of each library pool by shotgun sequencing, which represented a fraction of 0.01% to 5.92% of high-quality reads mapping uniquely against the human reference genome (Table S2). In all libraries, mapping reads showed fragment length distributions and DNA damage patterns characteristic of authentic aDNA (*89*). These patterns were, however, less pronounced in those samples with very low endogenous content, due to the limited number of read alignments available.

After exome enrichment, the proportion of unique on-target reads was commonly found to reach >50 % of the trimmed high quality reads (Table S2). When normalizing the sequencing effort to 1 million random reads, the proportion of the sites for which there was at least one read within the target region (DoC $\geq$ 1) was typically 1 - 6%. At this sequencing effort, this translates into a median enrichment of on-target reads of 14.8 or 18.9-fold, for the Agilent and Nimblegen kits, respectively (Table S3). Although the number of samples is limited and sample-specific variation is observed, there was a tendency for the Nimblegen exome kit to

produce more on-target reads than the Agilent kit. The whole genome enrichment resulted in on-target rates ranging between 1.1 and 18.2%, which translates into a median enrichment factor of 7.54-fold (range: 3 - 228).

In Figure S1, we show, for all libraries, the medium lengths and 25-75% quartile ranges of the mapping fragments. For all samples the mapping fragments in captured libraries were significantly longer than in shotgun libraries (Kolmogorov–Smirnov test, $p \leq 2.93*10^{-6}$).

Complexity of ancient DNA libraries is reduced post capture

As captured ancient DNA libraries may sustain only relatively shallow sequencing, before PCR-generated duplicate reads became excessive, we investigated the clonality of the post capture libraries. Figure S3A shows that the clonality for certain libraries approached 50%. However, these samples were predicted to have very limited complexity. The remaining samples showed between 0.2% and 19.9% clonality.

When increasing the sequencing volume, there seemed to be a general trend for the clonality to be negatively correlated with the predicted endogenous input. However, the variation among samples and kits is too great to establish a statistically significant relationship. In our experiments, the level of clonality seemed to be less pronounced after whole genome capture (MYcroarray) as compared to the exome capture. Nevertheless, more samples are needed to substantiate this claim.

We also investigated the library complexity using PRESEQ version 2.0 (*90*). Using 5 million random trimmed reads, robust statistical predictions were possible for 80% of the libraries. More specifically, we predicted the total number of bases uniquely covered following a sequencing effort of up to $>1*10^{10}$ bases per library. Libraries captured using the MYcroarray WG kit generally showed only slightly lower complexity that that of shotgun-sequenced libraries. In contrast, exome capture commonly resulted in reduced complexity by several orders of magnitude (Figure S4). In the three cases that allowed for direct comparison, predicted complexity after exome capture was superior using the Agilent kit (samples R23, K, S60). Exhaustion was predicted in most exome-captured libraries by $2.5\text{-}10*10^{8}$ bases

(corresponding to ~4.2-16.6 *10^6 reads of average length). Overall, the predictions from PRESEQ reflect the results from the other assessment of library clonality.

Overlapping SNPs

To provide a measure of the applicability of target enrichment in population genetic analyses, we compared the number of SNPs in the captured and the non-captured data overlapping with the >644K SNPs in the Human Genome Diversity Panel (HGDP) (*91*). The HGDP and the target regions of the Agilent or Nimblegen exome kits shared a total of 21,487 and 24,910 sites, respectively. In Figure S5, we show the number of HGDP-overlapping SNPs for increasing sequencing effort, following shotgun sequencing or capture. We found that generally ~10 million post-capture reads were sufficient for obtaining >1,000 overlapping SNPs, which are sufficient for determining broad-scale continental ancestry of a sample (Figure S5 A, B or C). In contrast, shotgun sequencing could only achieve similar overlap from ~30-50 million reads. For one sample, even >34 million shotgun reads were insufficient to reach 100 SNPs (Figure S5D).

Upon capture, more HGDP SNPs were covered within the target region, as compared to an equal number of randomly chosen non-target HGDP SNPs. In contrast, shotgun sequencing resulted in balanced coverage of SNPs within or outside the target regions. This difference was statistically significant except in four cases with too few comparable SNPs (Fisher's exact test, Figure S6).

Conclusion of target enrichment experiments

In cases where shotgun sequencing is prohibited by very low endogenous content and/or availability of sample material, target enrichment may provide a powerful mean for obtaining sequencing information required for population-level genetic analyses. The data show that at 1 million reads the median fold enrichment was 14.8 - 18.9-fold for the exome kits and 7.53 for whole genome enrichment.

The difference in fragment length distributions between pre- and post-capture show that the enriched library represents a sub-sample of the shotgun sequencing library. Other studies have also documented that capture, while enriching, also increases average fragment lengths by up to 20 bp (*92–94*). Given these findings, we can expect that subsampling by capture of low-complexity libraries to be challenging.

Nevertheless, we generally found higher complexity upon whole genome capture compared to exome-captured libraries. Predicted complexity in whole genome-captured libraries was comparable to that of shotgun sequencing libraries. As the exome constitutes only a minor proportion of the whole genome, complexity was substantially lower in libraries enriched for the exome. We found a trend of lower clonality in libraries with higher predicted input. Although enrichment may be critically sample-dependant, for future aDNA studies we suggest to increase complexity of the capture reactions by maximizing the number of combined libraries and limit PCR amplification, which may introduce DNA polymerase-specific biases (*95*).

Sequencing >6.5 million reads of captured libraries (using exome or whole-genome) was sufficient to yield >1,000 SNPs overlapping with the HGDP panel. Whole genome capture resulted in ≤13,359 HGDP SNP. These results suggest that discrimination of samples with low endogenous content is possible after limited sequencing of captured libraries.
These experiments provide the basis for conducting future capture experiments using MYcroarray whole genome probes (at low stringency) on samples whose endogenous DNA content is too low for shotgun sequencing.

SOM2. Archaeological Overview

We obtained ancient genomic data from 41 ancient samples found in Vietnam, Laos, Thailand, Malaysia, Indonesia and the Philippines (Table S3, Figure S7). All samples were dated directly at Oxford Radiocarbon Accelerator Unit, where the carbon yield allowed. Dates for five samples (La368, Ma911, Vt739, Ma912, Vt778) with low carbon yields and high C/N ratios are provided with a cautionary note on the absolute accuracy of the measurement.

Based on archaeological and anthropological studies, the individuals samples fall into two broad groups: Austro-melanesian hunter-gatherers and East Asian Neolithic farmers. The oldest individual comes from Pha Faen (Laos) 6-10,000 years ago (*96*). The genome belongs to a tall individual (ca. 176 cm) who was identified as male based on an osteological assessment, and who was interred in a flexed position (a common Hoabinhian burial position). The individual was dated to ^{14}C 7,040 ± 38 and was not interred with any associated mortuary offerings (*96*). Gua Cha (Malaysia) and Ma Da Dieu (Vietnam) had two phases, the first was Hoabinhian hunter-gatherers followed by the arrival of Neolithic farmers. Two individuals from Gua Cha were from the former context, while a third is from the latter (*97*). We recovered ancient DNA from a complete skeleton, interred fully extended as is typical of other Neolithic occupants of this site. Ancient DNA was also extracted from the petrous bones of two individuals from the Hoabinhian occupation that were not associated with any complete skeleton. The two Ma Dai Dieu samples (^{14}C 3,788 ± 35 and 2,275 ± 24), though ca. 1.5 kya apart, correspond to the end of the period of occupation of the site.

Unlike the previously described sites, the Hoabinhian hunter-gatherer Tam Hang samples does not appear to be from Australo-melanesian hunter-gatherers (*1*, *2*, *98*). It is instead similar to local Neolithic/Bronze age populations (Dong Son). Stone tools were found across all layers, typically Hoabinhian with Sumatraliths (*6*, *99*). The teeth that yielded ancient DNA were from two of the six individuals who were interred in a flexed position in a shell midden at the site (*100*). Cord marked pottery was also found (*100*). The ^{14}C dates from charcoal indicate an occupation from 13 to 3 kya but direct ^{14}C dating of one of the human teeth that yielded ancient DNA proved problematic: one sample appeared highly contaminated and provided a date of 6,980 ± 40 BP, with the second hydroxyproline date of 2,308 ± 30 BP was deemed more accurate. Two subsequent attempts at dating a second and third tooth from the other individual with ancient DNA failed due to high C/N ratio. The relatively recent dating and Dong Son affinities suggest that the dated individual was not Hoabinhian, and instead belonged to a more recent intrusive burial.

At Tam Pa Ping, Northern Laos, nearby Tam Hang site, a tall male (^{14}C 2,865 ± 29) - morphologically similar to individuals from local Neolithic/Bronze Age populations - was found in an extended burial with a bronze axe between the ribs. The Bronze Age burial included cord mark pottery.

The coastal fishing and farming site of Hon Hai Co Tien (*101*) belongs to the Ha Long Culture and was dated to around 4ka. From the five samples yielding aDNA from this site, one fit with the expected age (^{14}C 3,755 ± 60), three failed dating (low %N, high C/N ratio - 4.8), and the final (^{14}C 223 ± 23) indicated a recent intrusive burial. The site contains mostly corded fine pottery (bracelets, necklaces and stone tools, like hand axes and hoes). Lead pieces used for fishing nets indicate a more recent period of occupation, reflected by the second ^{14}C date.

Nam Tun, Vietnam (^{14}C 2,549 ± 28) contained only surface pottery (*102*, *103*), similarly to Mai Da Dieu. At the early metal age site Nui Nap, Vietnam (^{14}C 2,264 ± 64, 2248 ± 24, 2242 ± 24, 2,179 ± 91) cord marked pottery was found.

Long Long Rak cave is a cemetery site located in highland Pang Mapha, northwest Thailand. It contained Iron age wooden coffin burials with individuals dated from ^{14}C 1,792 ± 25 to 1,687 ± 24, along with corded pottery, iron implements, wooden pot covered with lacquer, weaving loom, basket, fabric and ornaments made of plants and glass beads (*104*). This site dates between 1,940 ± 30 to 1,636 ± 44 BP (*105*). Isotope analysis has supported Long Long Rak as a farming population (*50*).

In Island Southeast Asia (ISEA), the Nagsabaran Site (situated in Cagayan Valley of Luzon) is one of the oldest sites in the Philippines, extending as early as 4,200 years BP in its deepest pottery-bearing layer. The site's deposits have yielded red-slipped pottery, Taiwan jade ornaments, and rice remains dating to the beginning of the Austronesian expansion into this region. One extended-position burial was discovered in the Neolithic layer prior to 2,500 years BP, while others from the overlaying Iron Age contexts (2,500–1,500 year BP) were in extended positions, flexed formats, and secondary jar burials (*106*, *107*). Many of those Iron Age burials included mandibles but not skulls, and sometimes the skulls were re-deposited

inside the jar burials. A similar practice of keeping skulls in pottery has been noted in one instance at the Lapita-age cemetery of Teouma in Vanuatu, dated about 2,900 years BP (*108*). One of the Nagsabaran bone samples at ^{14}C 1,877 ± 27 represented an Iron Age farming group in the Philippines. Jar burials, in addition to interments in a wide variety of burial positions (flexed, supine, prone, some with skulls removed) have also been found at the 3,000-2,000 BP burial ground of Pain Haka, on eastern Flores, Indonesia (*109*).

Although flexed-position burial had characterized the older hunter-gatherer sites of SEA, a later tradition of extended-position burial seems to have been introduced with the appearance of rice-farming societies. Curiously, the flexed-position burial tradition regained its popularity in several ISEA sites. For example, there are two flexed burials at the cave of Loyang Ujung Karang dated to ^{14}C 2,152 ± 26 and 1,917 ± 25, contemporaneous with the upper-layer findings at a nearby cave site of Loyang Mendale (*110*), collectively representing the late Neolithic to early Iron Age population of Sumatra. Shell middens and fine pottery characteristic of the Austronesian expansion were also found at the site of Loyang Ujung Karang.

A recent wooden coffin burials from Kinabatagan (^{14}C 299 ± 23) and Supu Hujung 4 (^{14}C 383 ± 23) on Borneo also yielded fine pottery (*111*).

SOM3. Mapping

Reads were trimmed using AdapterRemoval 2.2.2, in order to remove adapters, terminal N's (--trimns), low quality bases (-trmi qualities, --minquality 2) and short reads (--minlength 30). Mapping, library merging and local realignment were undertaken as described in (*54*). Merging to library level and duplicate removal was undertaken with *picard* (http://broadinstitute.github.io/picard/). Local realignment was done with GATK (*77*, *112*) and MD tags called using *samtools* calmd v1.5-2 (*86*).

SOM4. Principal Component Analysis

To identify position of the ancient SEA samples within PC2 from the lower panel of Figure 1C, we first removed Yoruba (Figure S8). PC2 now separates the Melanesians from the East Asians, although we note that the Group 1 samples (La368, Ma911) and Onge do not fall within this cline. Groups 2 and 5 are clustered closer to the Malaysians relative to Groups 3, 4 and 5, though further differentiation among clusters is very subtle. The 2240k panel has limited SEA populations, so, to assess which clusters from Figure 1B had a representative population in the higher resolution panel, we calculated a PCA with the same populations as in Figure 1B, but including the whole genomes from the Simons Genome Diversity Panel (*64*) (Figure S9). We show here that there are no SGDP samples overlapping with the clusters containing Group 2 and 4. Finally, we recreated Figure 1C with population labels (Figure S10), so as to provide increased resolution as to the placement of particular samples.

SOM5. ADMIXTURE fitting

Here, we describe some qualitative observations we made when performing inference using ADMIXTURE and *fastNGSadmix*, and which allowed us to better tailor our other models (Figure S11).

At K=2, we observe a blue component that is maximised in Yoruba, and a light pink component that is present in most EA and SEA populations, while Europe and South Asian (SA) populations are modeled as a mix of these two components. Certain SEA populations, like the Malay, Papuan 'Negritos', Onge, Jarawa and Melanesians are also a mix. The Tianyuan and the Group 1 individuals show an unusually large blue component for the region.

At K=4, a dark purple component is maximised in the Jehai and Melanesians. This component is also present at >50% frequency in many ISEA populations (including Group 5). On the mainland, this component reaches ~50% frequency in the Mlabri, Htin and the Group 2 samples.

At K=5 the component that is dark purple in Melanesians and western Indonesian at K=4 now becomes black, and is also seen in Group 6. The Onge and Group 1 samples now share a blue+black+purple profile.

At K=6, a dark green East Asian component is maximised in the Ryukyuan and Japanese, while a pink East Asian component is now maximised in the Ami. While the dark green component is present in almost all EA, SEA and some SA populations, it is absent in the Mlabri, the Jehai, and most ISEA individuals. When looking at the ancient samples, we also observe that it is absent in In662 (Group 5) and all Group 2 samples.

At K=7, the light green component is maximised in the Mlabri, followed by other Austroasiatic SEA populations and Group 2. This component is present in all SEA populations, and is at highest frequency in the populations lacking the dark green component at K=6.

At K=13, the dark pink component is maximized in the Hmong, and allows SEA populations to be separated into two groups, represented in the ancient samples by Groups 3 and 4. This component is restricted to Tai-Kadai-speaking SEA populations, whereas the Austroasiatic populations are primarily composed of light and dark green components. However, among the SGDP populations, we can also see Austro-Asiatic speaking Kinh and Cambodian populations that also have this dark pink component.

SOM6. f3 Statistics

To identify which present and ancient samples have the most shared drift with the ancient groups, we calculated f3-statistics of the form f3(Yoruba; Y, X), for all populations (X) and all ancient groups (Y). We grouped ancient samples according to their position in the PCA and their inferred ADMIXTURE ancestry components: Group 1 - Ma911, La368; Group 2 - La364, La727, La898, Ma912, Vt833, Vt880; Group 3 - Vt777, Vt779, Vt781, Vt796, Vt808; Group 4 - Th519, Th521, Th530, Th703; Group 5 - In661, In662; Group 6 - Ma554, Ma555, Phl534. Group 3.1 (Th531, Vt719) and Group 4.1 (Vt778) appeared similar in the PCA and

ADMIXTURE to other members of their assigned groups, but were either geographically or temporally distant from the other samples within the group.

Group 1 individuals, the ancient Hoabinhians, share the most drift with the ancient MSEA samples, followed by present-day Onge, followed by the Malaysian 'Negritos' and the Ami (Figure S12, Table S4).

Group 2 individuals share the most drift with the Austroasiatic groups Mlabri and Htin (Figure S13, Table S5). Out of the closest seven populations, four are Austroasiatic speakers, and the remaining three are Austronesian speakers. The three Austronesian populations are all from Java, and carry the largest mainland component of the ISEA present-day population. No ancient sample has similar amounts of shared drift.

Groups 3 and 4 show the most shared drift with each other. The closest present-day population for both is Ami, but there are notable differences among ancient groups (Figures S14, S15, Tables S6, S7). Group 3 has affinities to Tai-Kadai speakers, the Hmong Mien and Han, whereas Group 4 has affinities to Austroasiatic groups instead. When looking only at groups 3.1 and 4.1 (Figures S16, S17, Tables S8, S9), we observe that the Hmong Miao and the Austroasiatic Mlabri are the closest present-day populations to each of them, respectively.

The present-day populations that share the most drift with Group 5 are the Austroasiatic Htin and the Austronesian Ami, followed by present-day Indonesian populations (Figure S18, Table S10). Similarly, the highest two f3-statistics for ancient samples are Group 6 (Austronesian affinities), and Group 2 (Austroasiatic affinities). Group 6 clearly shares the most drift with Austronesian Taiwanese and Filipino populations (Figure S19, Table S11).

We can therefore ascertain that there is a strong association between:

a) Group 1 and Jehai
b) Group 2 and the Mlabri and Austroasiatic Populations
c) Group 6 and the Austronesian populations.

Despite a close relation between Groups 3 and 4, differential affinities to the Hmong and Austroasiatic populations are apparent in these two groups. Group 5 appears to have both Austronesian and Austroasiatic affinities, supporting theories of a mainland Austroasiatic migration prior to the Austronesian expansion.

SOM7. D-statistics

We calculated D-statistics (*113*, *114*) using *AdmixTools* (*36*). For all tests, we removed transitions to minimise outgroup attraction resulting from damage present in aDNA.

Here, we list all D-statistic tables, and denote X as the candidate or test population over which we cycle in each table, and then describe particular observations which are important to the points discussed in the main text.

Table S12: D(Papuan,Tianyuan;X,Mbuti)
Table S13: D(X,Tianyuan;Papuan,Mbuti)
Table S14: D(Papuan,X;Tianyuan,Papuan)
Table S15: D(Onge,Tianyuan;X,Mbuti)
Table S16: D(X,Tianyuan;Onge,Mbuti)
Table S17: D(Onge,X;Tianyuan,Onge)
Table S18: D(Tianyuan,Han;X,Mbuti)
Table S19: D(Mixe,Surui;X,Mbuti)

aDNA damage in Tianyuan

To explore the damage signal we detected in *qpGraph* (SOM9), we calculated D statistics of the form D(Tianyuan,Han,X,Mbuti) and found significant D-statistics when setting X to be archaic humans (Denisova Z =-3.6, Neanderthal Z = -3.8) and great apes (Gorilla Z= -5.5, Macaque Z = -5.1, Chimpanzee Z = -5.6 and Orangutang Z = -5.8). The stronger D statistics in the great apes compared to archaic humans suggests significant outgroup attraction as a result of damage, even in the absence of transitions (Table S18).

Relationship between Papuan, Tianyuan and EA/SEA/Ancients

We find support for Australians and Bougainville islanders forming a clade with Papuans, to the exclusion of Tianyuan (Table S14). In turn, many EA and SEA form a clade with Tianyuan, to the exclusion of Papuan (Table S13). Onge, Jarawa and Jehai do not form a clade with either Papuans or Tianyuan (Table S13, S14), but have a stronger affinity to Papuans than to Tianyuan (Z = 3 - 4.2, for D(Onge/Jarawa/Jehai, Tianyuan; Papuan, Mbuti)).

Relationship between Onge, Tianyuan and SEA

We find Onge, Jarawa and Jehai form a clade with Onge to the exclusion of Tianyuan, but no other EA or SEA population form a clade with Onge, to the exclusion of Tianyuan (Table S17).

Relationship to Surui and Mixe

We tested for a specific affinity in the Surui to our ancient samples, as was previously detected in Papuans, Onge and Tianyuan (*45*, *53*, *54*). For the 2240k panel, we find that D-statistics of the form D(Mixe, Surui, Group 1 individual, Mbuti) are high but non-significant (Z = -2.18 and -2.5, using Ma911 and La368, as the Group 1 representative, respectively) (Table S19).

SOM8. TreeMix fitting

To find well-fitting admixture graphs on which we could place different ancient SEA populations, we used TreeMix (*55*), allowing for either 0, 1, 2 or 3 migration arrows. We only used transversions and included, in all graphs, the following samples: the high-coverage Denisovan (*58*), Kostenki-14 (*60*), the Tianyuan individual (*45*), Papuans, Onge, and two East Asian populations (Han and Ami) genotyped on the 2240k panel . We call these groups the "base populations" in each of the graph legends. To account for damage in the ancient samples, only transversion were included. We explored various combinations of Group 1 ancient samples to overcome limited SNP overlap between the low coverage samples.

Figures S20-S32 show the graphs fitted using combinations of the base populations and additional samples:

Figure S20: base populations only

Figure S21: base populations + La368

Figure S22: base populations + Ma911

Figure S23: base populations + La368 + Ma911

Figure S24: base populations + La364 + Ma912

Figure S25: base populations + La368 + La364 + Ma912

Figure S26: base populations + Ma911 + La364 + Ma912

Figure S27: base populations + La368 + Ma911 + La364 + Ma912

Figure S28: base populations + La364 + Ma912 + Jehai

Figure S29: base populations + La364 + In661

Figure S30: base populations + La364 + In662

Figure S31: base populations + La364 + Ma554

Figure S32: base populations + La364 + Ma555

SOM9. qpGraph fitting

We used a previously estimated admixture graph of worldwide populations (*61*), and attempted to fit different SEA individuals in it via *qpGraph* (Patterson et al. 2012). We first built a skeletal framework that included Denisova (Meyer et al. 2012), Kostenki (Seguin-Orlando et al. 2014), Mbuti, Papuan, Onge and Ami. We note that the degree to which we can assess differential population relationships in the early splits among East Asian populations is limited. Indeed, without including Tianyuan, we find that the split between Onge, Papuans and Ami populations is effectively a trifurcation (Figure S33) (Lipson et al. 2017), though including Tianyuan suggests that East Asian populations (like Ami) are best modeled as a mixture of a Tianyuan-like component and sister component to Onge (Figure S34, worst-fitting Z = -3.564).

We used the skeletal framework from Figure S34, and then attempted to fit ancient SEA individuals with relatively high coverage (> 0.1X) onto it. We first fit La368 (Figure S35, worst-fitting Z = 3.372), then Ma911 (Figure S36, worst-fitting Z = 3.803), and then La368 jointly with La364 (Figure 3E, worst-fitting Z = 3.667).

We also fit Dai (a present-day SEA population) as a mixture of La364 and an EA component, based on our previous inferences using *ADMIXTURE/fastNGSadmix* and *TreeMix* (Figure S37, worst-fitting Z = 3.66).

While including transitions for ancient samples introduces bias, removing the transitions also removes a large number of informative sites when samples are of low coverage. We therefore repeated analyses with and without transitions for each fit, to verify our graphs were consistent. Here, we only show graphs using data excluding transitions. For all graphs with or without transitions, including Tianyuan led to at least one Z score having a value larger that 3. All worst f4 Z-scores included Tianyuan, even in very simple trees. Adding an admixture event from more basal points to Tianyuan progressively improved this fit, with the best fit being 1% admixture from the branch leading to chimpanzee. This was also observed in *TreeMix*, and in both cases was not improved by removing transitions to account for C-to-T damage. We do not believe this reflects an archaic admixture event, but is more likely a consequence of outgroup attraction induced by highly damaged DNA or poor overlap in coverage among samples. We therefore also tested each *qpGraph* model with an artificial "damage admixture" event, inducing a 1% admixture from the chimpanzee branch to Tianyuan, to assess best fitting models without bias introduced from Tianyuan. In all cases, the best fits for the Southeast Asian individuals considered in this study were consistent, regardless of whether this artificial event was added or not.

SOM10. Measurements of archaic ancestry

We first aimed to determine the amount of Neanderthal ancestry in different Southeast Asian populations. For this we computed an F4 ratio of the form f4(X, Yoruba; Altai Neanderthal Chimpanzee)/f4(Mezmaiskaya Neanderthal, Yoruba; AltaiNea, Chimpanzee) (Green et al. 2010), which is meant to measure the proportion of Neanderthal ancestry in a non-African population X, using an African population (Yoruba) as a non-admixed baseline. This statistic is particularly elevated in several ancient samples (Figure S38). However, this statistic may also

be confounded by Denisovan ancestry that could be present in these populations, due to shared ancestry between Denisovans and Neanderthals.

We thus aimed to determine the amount of Denisovan ancestry in different ancient Southeast Asian populations, as a proportion of the ancestry found in present-day Papuan / Aboriginal Australian populations, using Han or French as baseline non-Denisovan-admixed populations. For this, we computed an F4 ratio statistic (pD) (*115*) of the form: pD(X) = f4(Denisova, Mbuti; X, Y) / f4(Denisova, Mbuti; Papuan, Y) where X is a population of interest, and Y was either Han (Figure S39) or French (Figure S40). For this analysis, we only used transitions to avoid possible biases due to ancient DNA damage, and excluded Ma525 due to its low coverage. pD(X) estimates a quantity proportional to the percentage of total Denisova ancestry in X, computed as a fraction of the total excess Denisovan ancestry found in Papuans, relative to a baseline population (Han or French). We observe that Ma554 shows a high proportion of relative fractional Denisovan ancestry when using Han as the baseline (37.8% of the total Denisovan ancestry found in Papuans), and the same pattern is observed if we replace Han for French as the baseline population. In particular, the amount of fractional Denisovan ancestry in Papuans is almost equivalent to the one observed in Tianyuan (Figures S39, S40).

Tianyuan is known to harbor a large proportion of Neanderthal ancestry, so it is possible that the large amount of Denisovan ancestry observed in Ma554 is caused by high amounts of Neanderthal ancestry (due to shared genetic affinity between the Neanderthal and Denisovan populations, to the exclusion of modern humans). We observe, however, that while Neanderthal ancestry in Tianyuan is elevated relative to other present-day and ancient Asian populations, this is not so much the case for Ma554 (Figure S41). This suggests that the high archaic ancestry in Ma554 should be attributed to Denisovans, rather than Neanderthals. Although it is located in the mainland, Ma554 exhibits Denisovan ancestry levels comparable to present-day Indonesian populations (Figure S41) (*115*). Given its affinity to other Austronesian populations, it is possible that Denisovan ancestry in this individual is linked to shared ancestry with Indonesian populations with elevated Denisovan ancestry, which later got replaced in the mainland by populations with considerably more reduced Denisovan ancestry.

Supplementary Figures:

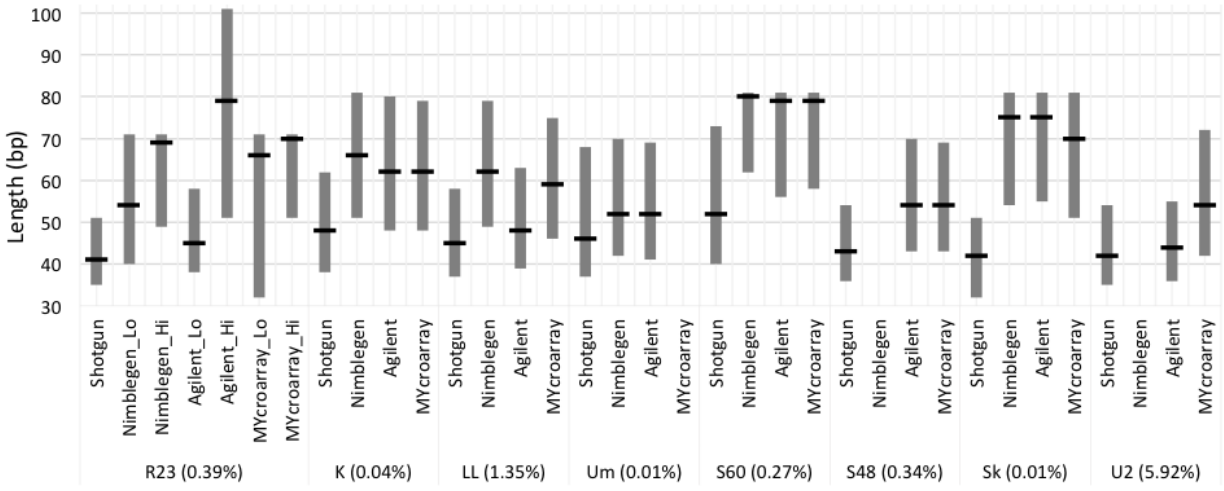


Figure S1. Length of sequenced reads. The median length (—) of aDNA libraries reads are shown for each sample before (Shotgun) and after enrichment of exome (Nimblegen or Agilent) or whole genome (MYcroarray). The 25- and 75- percentile ranges are indicated (bar).

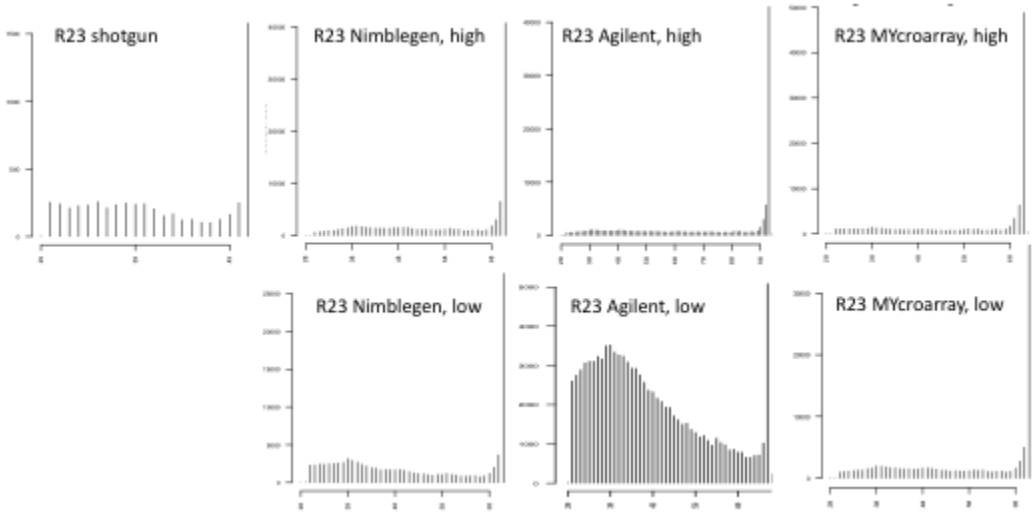


Figure S2. DNA fragment length distribution of libraries of control sample R23. Each panel represents the results from shotgun sequencing or capture using kits and conditions as indicated above. Note that the axes may show different ranges.

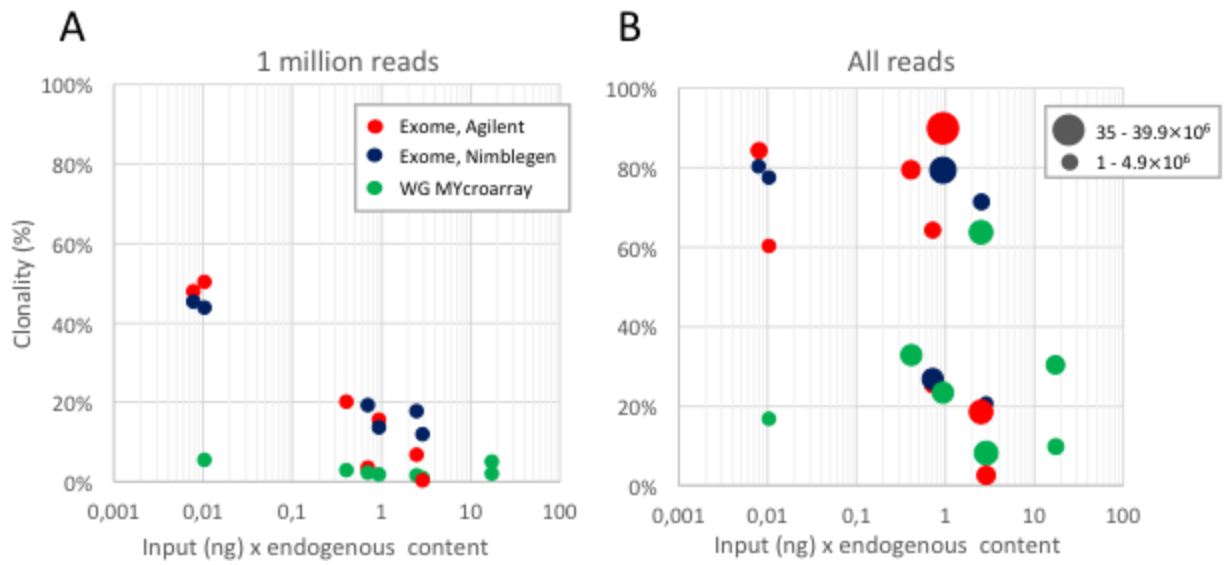


Figure S3. Clonality of captured aDNA libraries using one million trimmed reads (A) or the total sequence data generated (1.8 - 39.4 million reads) (B). The proportion of clonal reads are expressed as a function of the predicted complexity; (i.e. The fraction of the endogenous library DNA). In Panel B, the total number of produced reads is indicated by the relative size of the data mark. WG: Whole genome.

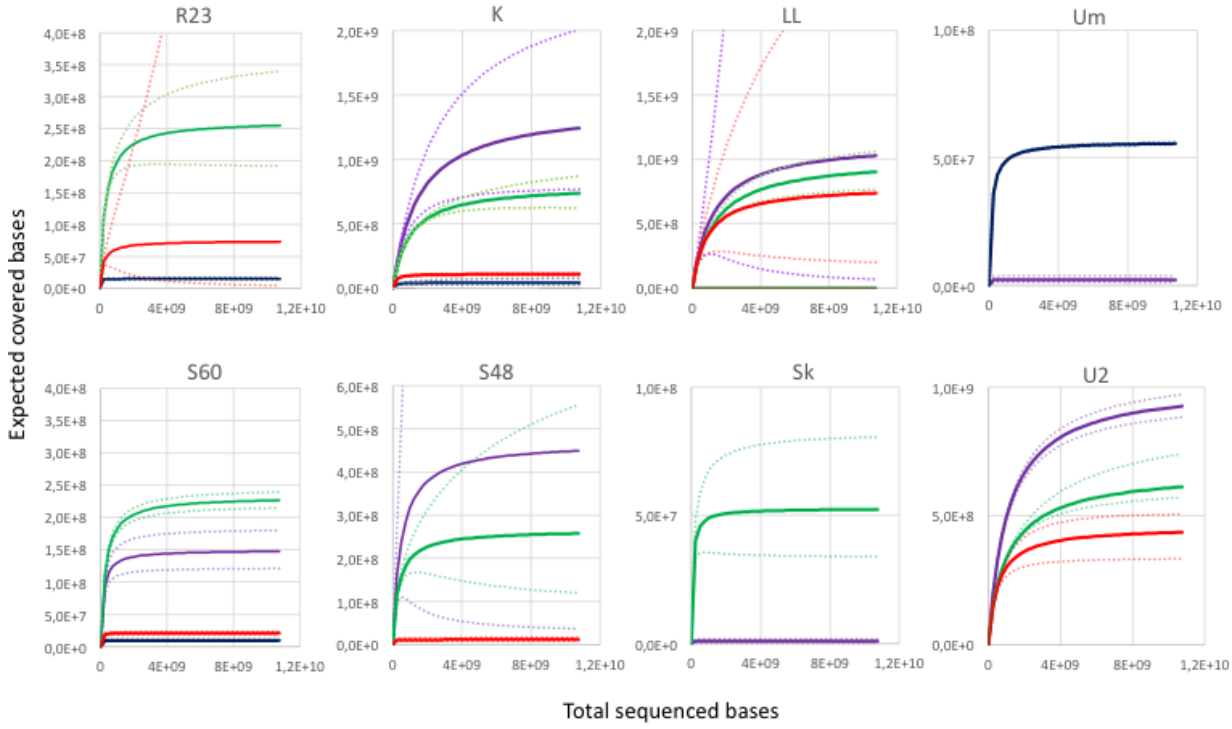


Figure S4. Predicted complexity in libraries pre- and post-capture, using PRESEQ. The number of expected covered bases (solid lines) is shown as a function of sequencing volume pre- (purple) and post-capture of whole genome (green) or exomes (red and blue for Agilent and Nimblegen, respectively). Confidence intervals (95%) are also indicated (dashed lines). Note that the y-axes may have different ranges.

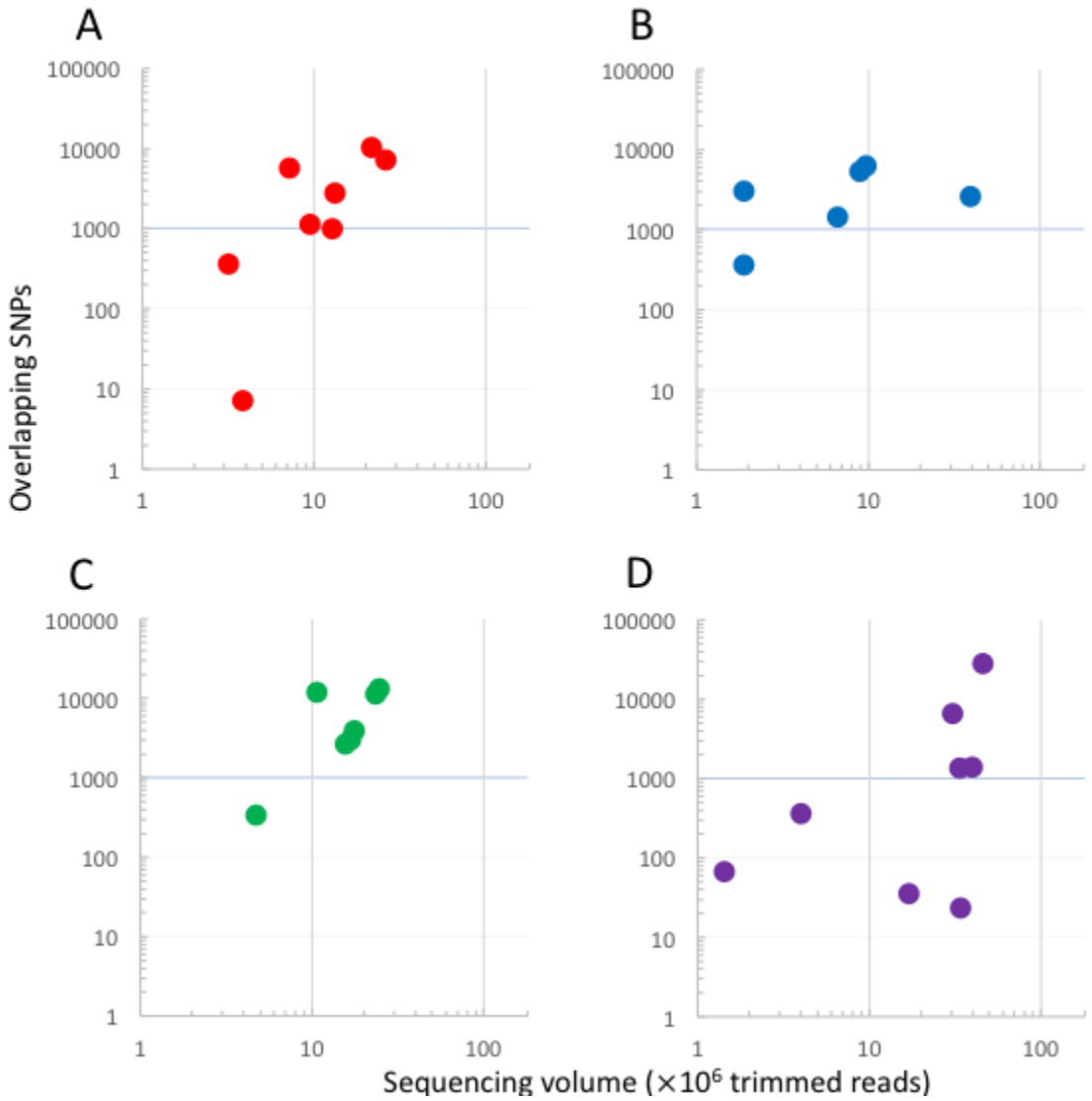


Figure S5. The number of SNPs overlapping with the Human Genome Diversity Project panel is shown as function of sequencing effort (expressed as trimmed reads). Libraries were target-enriched using the exome+UTR kits from Agilent or Nimblegen (A or B, respectively) or the whole genome enrichment kit from MYcroarray (C) or shotgun sequenced (D).

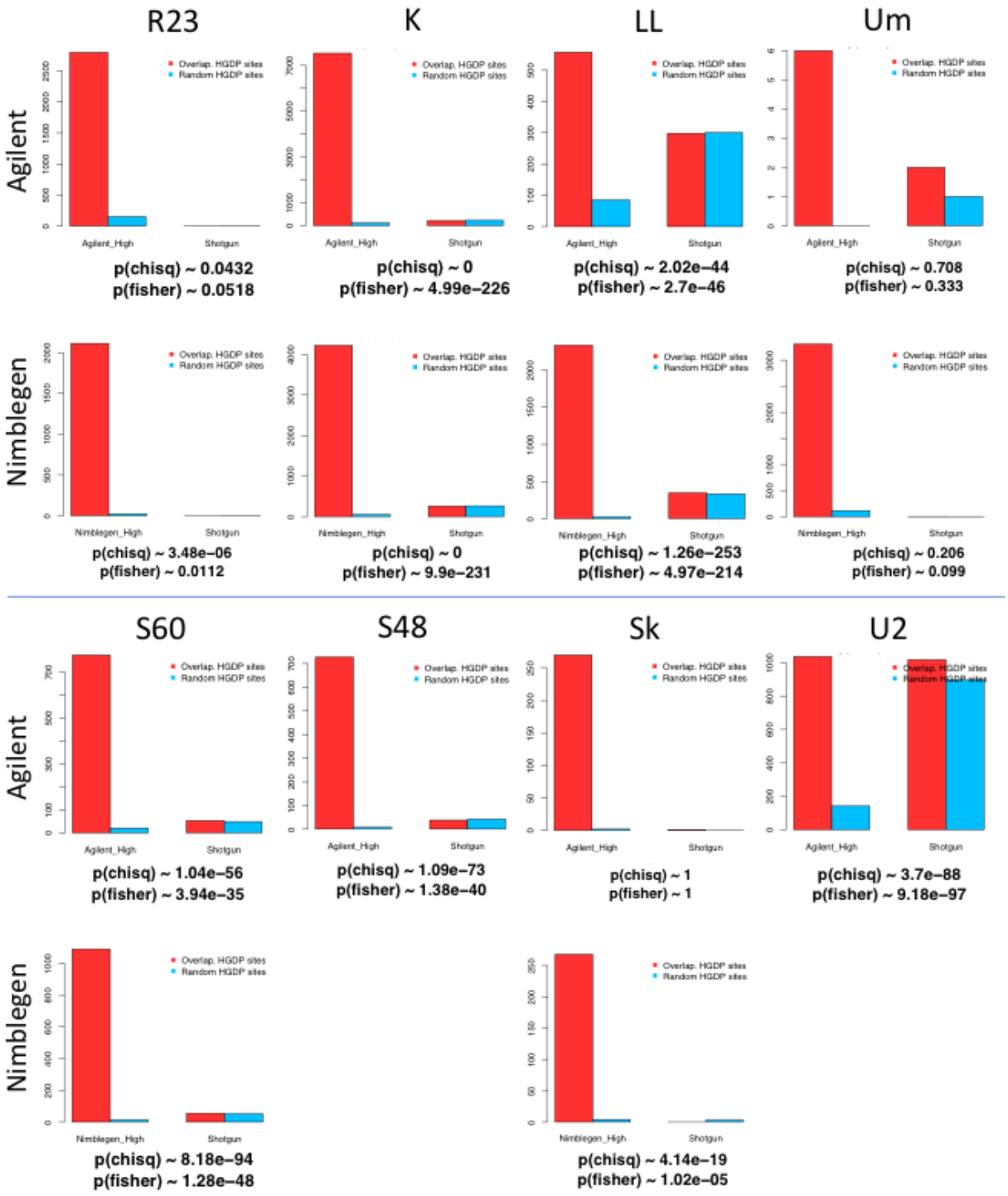


Figure S6. For each sample, indicated above, we show the number of HGDP SNPs covered within the target region (red), and randomly chosen outside (blue) after exome capture (Nimblegen or Agilent as indicated). Likewise, we show the same SNP coverage after shotgun sequencing, as well as the results of statistical testing (chi^2 and Fisher's exact test, p-values indicated below each panel). Note that the y-axes may have different ranges.

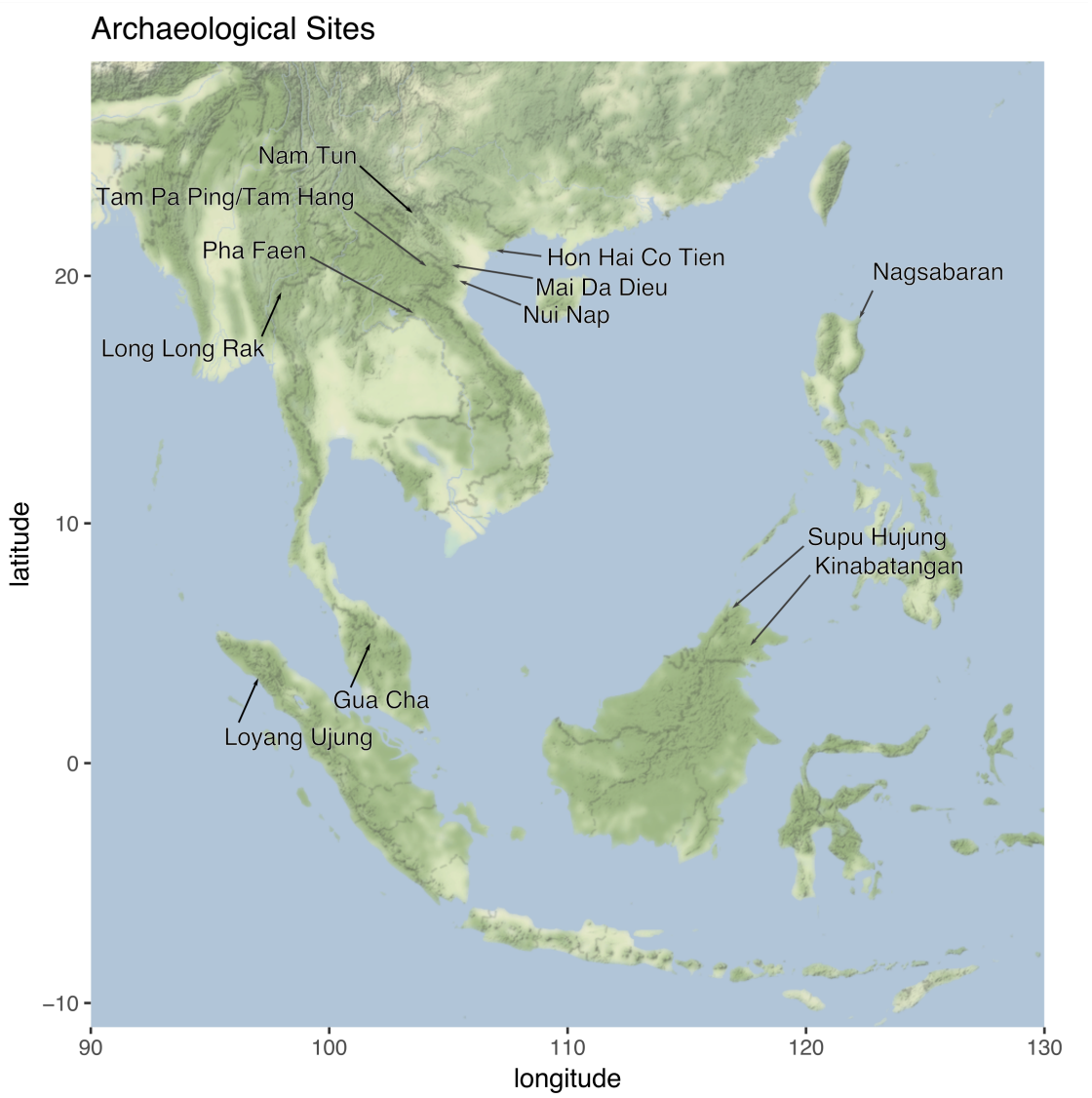


Figure S7. Location of archaeological sites from which we obtained ancient genomic data.

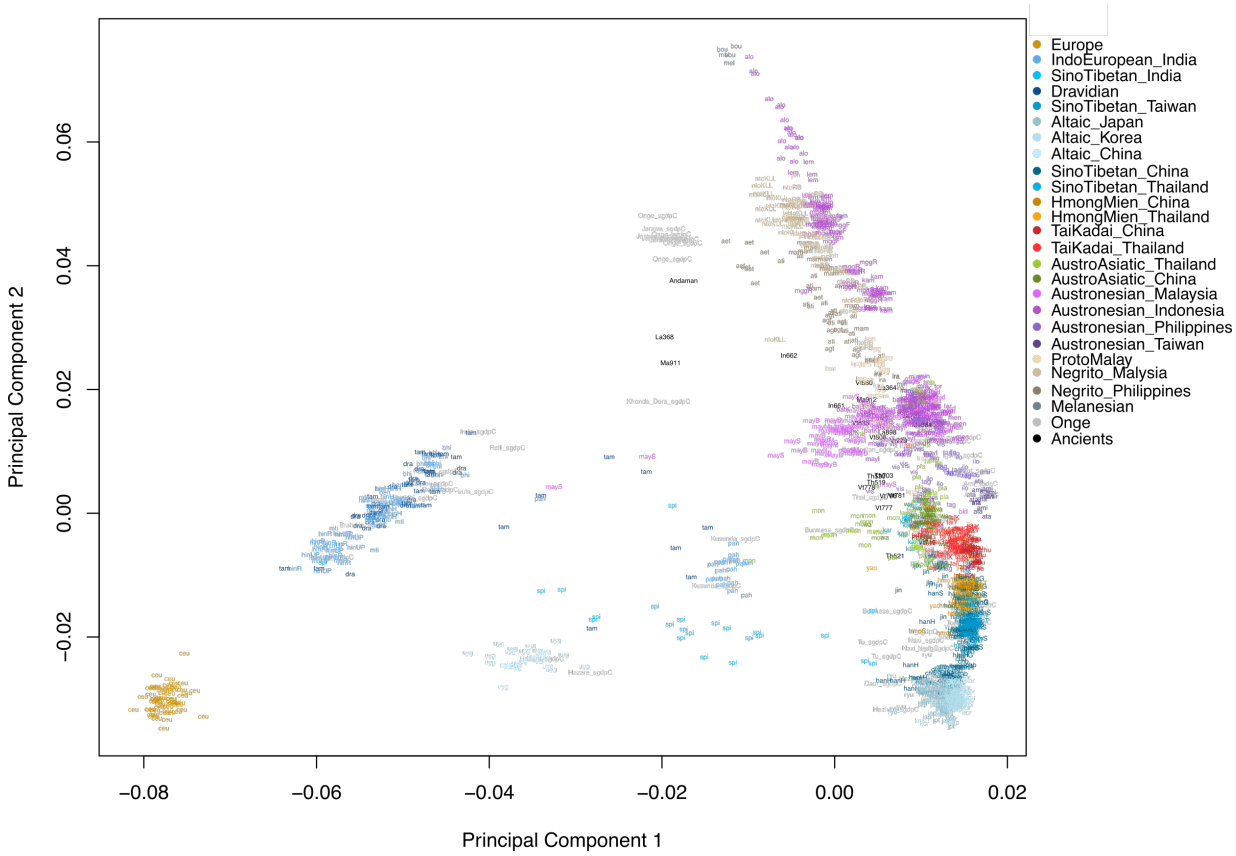


Figure S8. PCA of worldwide populations from Pan-Asia Panel, excluding Yoruba, and including Onge.

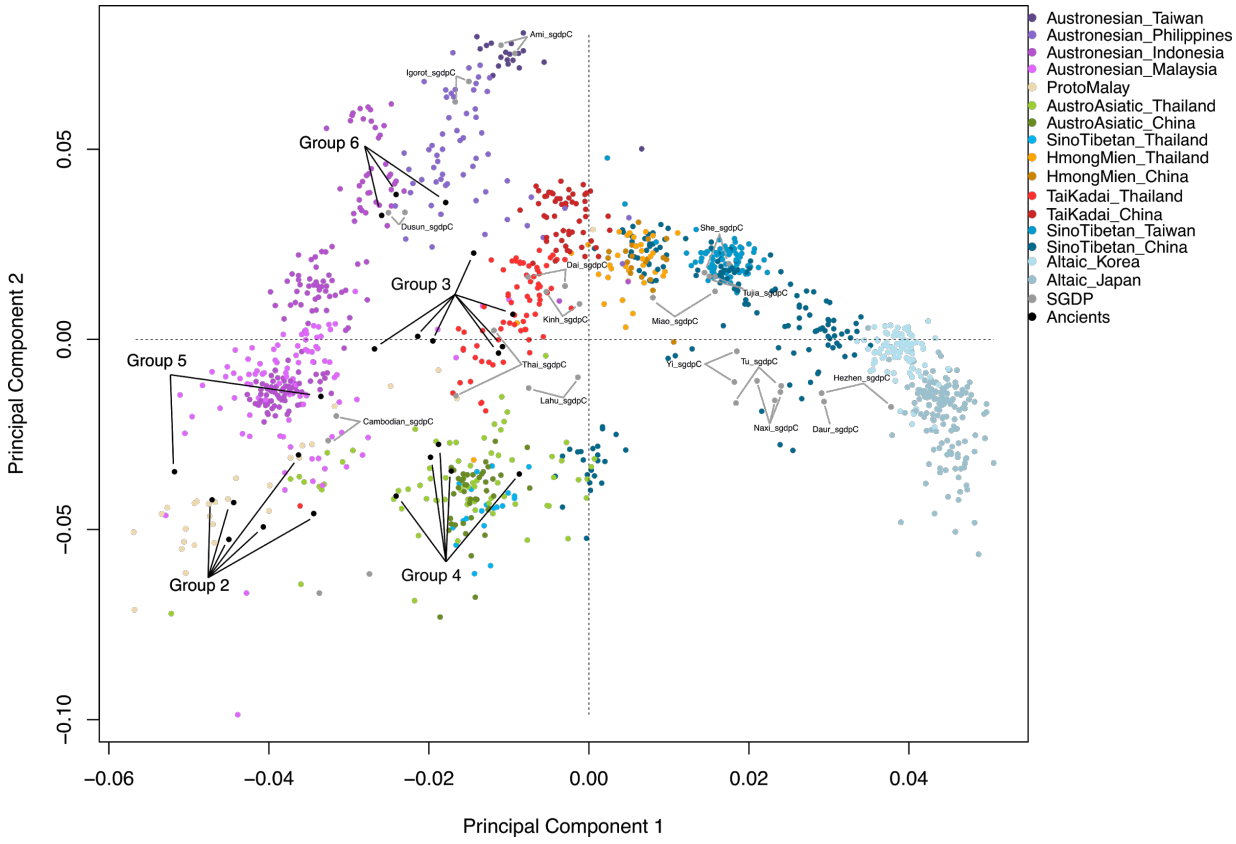


Figure S9. PCA from Figure 1B including SGDP individuals, showing the lack of present-day whole genome (SGDP) data representing clusters encompassing Group 2 and 3 samples.

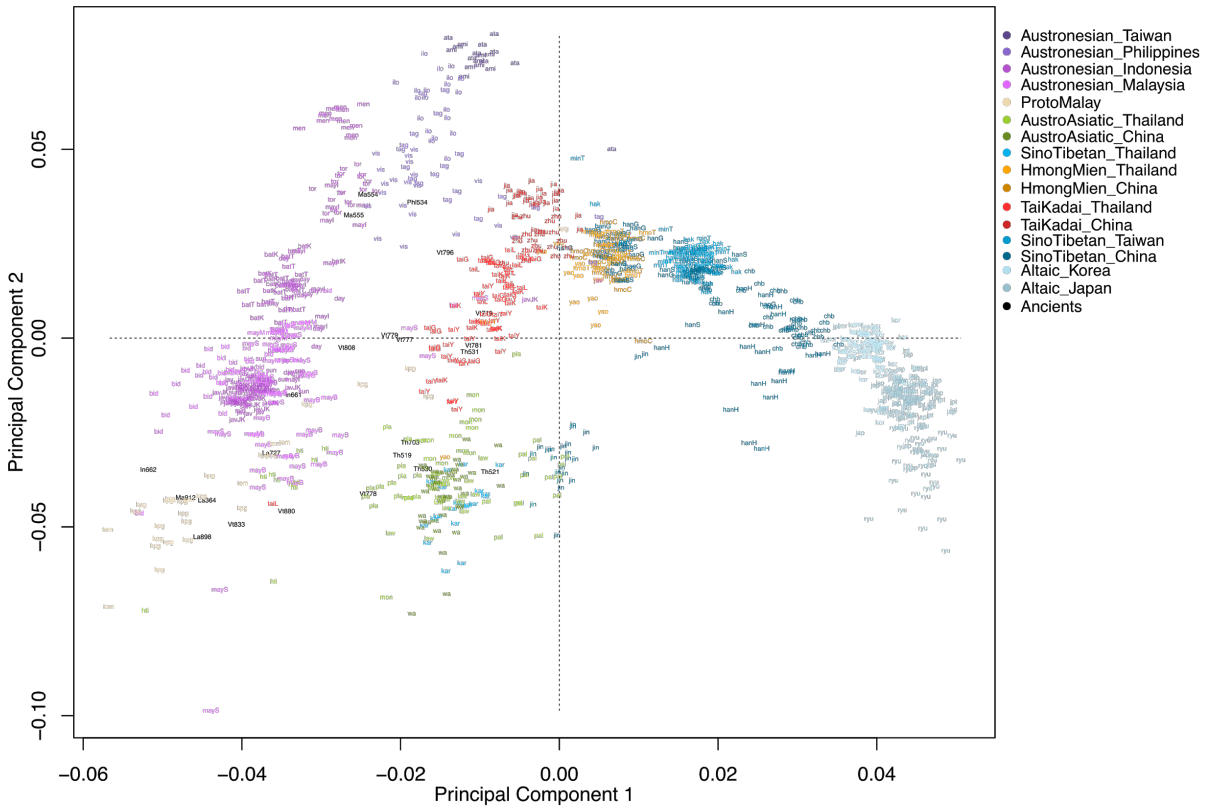


Figure S10. PCA from figure 1B with population labels.

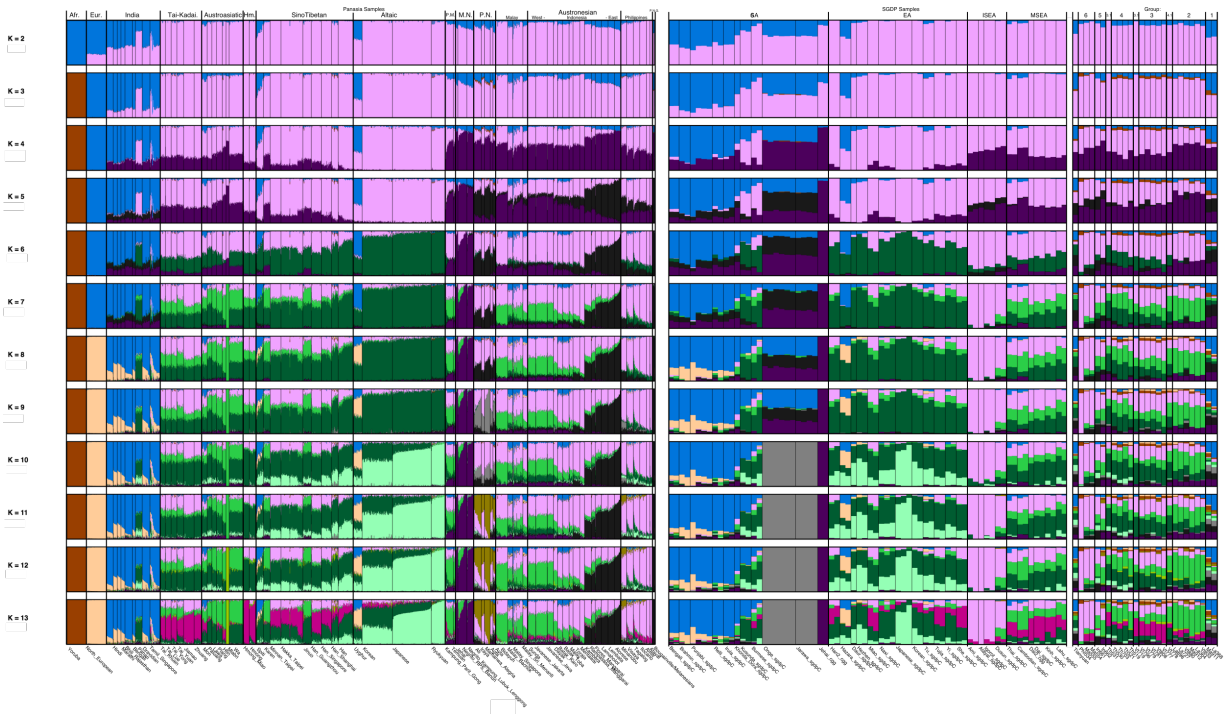


Figure S11. *fastNGSadmix plo*t of Pan-Asia samples, SGDP samples, Tianyuan, and ancient samples from this study, ranging from K=2 (top) to K=13 (bottom).

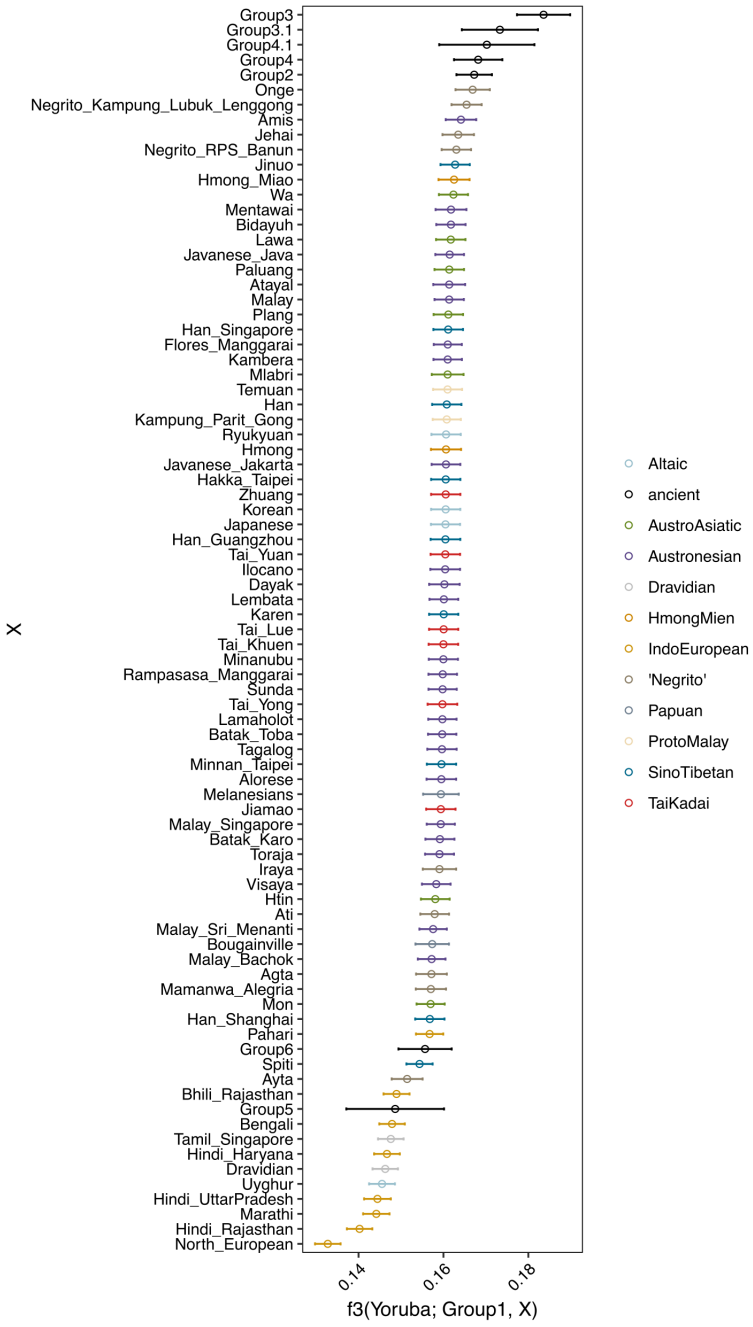


Figure S12. Values and standard errors of outgroup-f3 statistics using Yoruba as the outgroup, measuring shared drift between Group 1 individuals (La368, Ma911) and other individuals or populations.

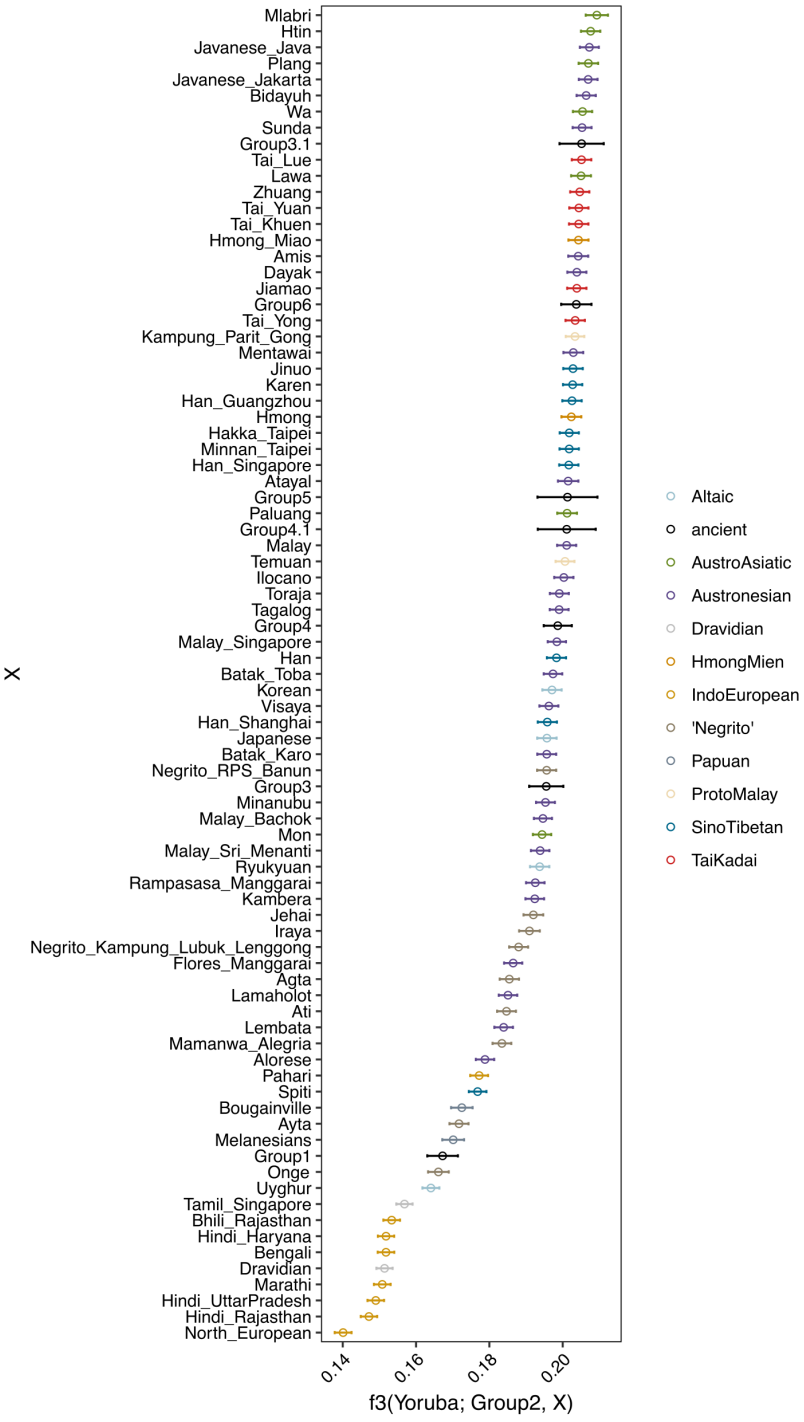
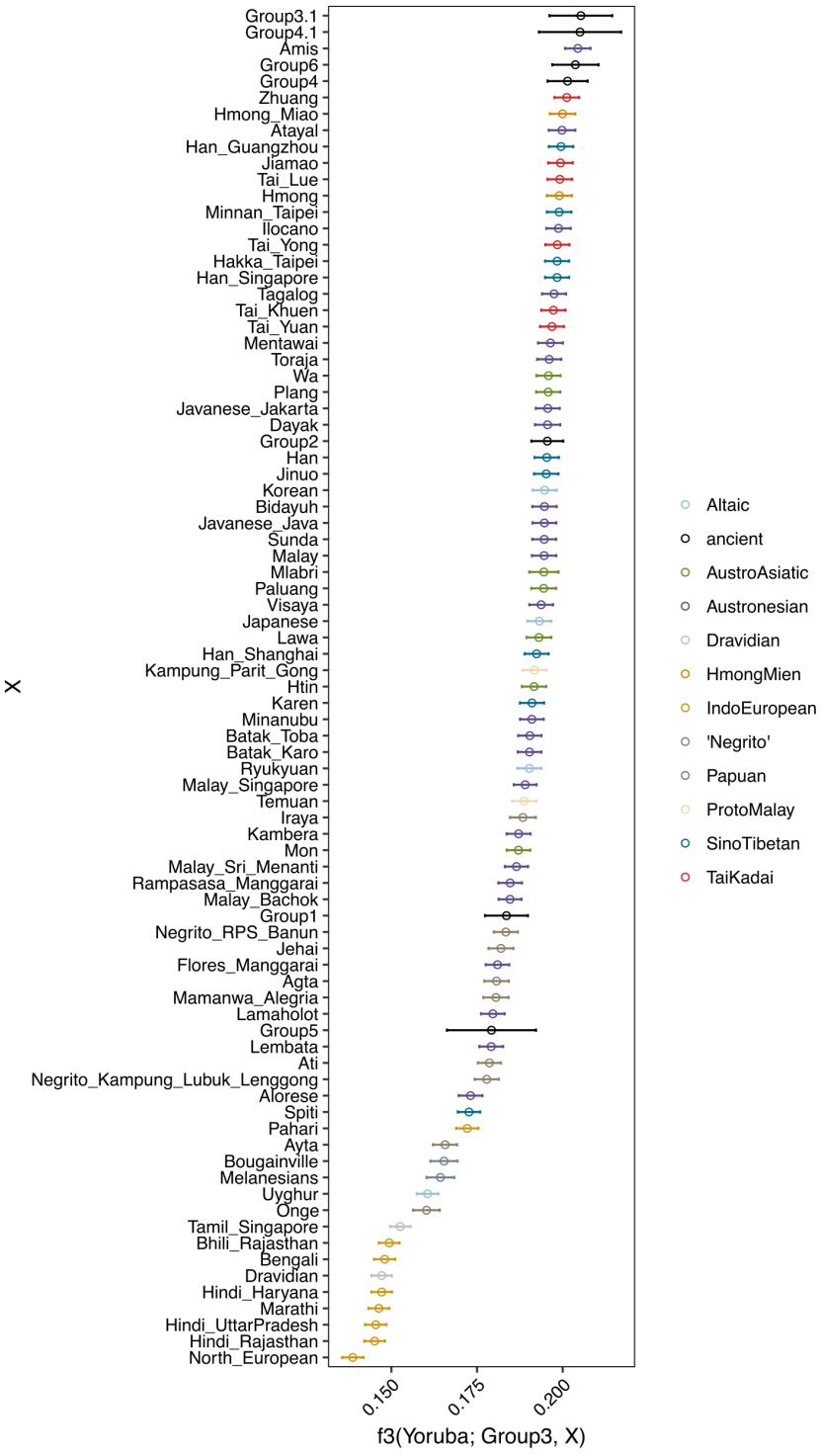


Figure S13. Values and standard errors of outgroup-f3 statistics using Yoruba as the outgroup, measuring shared drift between Group 2 individuals (La364, La727, La898, Ma912, Vt833, Vt880) and other individuals or populations.



643

644 **Figure S14.** Values and standard errors of outgroup-f3 statistics using Yoruba as the outgroup,
645 measuring shared drift between Group 3 individuals (Vt777, Vt779, Vt781, Vt796, Vt808) and
646 other individuals or populations.

647

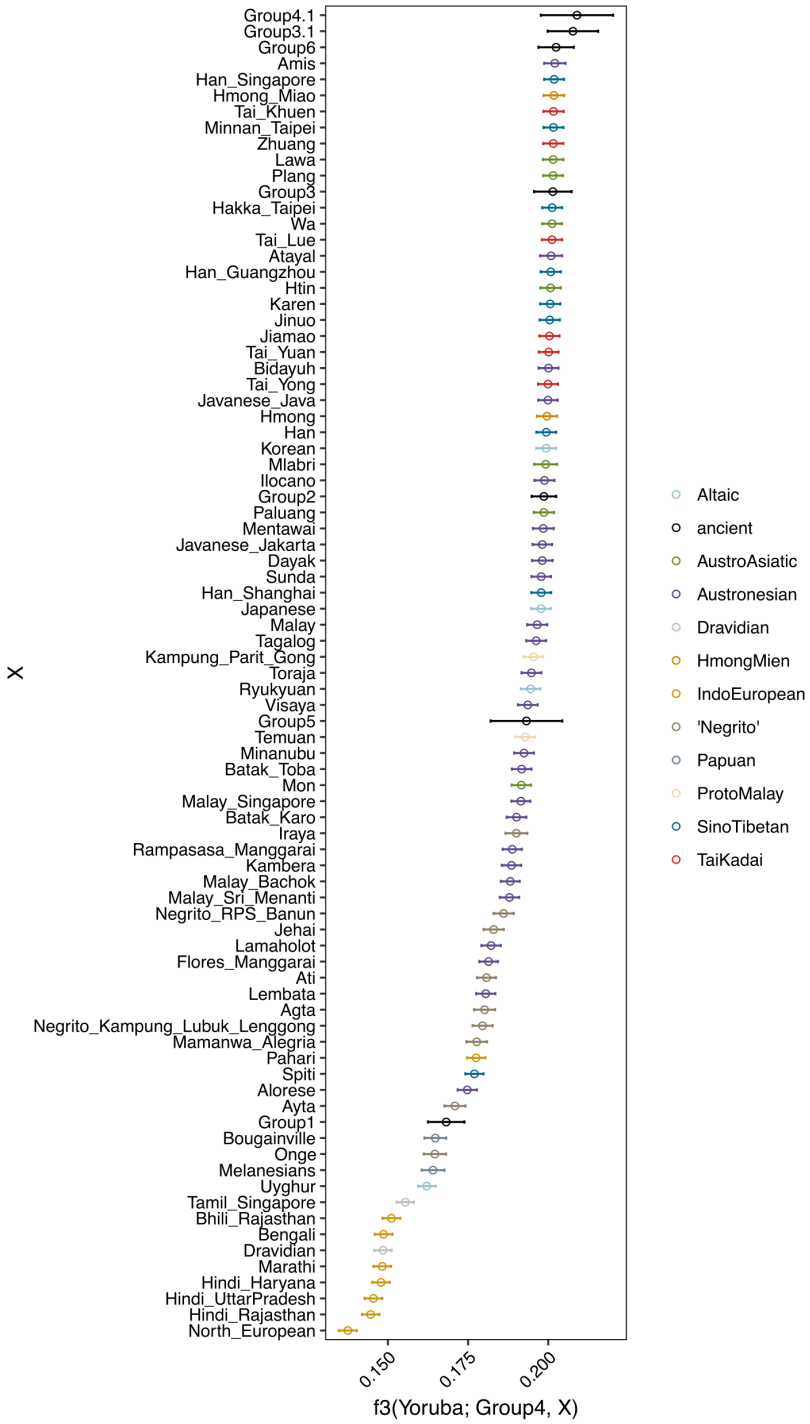
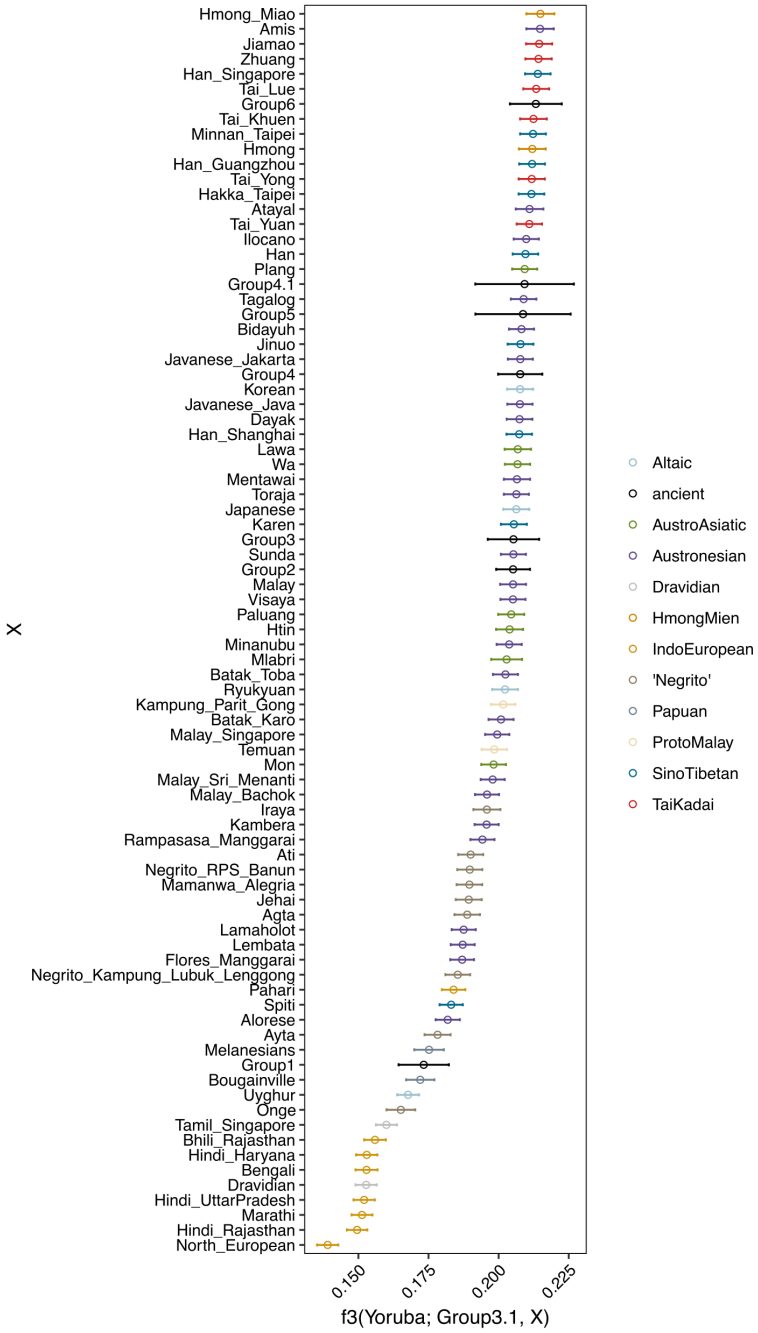


Figure S15. Values and standard errors of outgroup-f3 statistics using Yoruba as the outgroup, measuring shared drift between Group 4 individuals (Th519, Th521, Th530, Th703) and other individuals or populations.



653

654 **Figure S16.** Values and standard errors of outgroup-f3 statistics using Yoruba as the outgroup,
655 measuring shared drift between Group 3.1 individuals (Th531, Vt719) and other individuals or
656 populations.

657

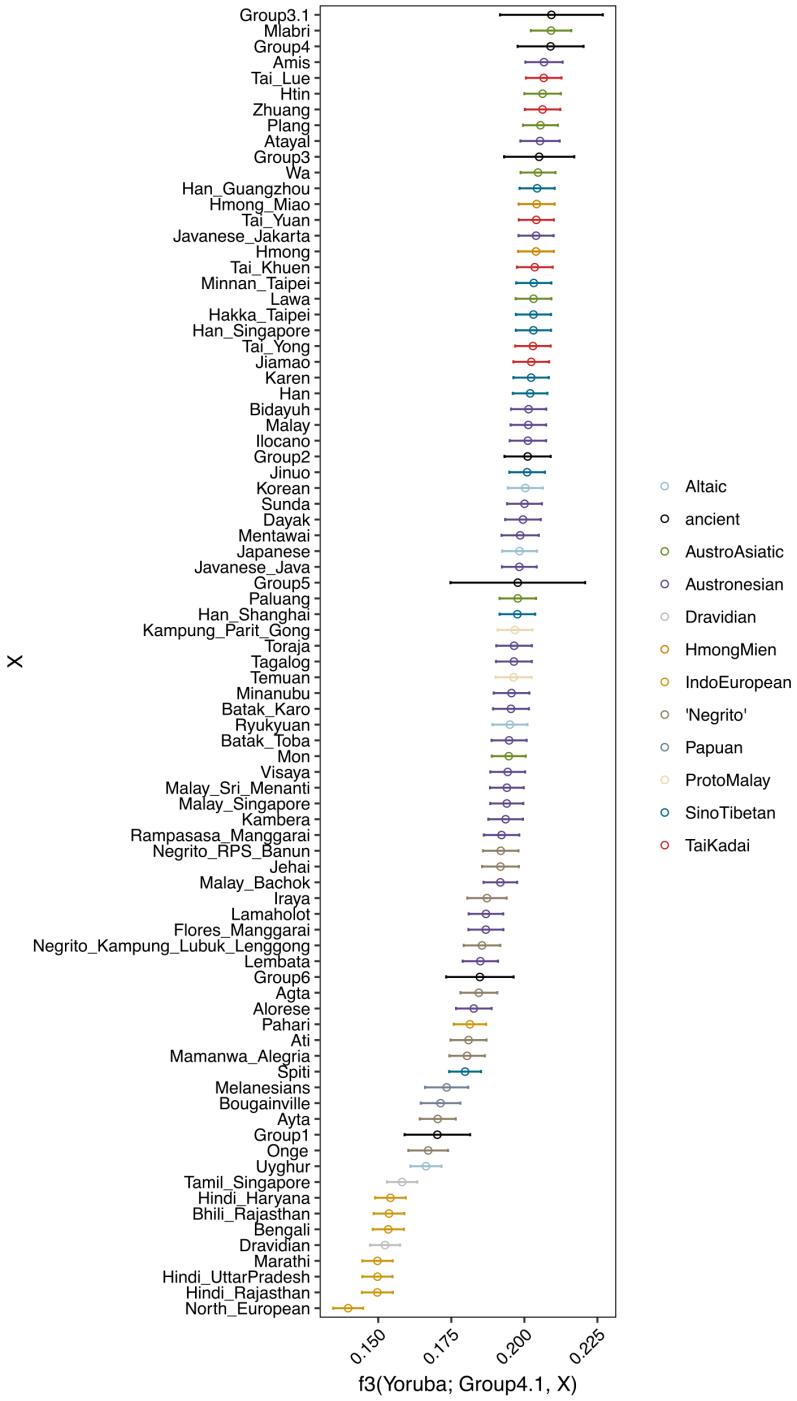


Figure S17. Values and standard errors of outgroup-f3 statistics using Yoruba as the outgroup, measuring shared drift between the Group 4.1 individual (Vt778) and other individuals or populations.

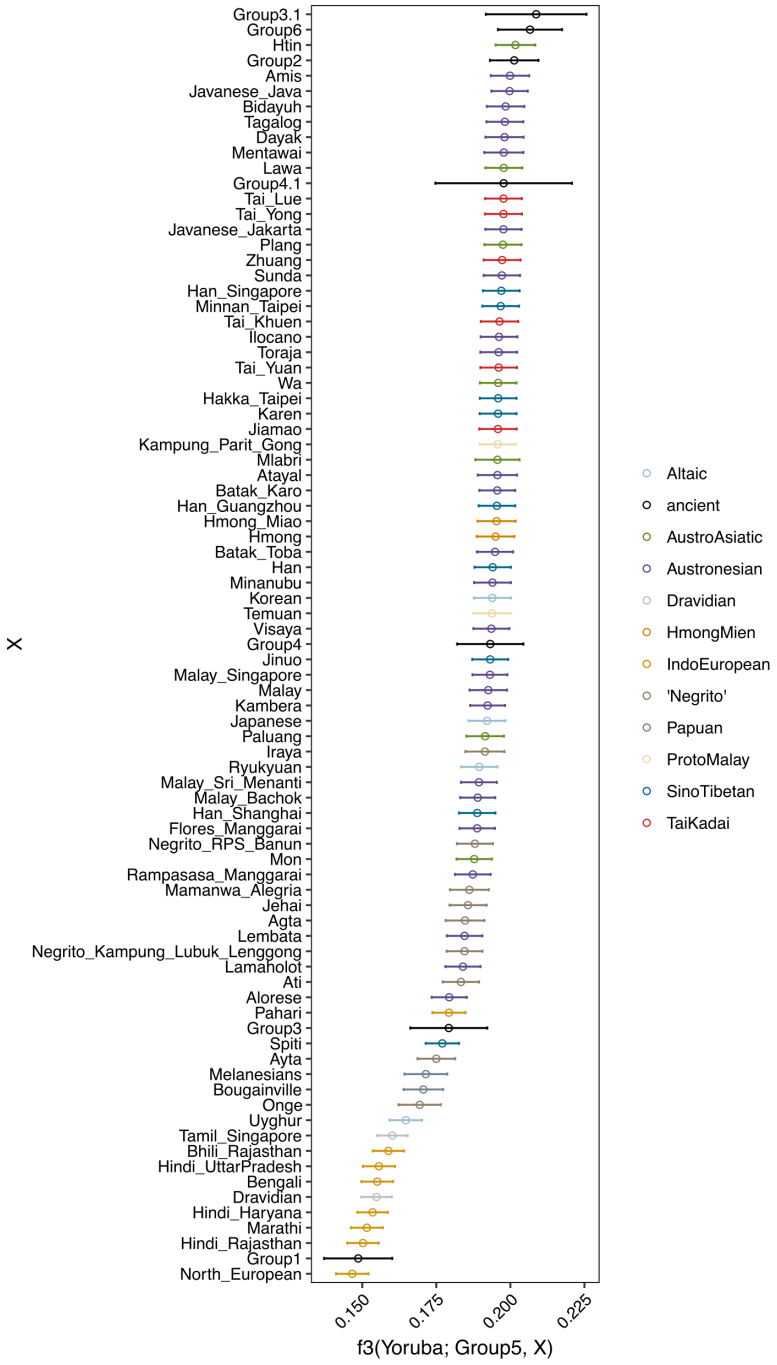


Figure S18. Values and standard errors of outgroup-f3 statistics using Yoruba as the outgroup, measuring shared drift between Group 5 individuals (In661, In662) and other individuals or populations.

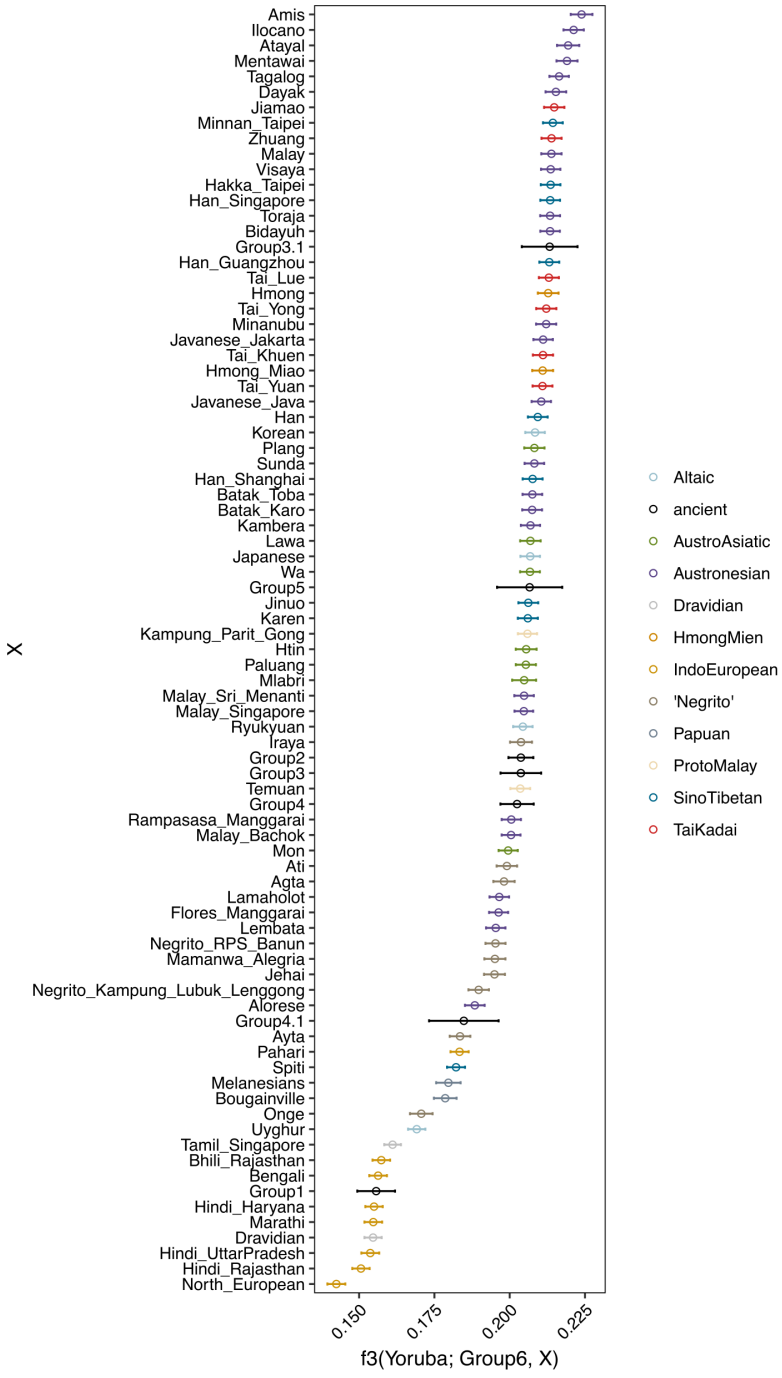
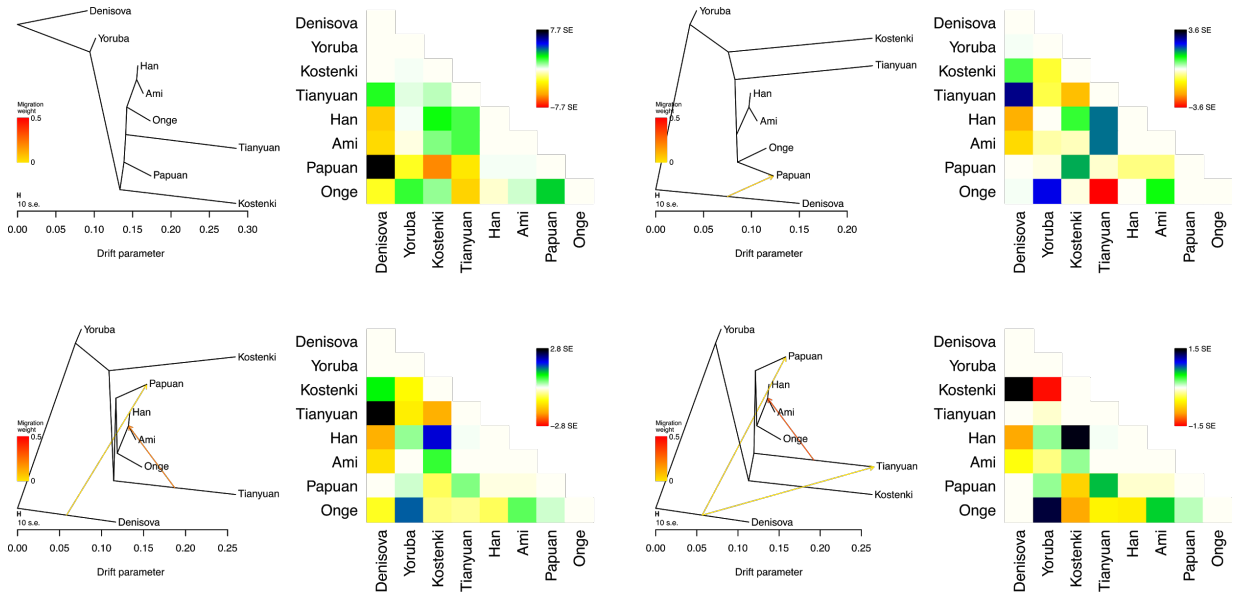


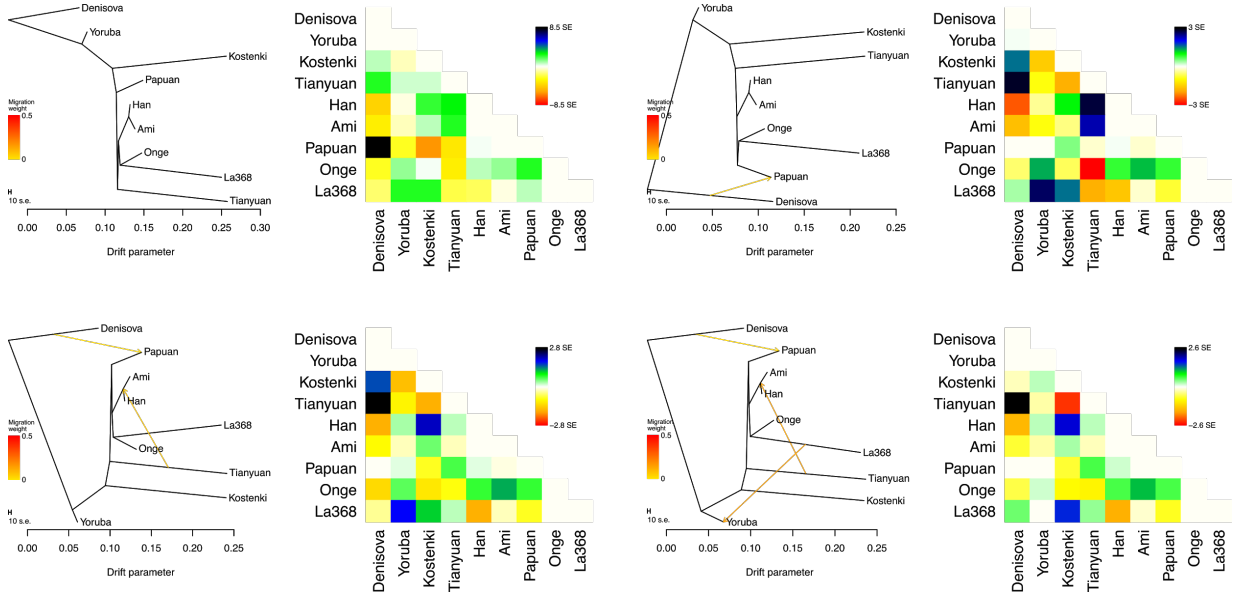
Figure S19. Values and standard errors of outgroup-f3 statistics using Yoruba as the outgroup, measuring shared drift between Group 6 individuals (Ma554, Ma555, Phl534) and other individuals or populations.



676

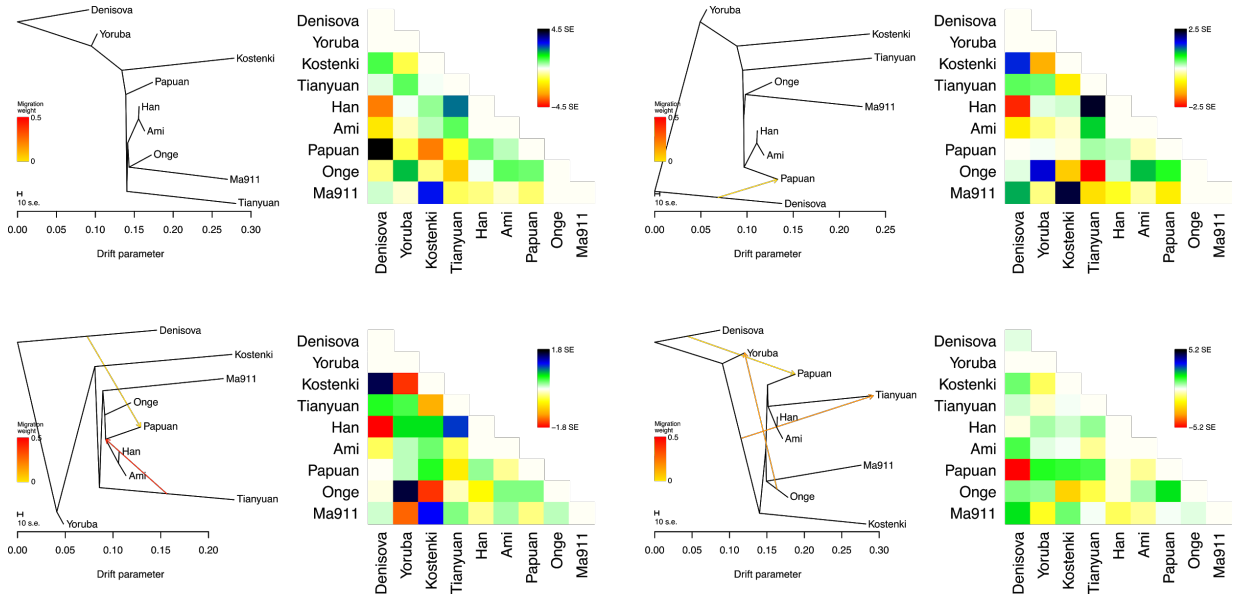
677 **Figure S20.** *TreeMix* admixture graphs modelling relationships among the "base populations":
678 Onge, Tianyuan Papuan and ancestral populations (Kostenki, Yoruba, Denisova) (422,211
679 SNPs).

680



681

682 **Figure S21.** *TreeMix* admixture graphs modelling relationships between the "base
683 populations" and La368 (Group 1) (189,694 SNPs).

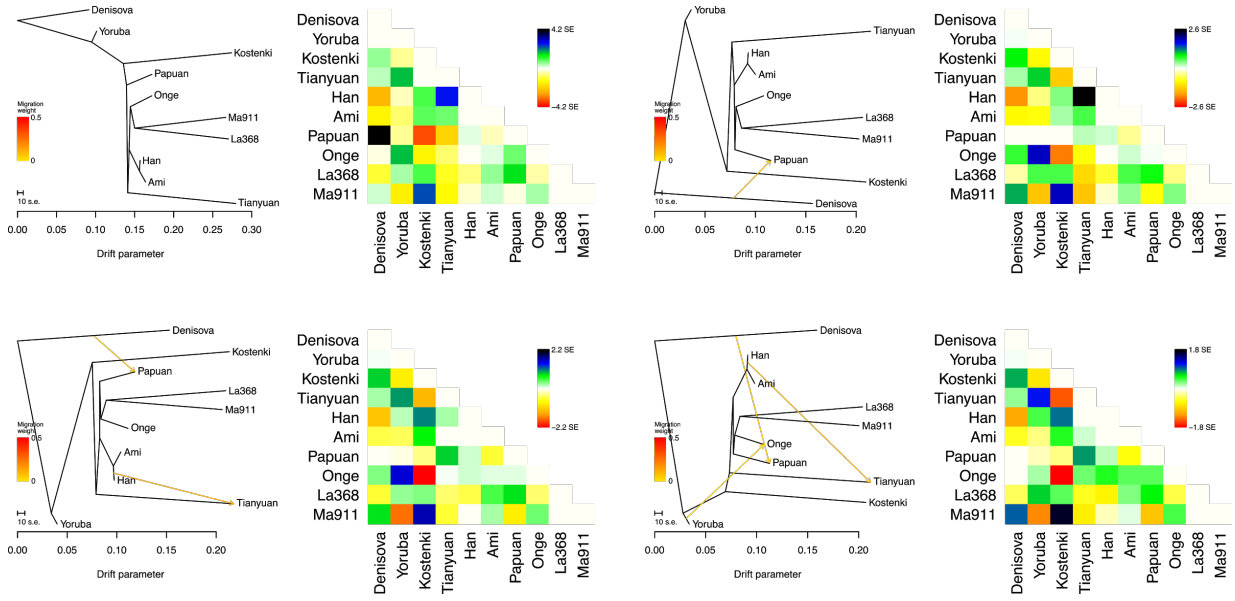


684

685 **Figure S22.** *TreeMix* admixture graphs modelling relationships between the "base

686 populations" and Ma911 (Group 1) (47359 SNPs).

687

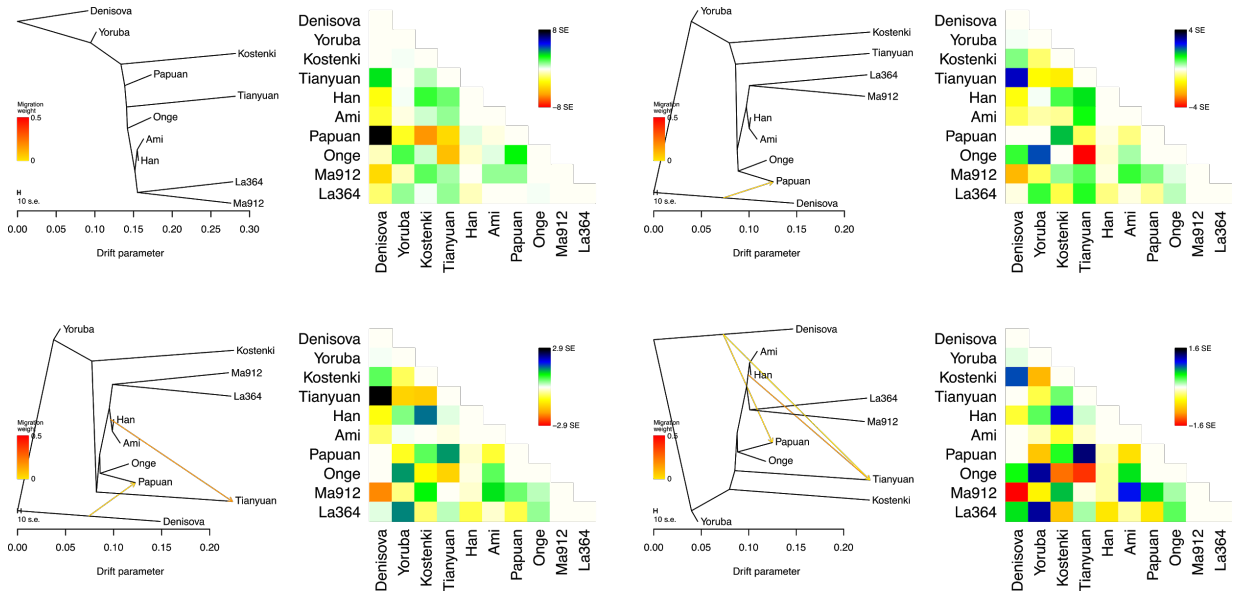


688

689 **Figure S23.** *TreeMix* admixture graphs modelling relationships between the "base

690 populations" and both La368 and Ma911 (Group 1) (24,324 SNPs).

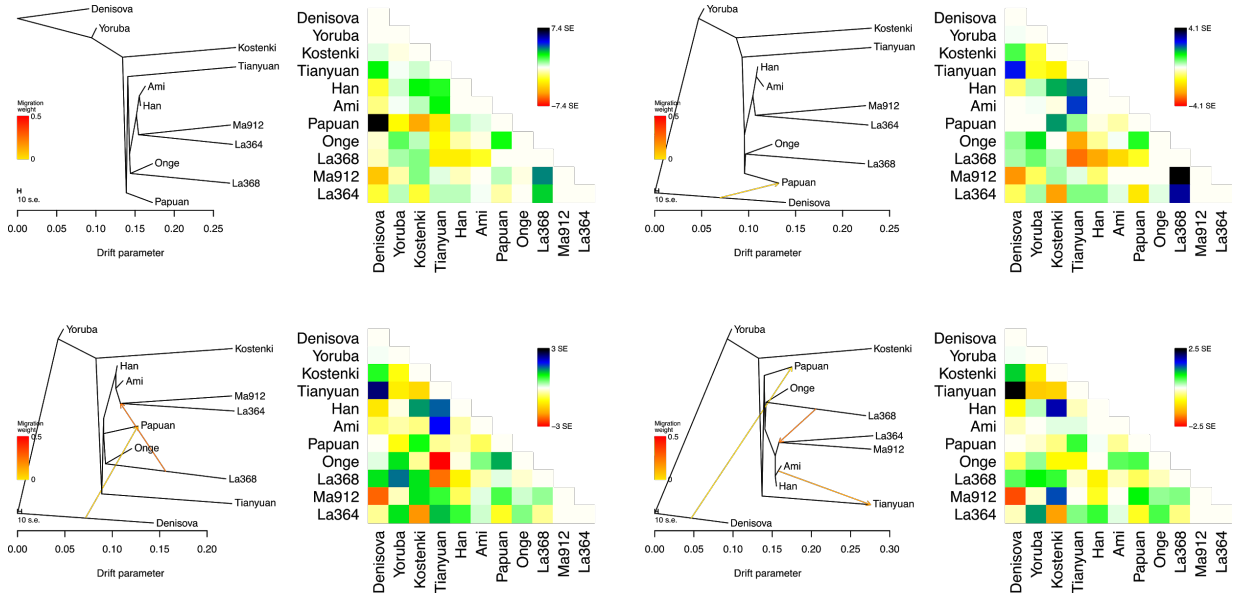
691



692

693 **Figure S24.** *TreeMix* admixture graphs modelling relationships between the "base
694 populations" and the two Group 2 individuals (La364 and Ma912) (274,352 SNPs).

695



696

697 **Figure S25.** *TreeMix* admixture graphs modelling relationships between the "base
698 populations" and the Laotian sample from Group 1 (La368) and the two Group 2 samples
699 (La364 and Ma912) (129,039 SNPs).

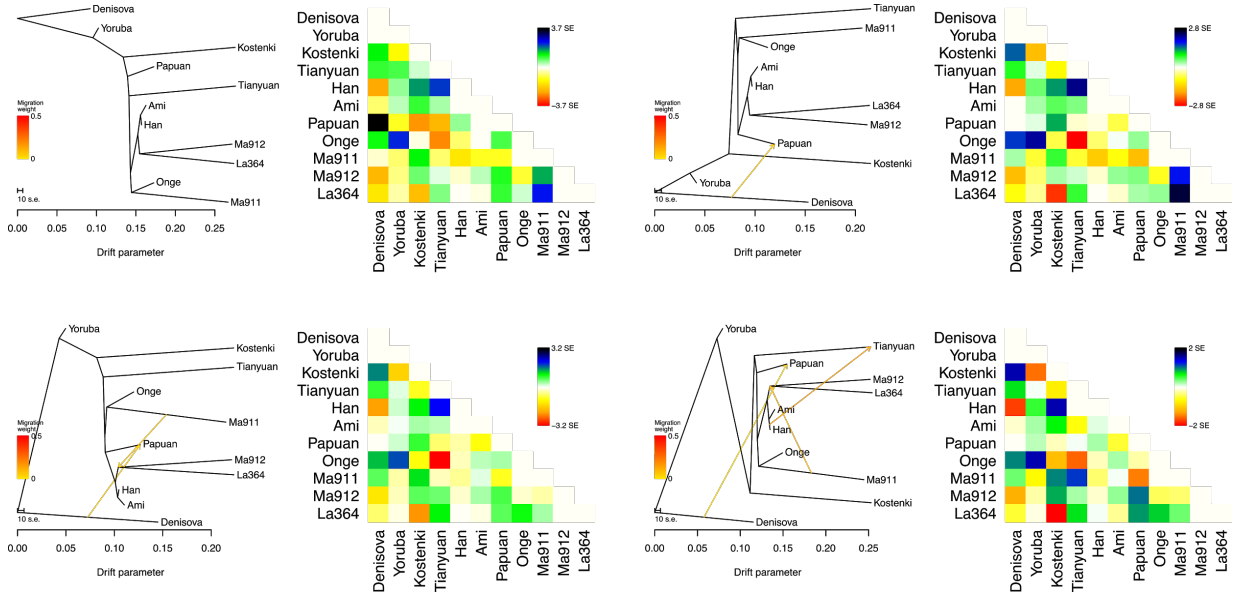


Figure S26. *TreeMix* admixture graphs modelling relationships between the "base populations" and the Malaysian sample from Group 1 (Ma911) and the two Group 2 samples (La364 and Ma912) (32,465 SNPs).

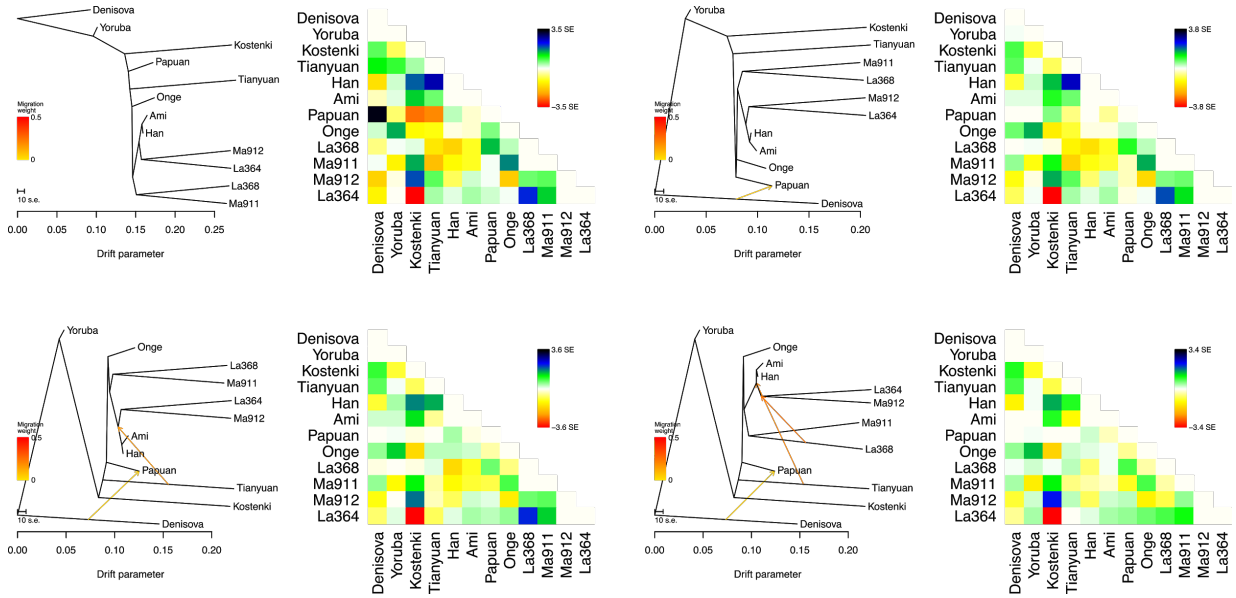
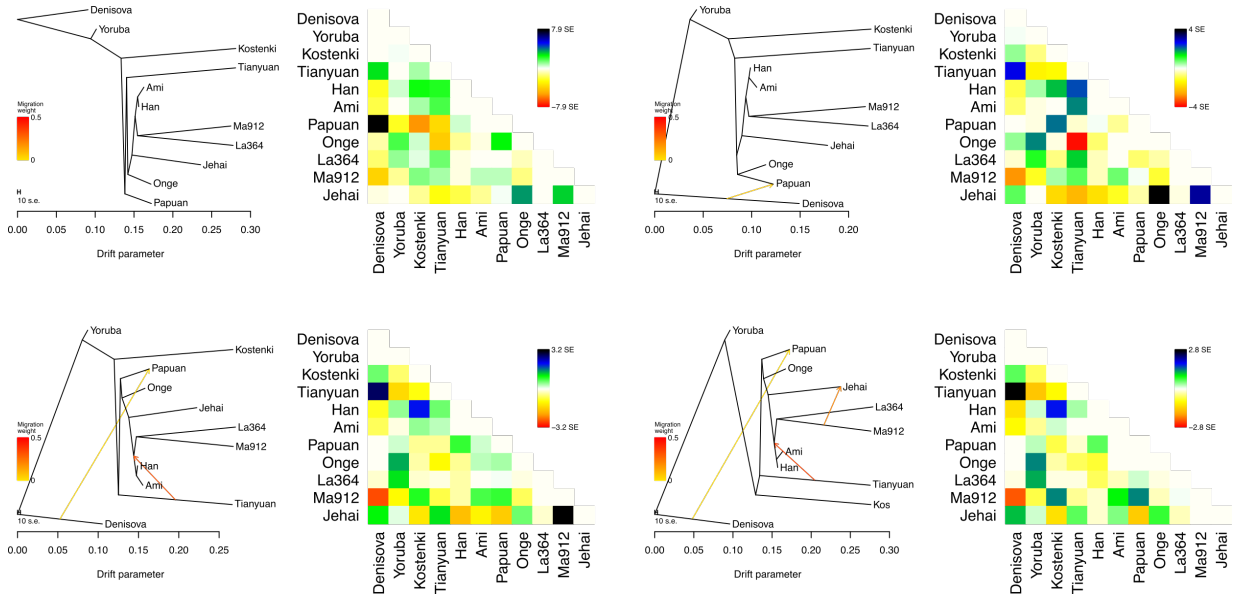


Figure S27. *TreeMix* admixture graphs modelling relationships between the "base populations" and the two samples from Group 1 (La368 and Ma911) and the two samples from Group 2 (La364 and Ma912) (17,286 SNPs).



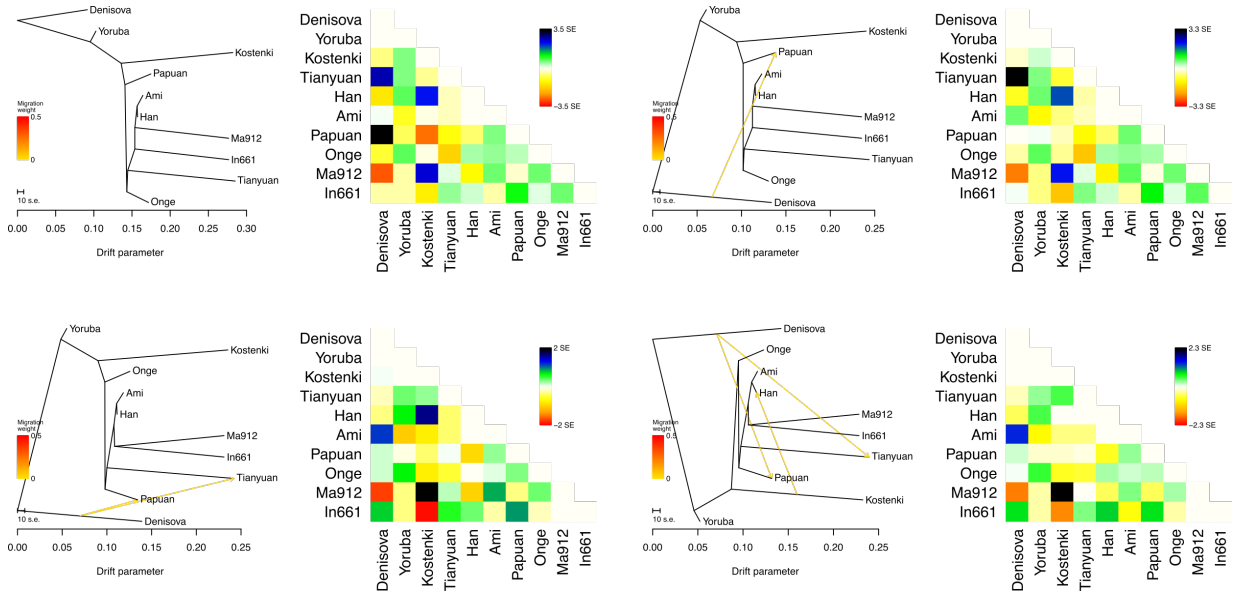
710

711 **Figure S28.** *TreeMix* admixture graphs modelling relationships between the "base

712 populations", the two Group 2 samples (La364 and Ma912) and Jehai (275,053 SNPs).

713

714



715

716 **Figure S29.** *TreeMix* admixture graphs modelling relationships between the "base

717 populations" and the Malaysian samples from Group 2 (Ma912) and an Indonesian Group 5

718 sample (In661) (18,594 SNPs)

719

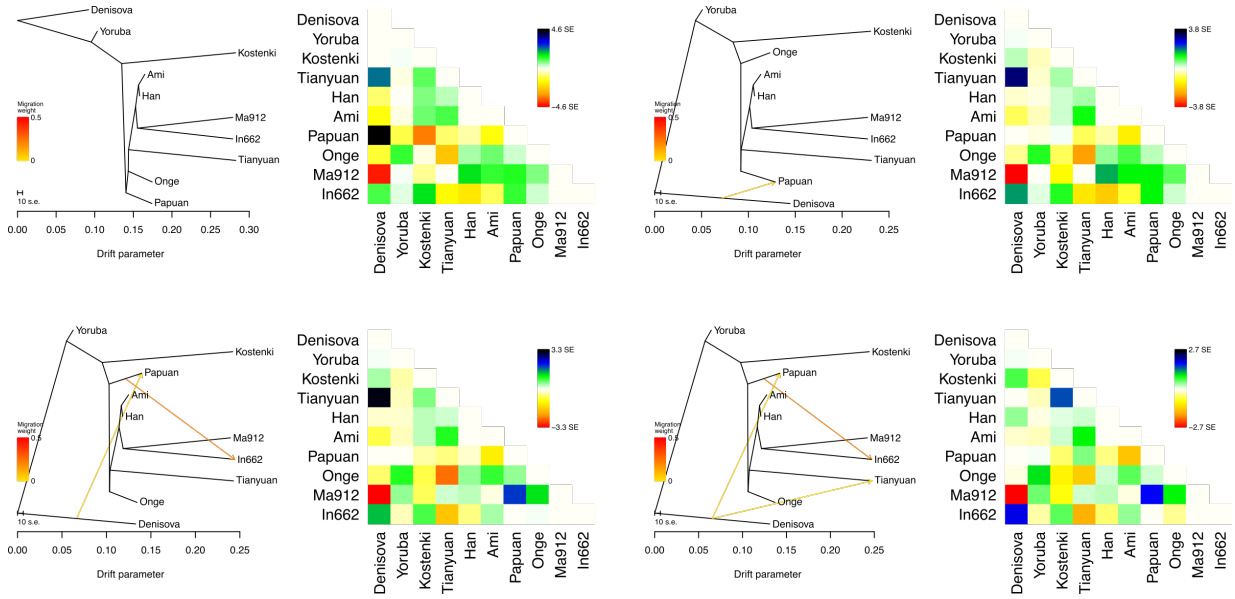


Figure S30. *TreeMix* admixture graphs modelling relationships between the "base populations" and the Malaysian samples from Group 2 (Ma912) and an Indonesian Group 5 samples (In662) (33,895 SNPs)

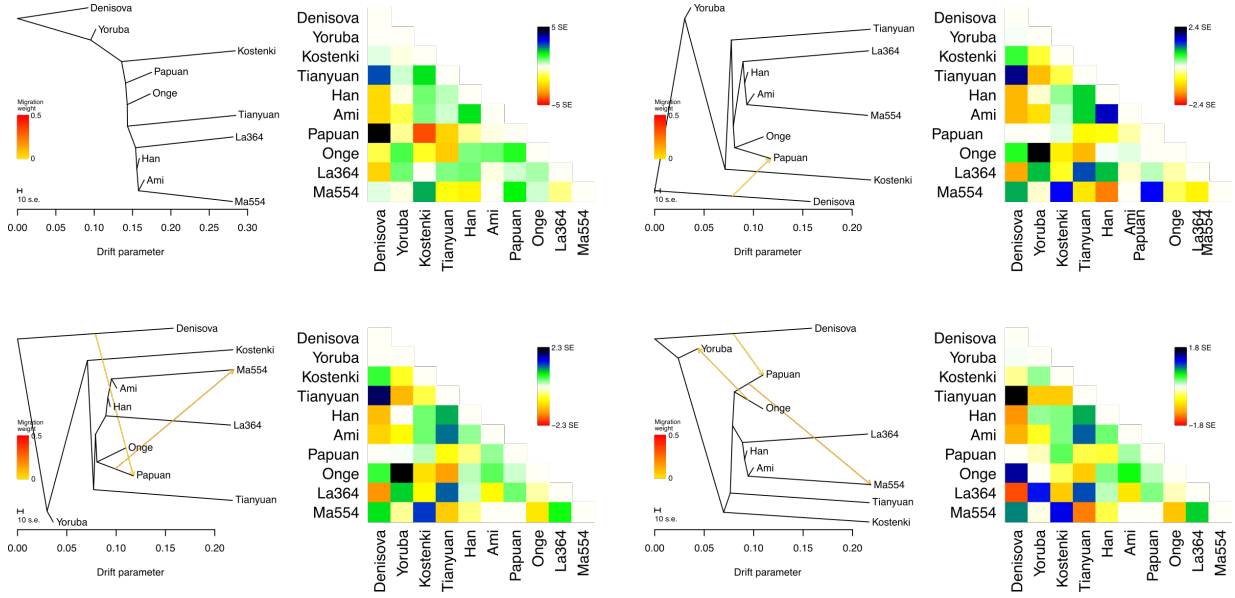
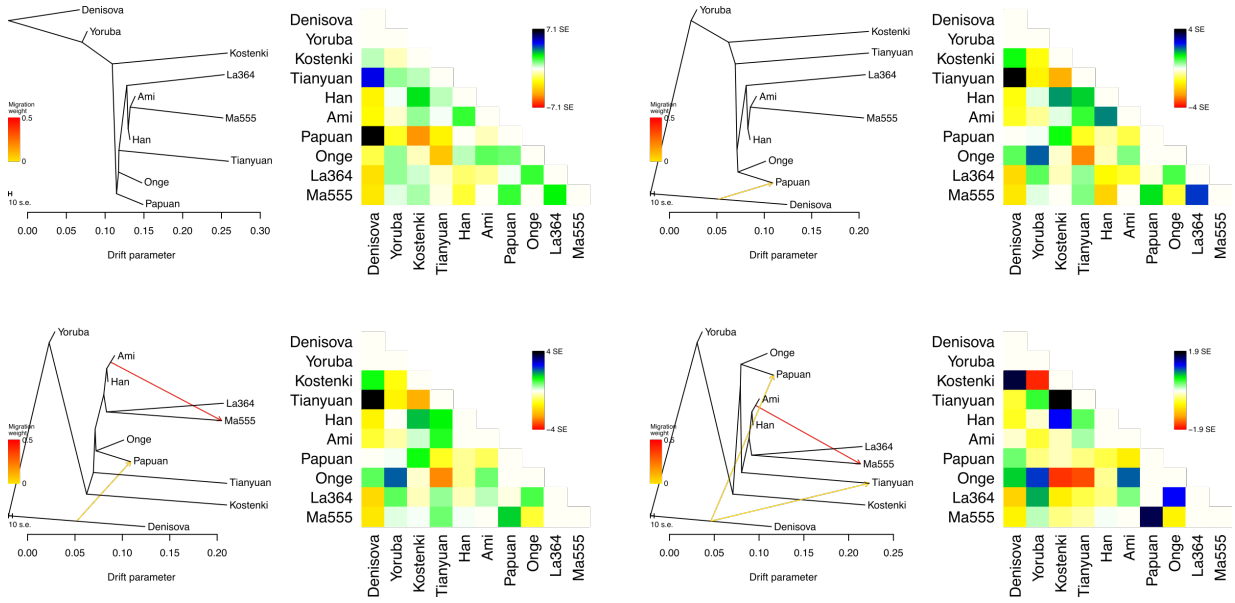


Figure S31. *TreeMix* admixture graphs modelling relationships between the "base populations" and the Laotian sample from Group 2 (La364) and a Group 6 Sample from Borneo (Ma554) (54,144 SNPs)



730

731 **Figure S32.** *TreeMix* admixture graphs modelling relationships between the "base

732 populations" and the Laotian sample from Group 2 (La364) and a Group 6 Sample from

733 Borneo (Ma555) (144,272 SNPs)

734

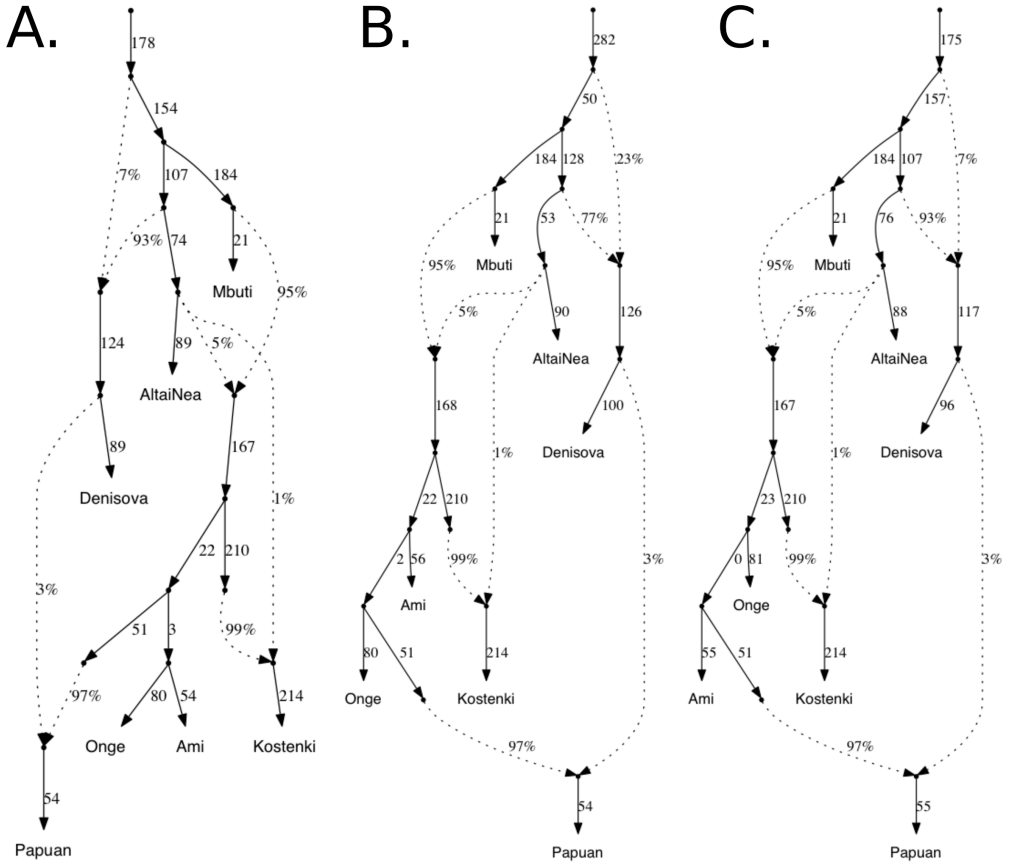


Figure S33. Without including Tianyuan, we are unable to resolve the Papuan-Onge-Ami trifurcation: all three possibilities for the graph give good fits: A) worst-fitting Z = -1.774; B) worst-fitting Z = -2.353; C) worst-fitting Z = -2.450.

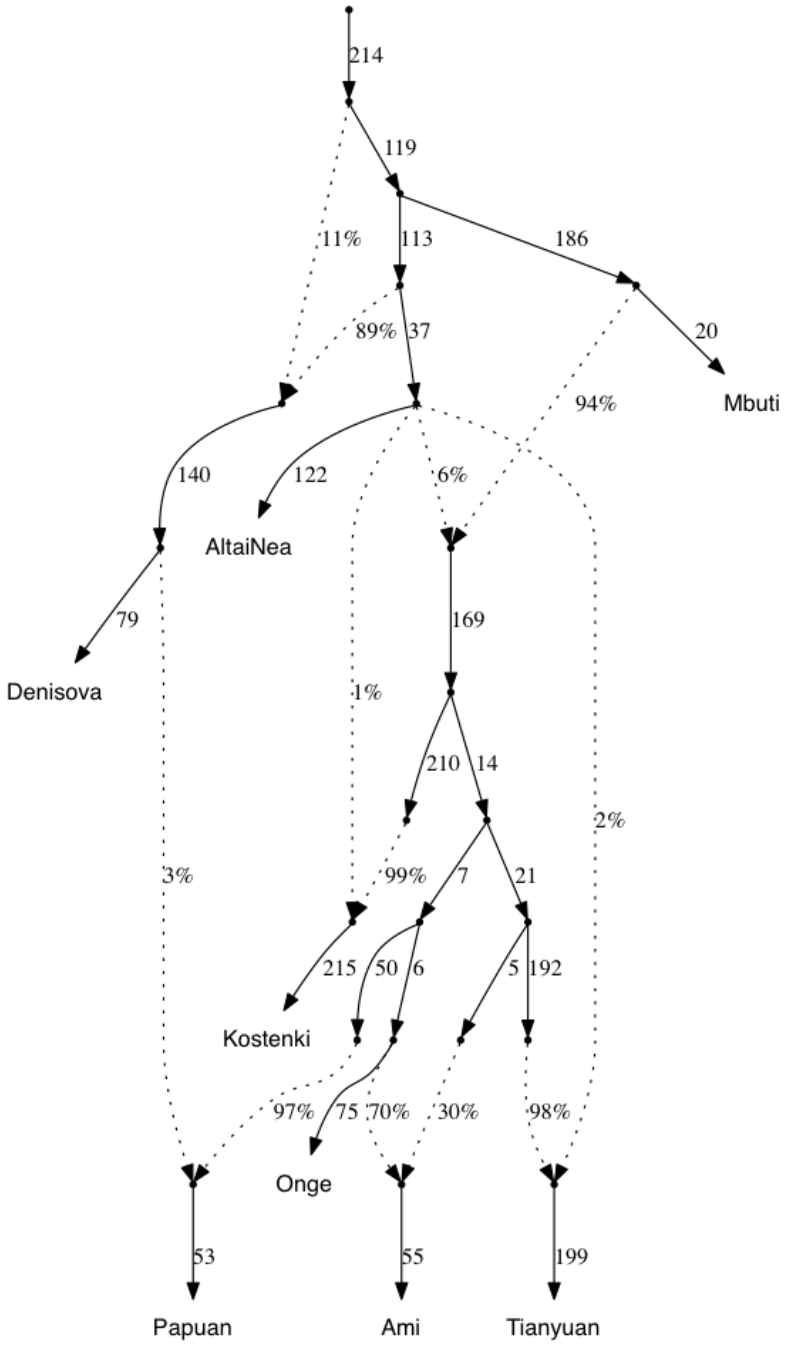
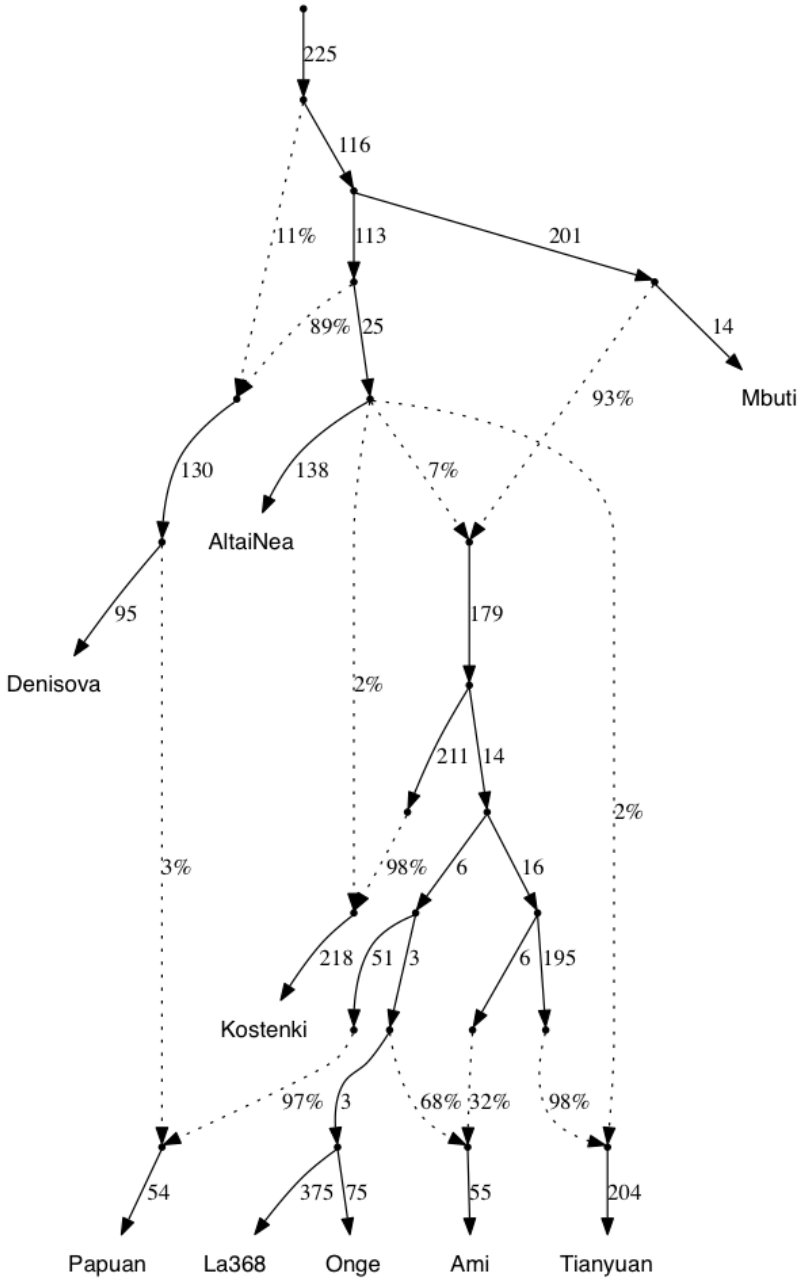


Figure S34. Including Tianyuan, we find that the best fit is found when modeling Ami as a mixture of a sister group to Onge and a sister group to Tianyuan (worst-fitting Z = -3.564).



745

746 **Figure S35.** La368 is best modeled as a sister group to Onge (worst-fitting Z = 3.372).

747

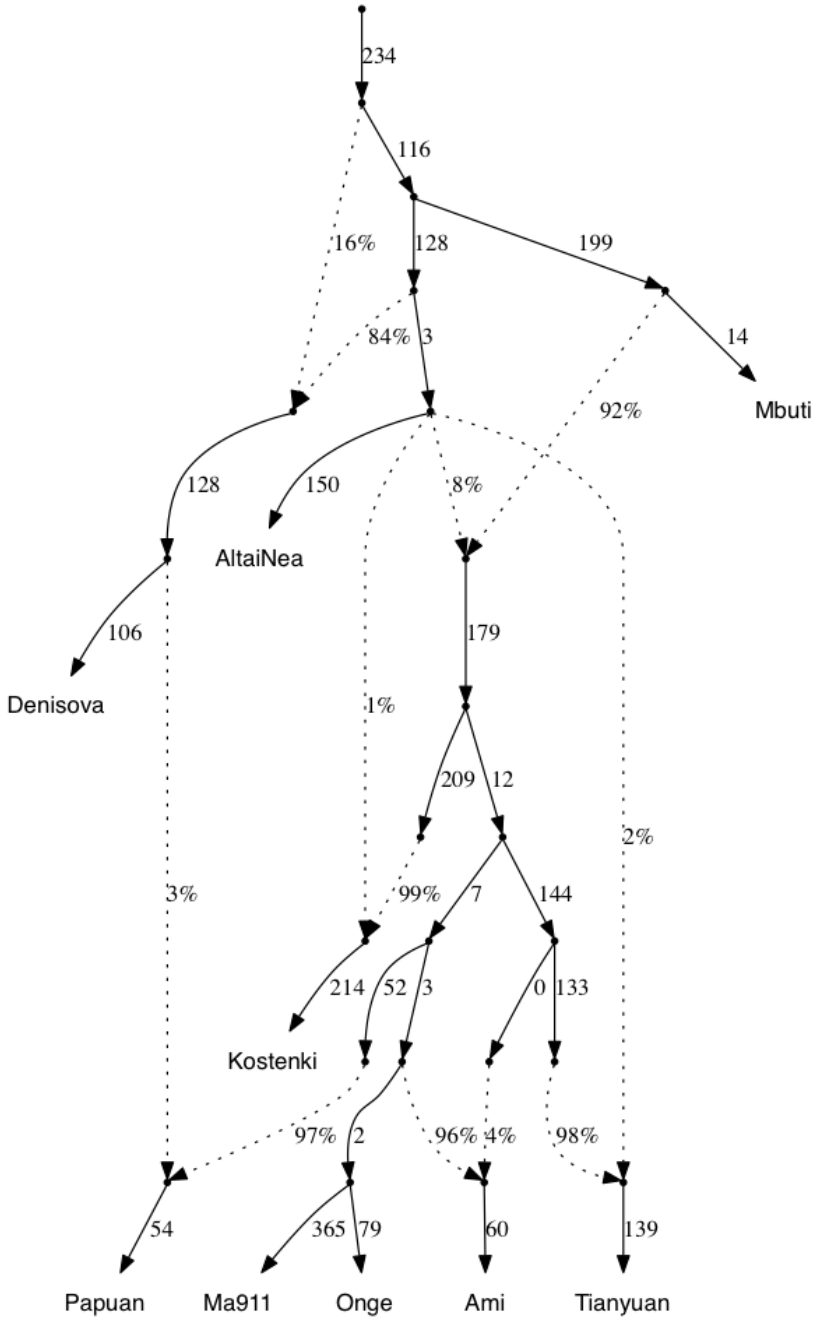


Figure S36. Ma911 is best modeled as a sister group to Onge (worst-fitting Z = 3.803).

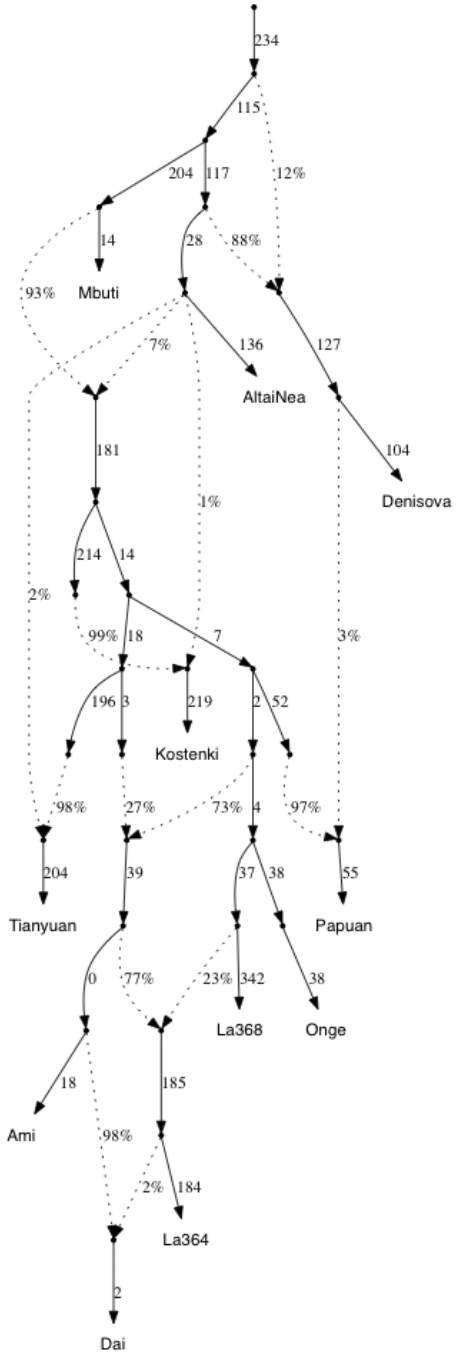


Figure S37. La364 is best modeled as a mixture of a sister group to La368 (Group 1) and an East Asian component (related to Ami). In turn, present-day Dai is best modeled as a mixture of a sister group to La364 (Group 2) and an additional East Asian component (worst-fitting Z = 3.66).

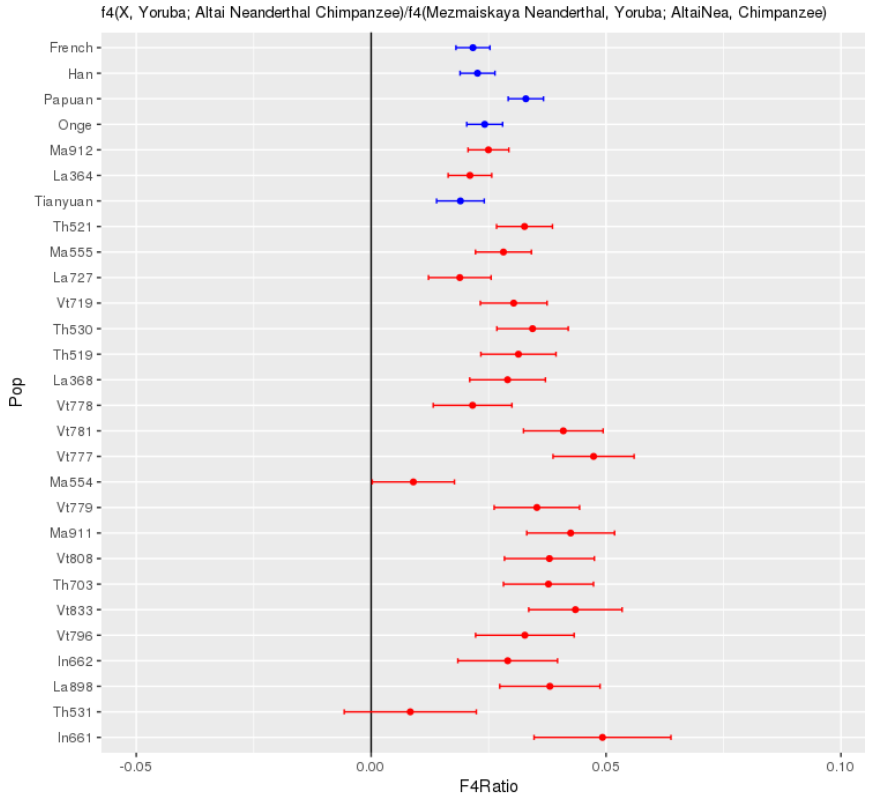
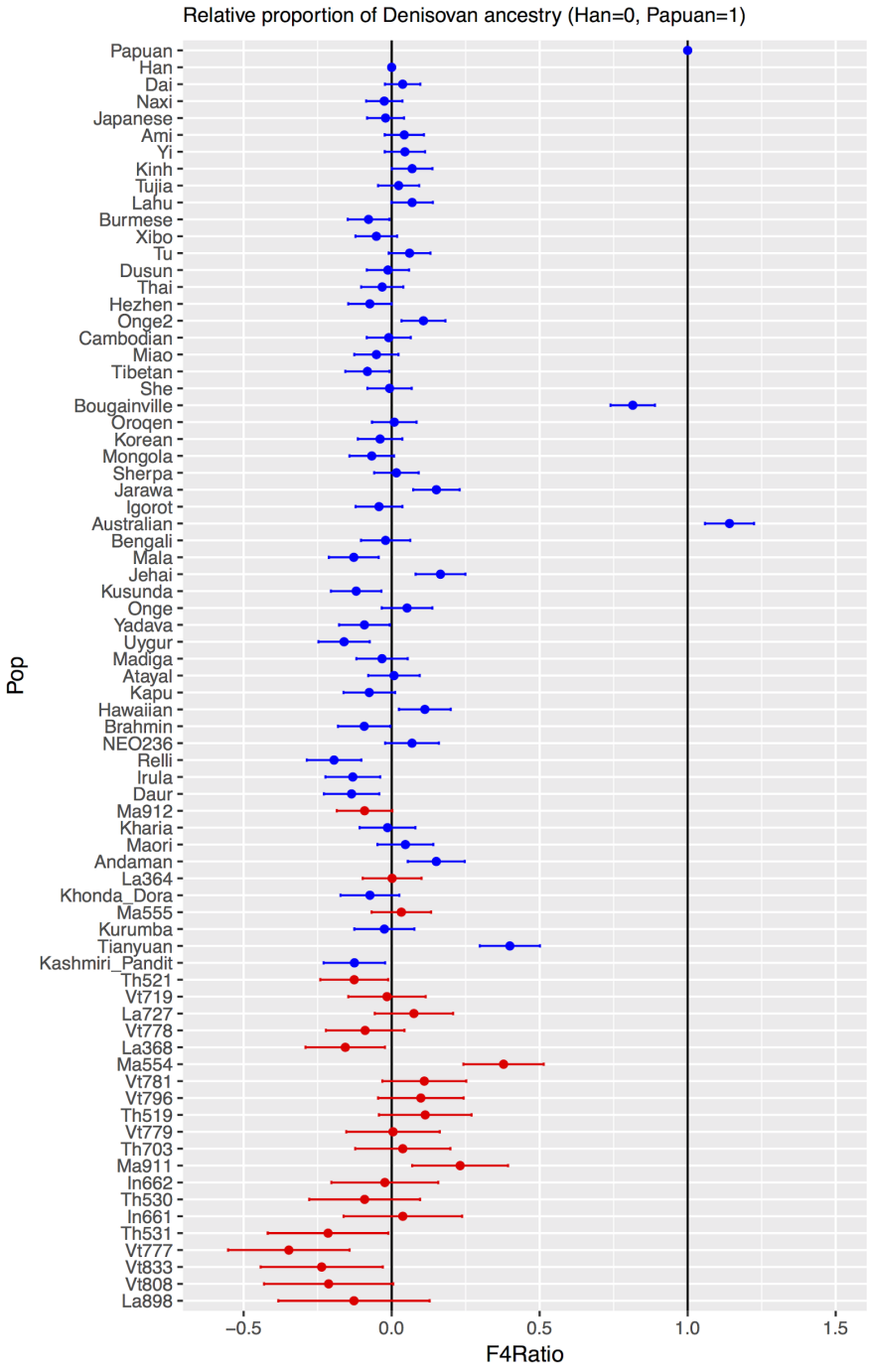


Figure S38. We computed an F4 ratio of the form f4(X, Yoruba; Altai Neanderthal Chimpanzee)/f4(Mezmaiskaya Neanderthal, Yoruba; AltaiNea, Chimpanzee) (Green et al. 2010). This serves to measure the proportion of Neanderthal ancestry in population X, using an African population (Yoruba) as a baseline non-admixed population, but this statistic may be confounded by extra Denisovan introgression that may be present in population X. The red-colored individuals are the samples from this study, and the individuals are ordered based on the size of the standard error.



769

770 **Figure S39.** We computed an F4 ratio of the form f4(Denisova, Mbuti; X, Han)/f4(Denisova,
771 Mbuti; Papuan, Han). This serves to measure the amount of Denisovan ancestry in population
772 X as a relative proportion of that ancestry found in Papuans, using Han as a baseline non-
773 admixed population. The red-colored individuals are the samples from this study, and the
774 individuals are ordered based on the size of the standard error.

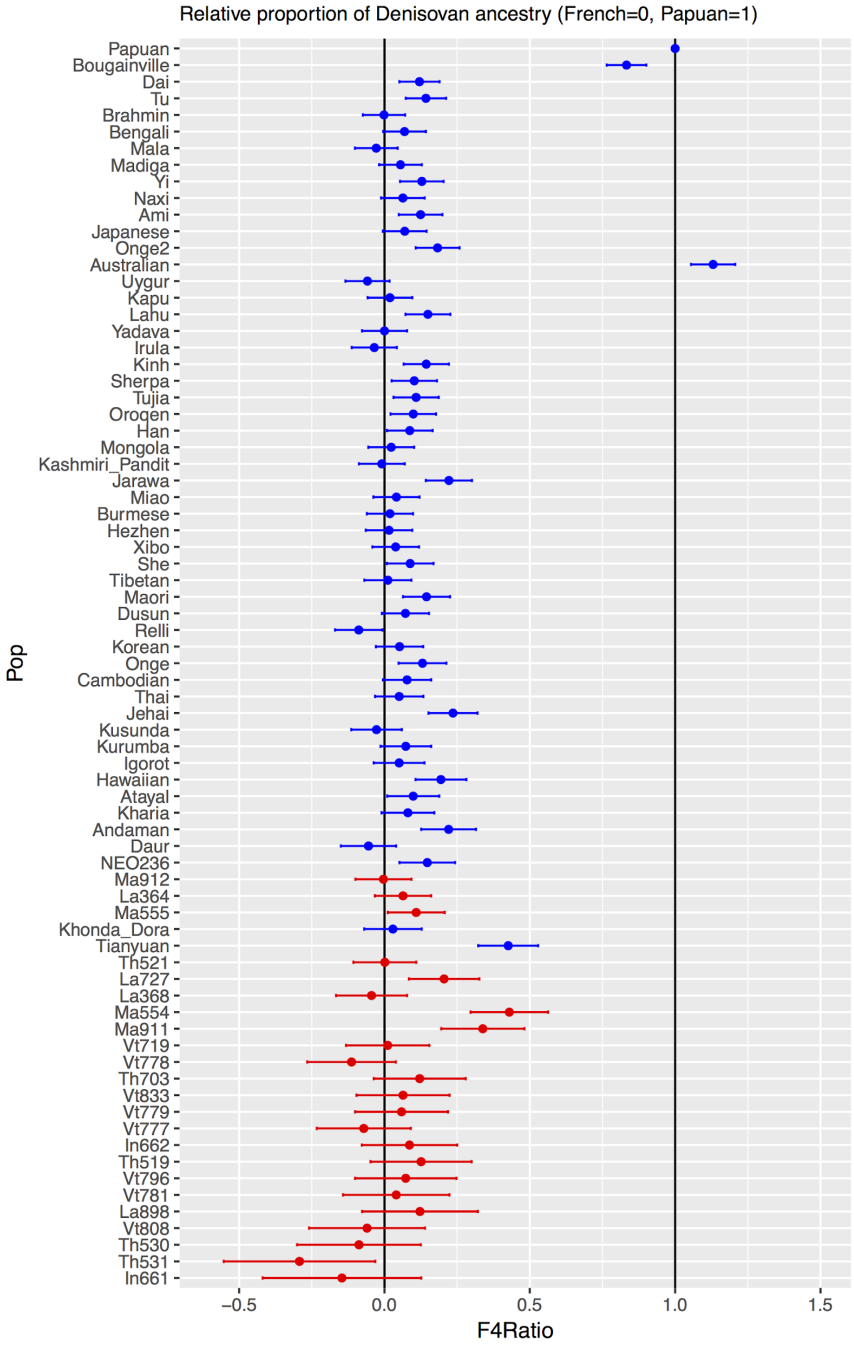
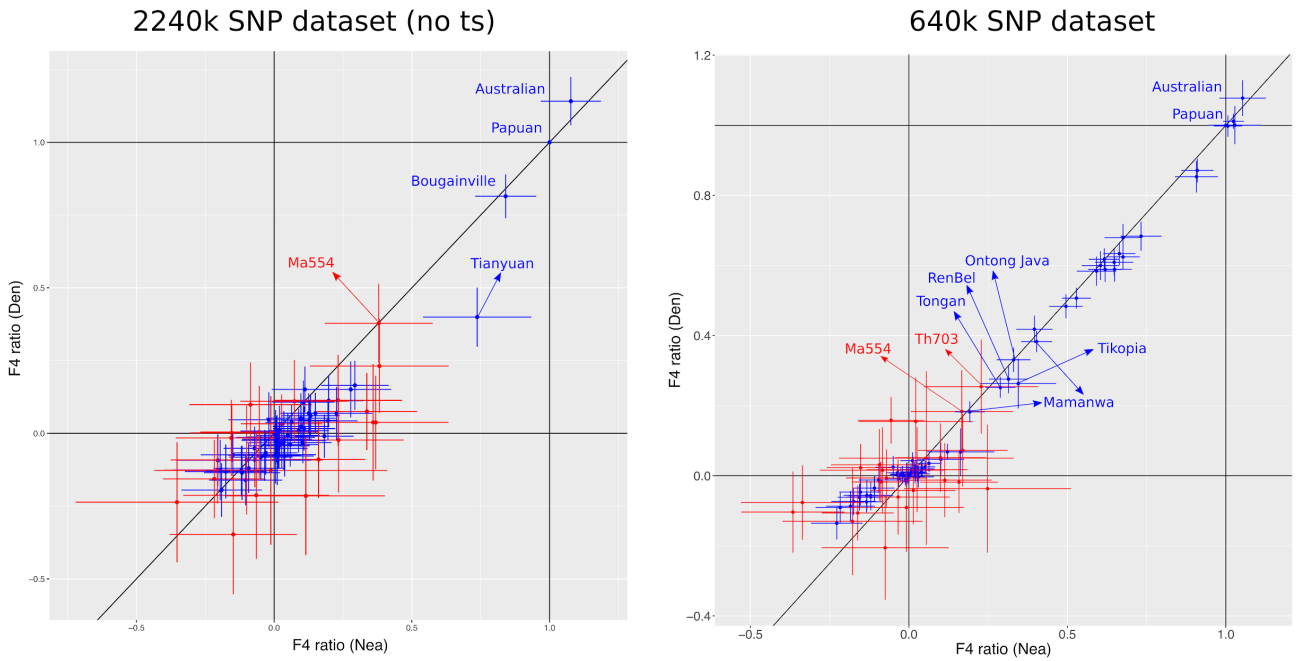


Figure S40. We computed an F4 ratio of the form f4(Denisova, Mbuti; X, French)/f4(Denisova, Mbuti; Papuan, French). This serves to measure the amount of Denisovan ancestry in population X as a relative proportion of that ancestry found in Papuans, using French as a baseline non-admixed population. The red-colored individuals are the samples from this study, and the individuals are ordered based on the size of the standard error.



781

782 **Figure S41.** We compared two F4 ratios. X-axis: f4(Denisova, Mbuti; X, Han)/f4(Denisova,
783 Mbuti; Papuan, Han). Y-axis: f4(Altai Neanderthal, Mbuti; X, Han)/f4(Denisova, Mbuti;
784 Papuan, Han). The X-axis ratio serves to measure the amount of Denisovan ancestry in
785 population X as a relative proportion of that ancestry found in Papuans, using Han as a baseline
786 non-admixed population. The Y-axis ratio serves to measure the amount of Altai Neanderthal
787 ancestry in population X as a relative proportion of that ancestry found in Papuans, using Han
788 as a baseline non-admixed population, but is confounded by Denisovan ancestry found in
789 Southeast Asian and Oceanian populations. When computing these two ratios in the 2240k
790 dataset (removing transitions), we observe that the high Altai Neanderthal ancestry in Ma554
791 can be entirely explained by high Denisovan ancestry (and shared ancestry between
792 Neanderthals and Denisovans), while this is not the case in Tianyuan, who has high Altai
793 Neanderthal ancestry that cannot be explained by high Denisovan ancestry. When computing
794 these ratios in the 640k SNP dataset, we find that the levels of Denisovan ancestry in Ma554 is
795 compared to the levels in Mamanwa, but that Th503 shows slightly higher levels of Denisovan
796 ancestry. Differences across these datasets may be attributable to the low coverage of the
797 ancient genomes. The red-colored individuals are the samples from this study.

798