Investigating demic versus cultural diffusion and sex bias in the spread of Austronesian languages in Vietnam

Dinh Huong Thao, Tran Huu Dinh, Shigeki Mitsunaga, La Duc Duy, Nguyen Thanh Phuong, Nguyen Phuong Anh, Nguyen Tho Anh, Bui Minh Duc, Huynh Thi Thu Hue, Nguyen Hai Ha, et al.

PLOS ONE

RESEARCH ARTICLE

Investigating demic versus cultural diffusion and sex bias in the spread of Austronesian languages in Vietnam

Dinh Huong Thao[1], Tran Huu Dinh[1], Shigeki Mitsunaga[2], La Duc Duy[1], Nguyen Thanh Phuong[1,2], Nguyen Phuong Anh[1], Nguyen Tho Anh[1], Bui Minh Duc[1], Huynh Thi Thu Hue[1], Nguyen Hai Ha[1], Nguyen Dang Ton[1], Alexander Hübner[3], Brigitte Pakendorf[4], Mark Stoneking[3,5]*, Ituro Inoue[2]*, Nguyen Thuy Duong[1]*, Nong Van Hai[1]*

1 Institute of Genome Research, Vietnam Academy of Science and Technology, 18 Hoang Quoc Viet, Cau Giay, Hanoi, Vietnam, **2** Division of Human Genetics, National Institute of Genetics, Shizuoka, Japan, **3** Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, Leipzig, Germany, **4** Dynamique du Langage, UMR5596, CNRS & Université de Lyon, Lyon, France, **5** Laboratoire de Biométrie et Biologie Evolutive, Université Lyon 1, CNRS, UMR 5558, Villeurbanne, France

* vhnong@igr.vast.vn (NVH); tdnguyen@igr.vast.vn (NTD); itinoue@nig.ac.jp (II); stonekg@eva.mpg.de (MS)

Citation: Thao DH, Dinh TH, Mitsunaga S, Duy LD, Phuong NT, Anh NP, et al. (2024) Investigating demic versus cultural diffusion and sex bias in the spread of Austronesian languages in Vietnam. PLoS ONE 19(6): e0304964. https://doi.org/10.1371/journal.pone.0304964

Editor: Christian Reepmeyer, German Archaeological Institute: Deutsches Archaologisches Institut, GERMANY

Received: February 25, 2024

Accepted: May 21, 2024

Published: June 17, 2024

Data Availability Statement: The mtDNA sequences generated during the current study have been deposited in GenBank (https://www.ncbi.nlm.nih.gov/genbank/) under accession numbers PP654481 - PP654849.

Funding: This research was funded by the Ministry of Science and Technology, Vietnam (TL.CN.60/19).

Competing interests: The authors declare that they have no conflict of interest.

Abstract

Austronesian (AN) is the second-largest language family in the world, particularly widespread in Island Southeast Asia (ISEA) and Oceania. In Mainland Southeast Asia (MSEA), groups speaking these languages are concentrated in the highlands of Vietnam. However, our knowledge of the spread of AN-speaking populations in MSEA remains limited; in particular, it is not clear if AN languages were spread by demic or cultural diffusion. In this study, we present and analyze new data consisting of complete mitogenomes from 369 individuals and 847 Y-chromosomal single nucleotide polymorphisms (SNPs) from 170 individuals from all five Vietnamese Austronesian groups (VN-AN) and five neighboring Vietnamese Austroasiatic groups (VN-AA). We found genetic signals consistent with matrilocality in some, but not all, of the VN-AN groups. Population affinity analyses indicated connections between the AN-speaking Giarai and certain Taiwanese AN groups (Rukai, Paiwan, and Bunun). However, overall, there were closer genetic affinities between VN-AN groups and neighboring VN-AA groups, suggesting language shifts. Our study provides insights into the genetic structure of AN-speaking communities in MSEA, characterized by some contact with Taiwan and language shift in neighboring groups, indicating that the expansion of AN speakers in MSEA was a combination of cultural and demic diffusion.

Introduction

The Austronesian language family (AN), encompassing 1256 languages [1] spoken by approximately 360 million people, stretches from Madagascar to Hainan, Southeast Asia, Taiwan, and Near and Remote Oceania [2]. The ancestors of Austronesian-speaking peoples are thought to have originated in the Yangtze River Delta 9–6 thousand years ago (kya) [3,4] and then spread

to Taiwan. Nine out of ten AN primary sub-branches are exclusive to Taiwan, while all AN languages outside of Taiwan belong to just a single primary sub-branch (Malayo-Polynesian, consisting of more than 1200 languages), strongly suggesting that Taiwan was the source of the Austronesian expansion [2].

AN-speaking groups in Island Southeast Asia (ISEA) have been extensively examined from cultural and biological perspectives, contributing valuable data for elucidating the history of this region [5–17]. However, ethnic groups in Mainland Southeast Asia (MSEA) that speak AN languages have not yet received the same attention [18–25]. In MSEA, AN-speaking groups are found in Vietnam, Thailand, and Cambodia but account for only a small proportion of the population (e.g., about 1.32% % of ~100 million people in Vietnam) (www.gso.gov.vn; accessed the *General Statistics Office of Vietnam* in July 2023) [1]. A crucial question concerning the spread of AN-speaking groups in MSEA is the extent to which this was a process of demic diffusion (i.e., migration of AN-speaking groups from elsewhere spreading both their languages and their genes) vs. cultural diffusion (i.e., existing MSEA groups adopting an AN language with little genetic mixing with AN-speaking groups from elsewhere). Historical records point to the appearance of AN speakers along the coast of Indochina and the Gulf of Thailand around the 5th century BCE [26]; the close relationship of the AN languages of Vietnam with the Malayic branch of the family [2] points to northwest Borneo as the source of this migration [26]. However, whether the nature of the subsequent diffusion of AN languages was demic or cultural continues to be an open question. Here, we address this question, and potential sex bias in the spread of AN ancestry, by analyzing mtDNA and Y chromosome variation in AN-speaking groups from Vietnam.

Vietnam (VN), with its long coastline, occupies a key geographical position in MSEA and is home to 54 ethnic groups speaking languages classified into five language families: Austroasiastic (AA), Tai-Kadai (TK; also known as Kra-Dai), Hmong-Mien (HM), Sino Tibetan (ST) and Austronesian (AN). There are five recognized AN-speaking groups in Vietnam: Cham, Churu, Ede, Giarai, and Raglay, together accounting for ~1.32% of the national census size (www.gso.gov.vn; accessed the *General Statistics Office of Vietnam* in July 2023). It is thought that the first AN-speaking group to arrive in Vietnam were the ancestors of the Cham on the South Central Coast in 500 BCE, who probably originated in Borneo [26]. From there, the Cham rapidly extended their territory and established the Champa kingdom [27]. In the process, their languages underwent profound contact-induced changes due to the language shift of the autochthonous populations who were politically subordinate to the Cham [28,29]. Modern Austronesian-speaking communities in Vietnam (VN-AN) mainly occupy the mountainous Central Highlands and the South Central coastline. Their social customs, traditions, and family dynamics are related to the ancient Champa and comparable to their counterparts in ISEA [27]. To date, the maternal genetic ancestry of the Vietnamese Cham was described based only on the mtDNA HVS region [20], which provides limited resolution, while complete mtDNA genome sequences are available for two groups of Cham from Cambodia [19]. Recent studies examined both uniparental markers and genome-wide data for two AN-speaking groups, Ede and Giarai [18,21,23,24]. These findings were compared with other neighboring populations from different language families but not with other AN-speaking ethnicities on the mainland due to data scarcity, hindering attempts to trace the dispersal of the AN languages in MSEA.

Here, we present a novel set of complete mtDNA genome sequences of 369 individuals and Y chromosome haplotypes based on genotypes for 847 Y-chromosomal SNPs in 170 individuals, encompassing all five AN-speaking ethnic minorities and five neighboring AA groups in Vietnam (Fig 1). This new dataset is analyzed with comparative data from AN-speaking groups from elsewhere in ISEA and MSEA, and with non-AN groups from Vietnam, in order

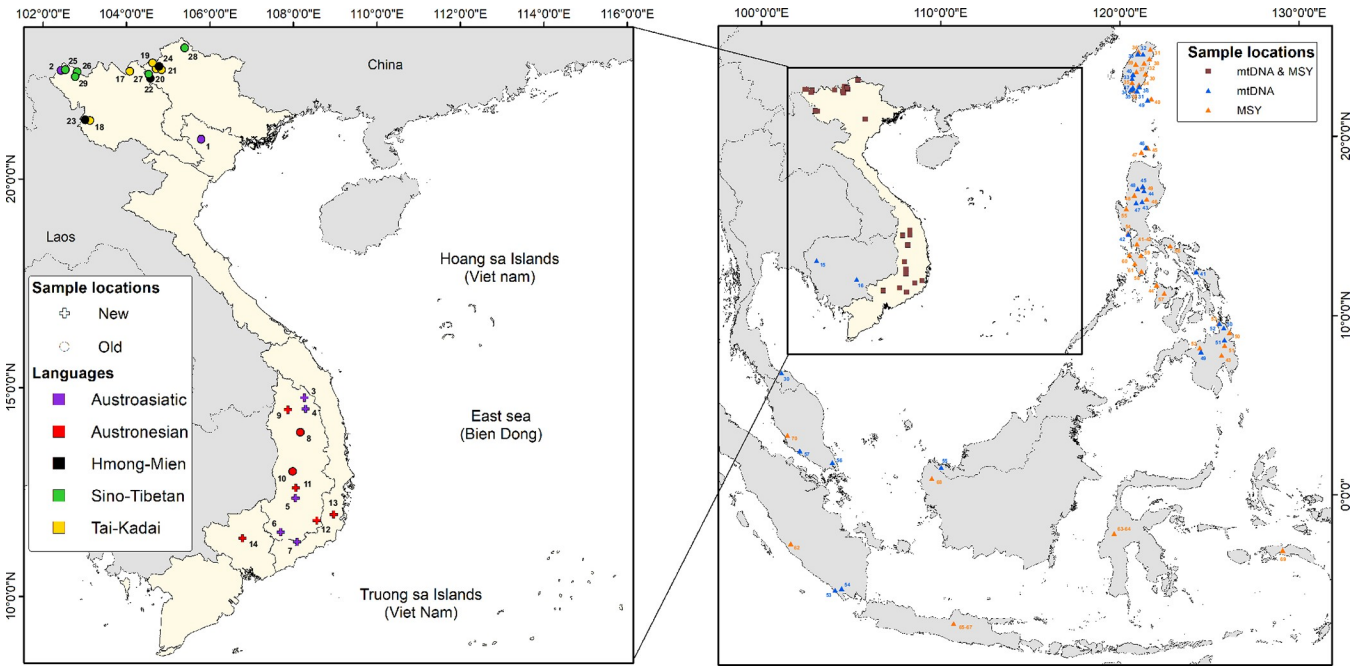

Fig 1. Geographical map displaying sampling locations. Triangles mark the sampling sites of all MSEA and ISEA populations included in the analysis. In the inset, crosses and circles mark the sampling locations of the 10 new and 17 previously published Vietnamese populations, respectively. For the Vietnamese populations, labels are color-coded by language family, as indicated in the legend. All the populations outside of Vietnam included here speak AN languages, and their locations are color-coded by mtDNA and MSY. The comparative populations for mtDNA and MSY analyses are numbered according to S1 and S4 Tables, respectively. The map was generated with QuantumGIS (Version 3.34), Opensource Geospatial Foundation (https://www.qgis.org/), and country administrative boundaries that were downloaded from https://www.diva-gis.org/.

https://doi.org/10.1371/journal.pone.0304964.g001

to evaluate the relative roles of demic vs. cultural diffusion, and potential sex bias, in the spread of AN languages in Vietnam. Patterns of genetic differentiation and haplotype sharing between populations provide insights into the question of demic vs. cultural diffusion in the spread of AN languages in Vietnam. With demic diffusion, we would expect to see closer relationships between VN-AN groups and non-VN-AN groups; with cultural diffusion, we would expect to see closer relationships between VN-AN groups and VN-non-AN groups. Moreover, differences between relationships based on mtDNA vs. the MSY would indicate differences in the maternal vs. paternal relationships of VN-AN groups. In the remainder of the paper, when describing populations we use the language family abbreviations as shorthand for "speaking a language of family X", e.g. "AN group" means "a group speaking an Austronesian language", and "AA population" means "a population speaking an Austroasiatic language".

Materials and methods

Sample information

Blood samples were collected from 369 Vietnamese males belonging to five Austronesian-speaking groups (Churu, Raglay, Ede, Giarai, and Cham) and five Austroasiatic-speaking groups (Coho, Hre, Bana, Mnong, and Ma) from the central highlands of Vietnam (Fig 1 and S1 Table). These samples were recruited between December 30th, 2019 and November 30th, 2022. Since Ede and Giarai individuals were also included in a previous study [18], we here distinguish the two sample sets as Ede-I and Giarai-I (taken from Duong et al. 2018) and Ede-II and Giarai-II (this study). Similarly, in order to distinguish the Cham individuals from Vietnam (this study) from Cham individuals from Cambodia included in Kloss-Brandstätter et al.

(2021), we label them Cham-VN and Cham-CB, respectively, further distinguishing the Cambodian samples from Battambang (Cham-CB-Bat) from those from Kampong (Cham-CB-Kam). All participants gave written, informed consent to donate blood, were unrelated, and self-identified to have at least three generations of the same ethnicity. This study received ethical approval from the Institutional Review Board of the Institute of Genome Research, Vietnam Academy of Science and Technology (No: 2-2019/NCHG-HĐĐĐ); further details concerning the sampling are provided in S1 Text.

mtDNA sequencing

Genomic DNA was extracted with the GeneJET Whole Blood Genomic DNA Purification Mini Kit (ThermoFisher Scientific, USA) following the manufacturer's protocol. Construction of genomic libraries and capture enrichment for mtDNA were performed as described previously [30]. The libraries were sequenced on the NovaSeq 6000 platform (Illumina, USA) with 150 bp paired end reads. The reads underwent quality control and were processed as described previously [31]. Reads were aligned to the Reconstructed Sapiens Reference Sequence (RSRS) [32] using an in-house alignment program, and a multiple sequence alignment was performed using MAFFT v7.490 [33]. The entire mtDNA sequences have been deposited in GenBank (accession numbers PP654481-PP654849). The mtDNA haplogroups were classified by HaploGrep2 [34] with PhyloTree mtDNA tree Build 17 [35]. Haplogroups labeled with an asterisk exclude all downstream subhaplogroups, whereas haplogroup labels without an asterisk include all downstream subhaplogroups. For instance, M71 refers to all haplotypes belonging to haplogroup M71, while R* refers to haplotypes assigned to R, but not assigned to any of the defined subhaplogroups within R. For subsequent analysis, except for haplogroup identification, we excluded positions with missing nucleotide (Ns) and the following sites: poly-C stretch of hypervariable segment 2 (HVS-II; nucleotide positions (np) 303–317); CA-repeat (np 514–523); C-stretch 1 (np 568–573), 12S rRNA (np 956–965), historical site (np 3,107), C-stretch 2 (np 5,895–5,899), 9 bp deletion/insertion (np 8,272–8,289), and poly-C stretch of hypervariable segment 1 (HVS-I; np 16,180–16,195). For comparative analyses, we included 1598 complete mtDNA sequences of 47 populations speaking Austronesian languages from Taiwan, ISEA (Philippines, Indonesia, and Malaysia), and MSEA (Vietnam, Cambodia, and South Thailand) [10,18,19,36–40] (S1 Table).

Y-Chromosomal SNP genotyping and data analysis

A total of 170 samples from the 10 above-mentioned Vietnamese populations were genotyped using the Affymetrix Axiom Genome-Wide Human Origins array. We then extracted a total of 2088 SNPs on the non-recombining region of the Y chromosome (MSY) for this study (S1 and S2 Datasets); analyses of the autosomal SNP data for these individuals are part of a further study. For subsequent analysis, these SNPs were aligned with ~2.3 million bases of the MSY of 600 previously published Vietnamese samples [24], resulting in 2079 overlapping SNP positions (S3 and S4 Datasets). Positions with two or fewer supporting reads were marked as missing. After removing these sites, 847 SNPs remained; the final SNP genotypes and their positions on hg19 are provided in Supplementary Materials (S5 and S6 Datasets). Genotypes were then used to identify haplogroups by yhaplo v1.1.2 [41] using a stopping condition parameter "ancStopThresh" = 10. Haplogroups were determined to the maximum depth possible, given the phylogeny of ISOGG version 15.73 (http://www.isogg.org/) and the SNPs for which we had data. Labels denoted with an asterisk in the text and figures are paragroups that do not include subgroups. Two previously published samples, Kinh09 and Mang304, were excluded from this study as their haplogroups changed to non-informative ancestral

haplogroups with the reduced set of SNPs used in this study. For the comparative dataset outside of Vietnam, we calculated the frequencies of haplogroups of 1081 samples from 58 AN-speaking populations from Taiwan and ISEA (Philippines, Indonesia, and Malaysia) [42,43] based on reported Y-chromosomal SNP datasets (S4 Table). We use the term haplotype throughout this paper to refer to the Y chromosome SNP sequences and not to STR profiles.

Data analysis

For both uniparentally-inherited markers, we calculated the number of unique haplotypes for each population with R v4.2.2 [44], using the function haplotype (package: pegas v1.3 [45]). Summary genetic statistics of mtDNA and Y chromosome variation, including haplotype diversity (*H*), nucleotide diversity (π), and its variance, were calculated by the functions hap.div and nuc.div of the same package, respectively. To visualize π and *H*, we computed the percentage difference from the mean for each population. The mean number of pairwise differences (MPD) was obtained by averaging over the sum of nucleotide differences for each pair of sequences within a population (R function: dist.dna, package: ape v5.7–1 [46]) divided by the total number of pairs. The haplotype sharing within and between populations was estimated as the proportion of pairs of identical sequences shared between populations using an in-house R script, which is provided upon request. Arlequin version 3.5.2.2 [47] was used to calculate the pairwise genetic distance (Φ_{ST} distances) among the populations. For mtDNA markers, we used Φ_{ST} distances for computing a nonmetric multidimensional scaling (MDS) analysis (function: isoMDS, package: MASS v7.3–60.0.1 [48]). The correspondence analysis (CA) was computed based on haplogroup frequencies in R using libraries "vegan v2.6–4" [49] and "ca v0.71.1" [50].

Results

Haplogroup classification of mtDNA and Y-chromosomal SNPs

The entire mitochondrial genomes of 369 individuals were sequenced with an average read depth of 312X (range: 26–3357); 222 distinct sequences (haplotypes) were obtained and assigned to 65 haplogroups by HaploGrep2, all belonging to the two macro-haplogroups M and N (S2 Table). Of the 65 haplogroups, 17 (26.15%) were singletons (S2 Table). Overall, most sequences belonged to the macro-haplogroup M (54.74%), followed by haplogroups B (18.16%) and F1 (13.55%) (S3 Table and S1A Fig). However, there are notable differences in haplogroup frequencies between different MSEA-AN populations, and even between sub-samples from the same ethnolinguistic group, e.g. the Giarai-I and Giarai-II, and also the Cham from Vietnam and those from Cambodia (S2 Fig). The Giarai-I and Ede-I and -II have high frequencies (37–42%) of M71, which is practically absent in other MSEA-AN populations, including the Giarai-II, which were sampled in a location that is about 75 km away from Giarai-I. Similarly, M24b, which is practically absent in other AN populations from MSEA, is found in the Raglay at 35% frequency; haplogroup B, which is found at 17% frequency on average in the other MSEA-AN populations, is found in only very low frequencies in Giarai-I (4%) and Raglay (3%; S1 Table).

For the Y chromosome, a total of 18 polymorphic sites were found; these defined 17 haplotypes belonging to 16 haplogroups (S5 Table). Seven of the 16 haplogroups were shared between at least 2 populations (S6 Table). Of the 16 haplogroups, 5 (31.25%) are singletons. Overall, O1b is the overwhelmingly predominant haplogroup (74.71%), in particular subhaplogroup O1b1a1a (O-M95) (72.35%), followed by O2a (8.82%), R* (4.12%), and R1a (4.12%) (S5 Table and S1B Fig). However, there are some striking differences in haplogroup frequencies among AN populations. For example, O-M95 is very common in VN-AN populations

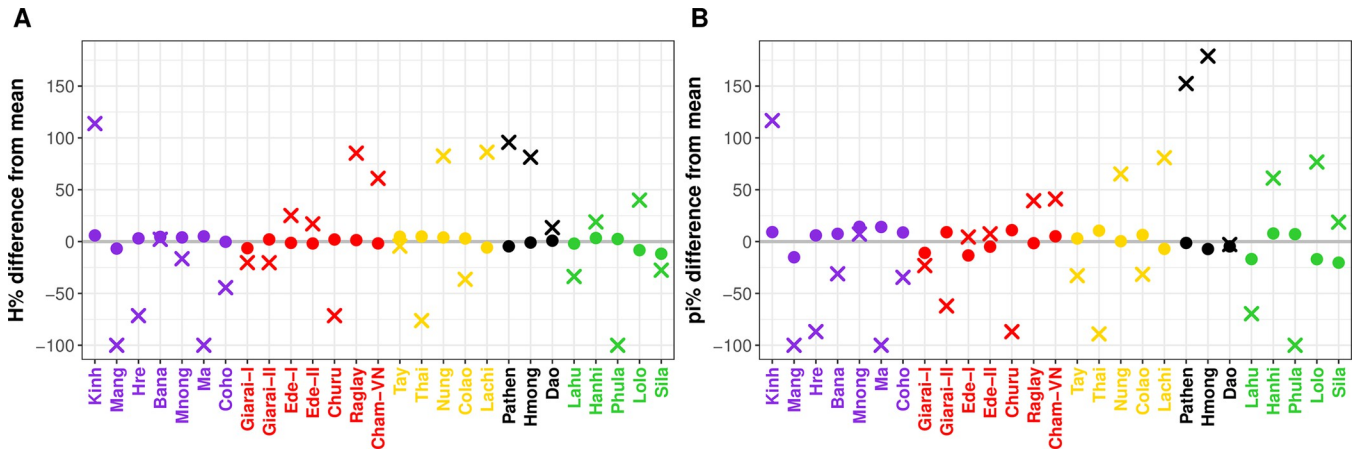

Fig 2. Genetic diversity indices shown as the percent difference from the mean. (A) Haplotype diversity. (B) Nucleotide diversity. Crosses and dots represent MSY and mtDNA data, respectively. Population labels are color-coded by language family, with Austroasiatic in purple, Austronesian in red, Tai-Kadai in yellow, Hmong-Mien in black, and Sino-Tibetan in green. The gray line indicates the mean across populations.

https://doi.org/10.1371/journal.pone.0304964.g002

(42–90%), but occurs at somewhat lower frequencies (14–50%) in Indonesian and Malaysian AN populations, lower still (10%) in a Taiwanese AN population, and at very low frequencies (2% and 5%) in two Philippine AN populations (S3 Fig). The Cham from Vietnam have a high frequency (31%) of J (specifically, sub-haplogroup J2a1), which is only present at very low frequencies in Philippine AN (3%) and Indonesian AN (2%) and is absent in Taiwanese and Malaysian populations (S3 Fig and S4 Table).

Genetic diversity

The nucleotide diversity (π) and haplotype diversity (H) for mtDNA sequences and Y-chromosomal SNPs were calculated for 958 individuals (369 newly sequenced here and 598 previously published individuals) and 768 individuals (170 newly genotyped here individuals and 598 previously published), respectively (Fig 2 and S7 Table). For mtDNA, the newly genotyped Vietnamese groups had an average haplotype diversity value of 0.9545, ranging from 0.926 (Ede-II and Cham-VN) to 0.991 (Ma). The average nucleotide diversity (π) was 0.00218 and did not vary much among groups; Mnong had the highest value (π = 0.00233), and Ede-II the lowest value (π = 0.00193). For the Y chromosome, the average values were 0.2947 and 0.005489 for haplotype and nucleotide diversities, respectively. Raglay had the highest haplotype diversity (H = 0.649), while Cham-VN had the highest nucleotide diversity (π = 0.00988). The most homogenous group was Ma (H = 0.00, π = 0.00), but the sample size (n = 6) was also the lowest for the Ma. The MPDs were strikingly low in Ma and Churu, less than 1.0, which probably reflects the very high frequencies of haplogroup O1b1a1a (O-M95) found in 6/6 Ma and 18/20 Churu individuals, respectively, indicating homogeneity of the Ma and Churu paternal ancestries. While there is a tendency for the VN-AN groups to have higher Y chromosomal diversity than their AA-speaking neighbors (Fig 2), as expected for matrilocal populations [51], mtDNA diversity does not seem to be lower in VN-AN groups than in other VN groups.

Population affinity

An mtDNA haplotype sharing matrix was computed for Vietnamese populations and for AN populations from MSEA and ISEA (Fig 3; a similar analysis is not shown for the Y

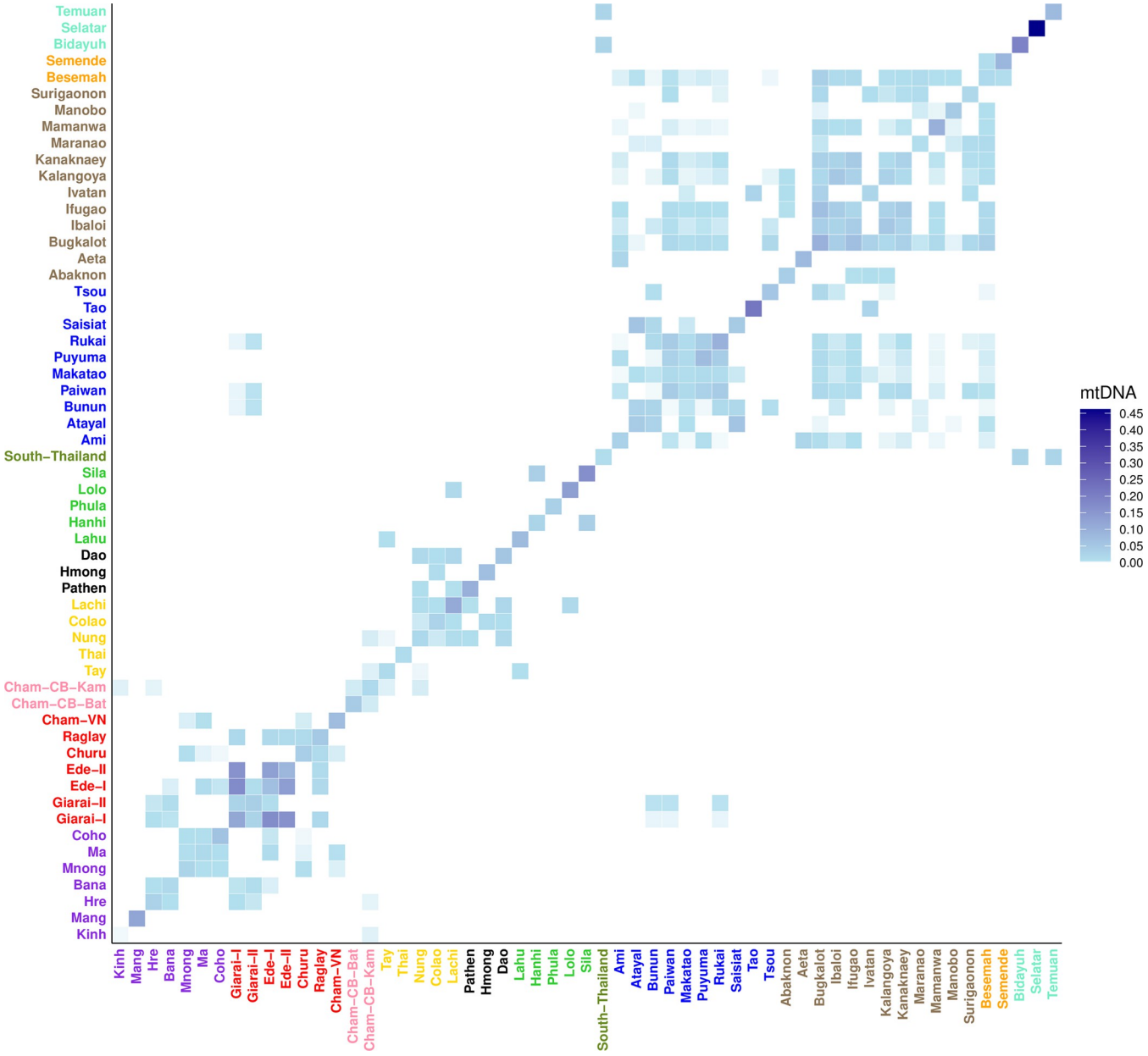

Fig 3. mtDNA haplotype sharing between Vietnamese and non-Vietnamese AN populations. mtDNA shared haplotype frequencies are represented by the blue gradient. Population labels are color-coded by language family with Austroasiatic in purple, Vietnamese Austronesian in red, Thai Austronesian in olive drab, Tai-Kadai in yellow, Hmong-Mien in black, Sino-Tibetan in lime, Cambodian Austronesian in pink, Taiwanese Austronesian in blue, Philippine Austronesian in brown, Indonesian Austronesian in orange, and Malaysian Austronesian in turquoise.

https://doi.org/10.1371/journal.pone.0304964.g003

chromosome as there are too few Y chromosome haplotypes to be informative). High values of sharing within groups are an indication of small population size and/or a recent bottleneck, while high values of sharing between populations suggest recent genetic contact or shared ancestry. Among VN-AN groups, within-group sharing is highest in Giarai-I (0.1168) and lowest in Giarai-II (0.0375), comparable to what is observed for non-AN-speaking VN groups (0.009–0.1206 for VN-AA, 0.0109–0.111 for VN-TK, 0.055–0.1 for VN-HM, and 0.0246–0.1675 for VN-ST; Fig 3). In non-VN-AN populations, higher amounts of intrapopulation

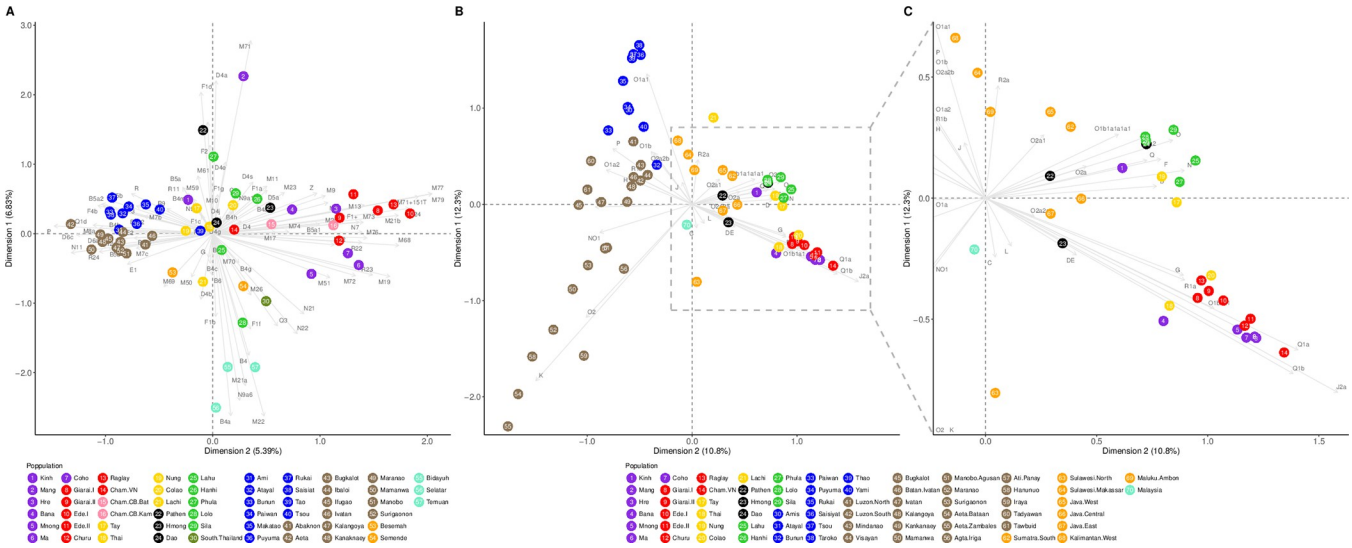

Fig 4. Correspondence analysis plot based on haplogroup frequencies of Vietnamese and non-Vietnamese populations. (A) mtDNA. (B, C) MSY, with the plot on the right (C) zooming in on the region indicated by the dashed rectangle in the full plot (B). Population labels are color-coded by language family with Austroasiatic in purple, Vietnamese Austronesian in red, Tai-Kadai in yellow, Hmong-Mien in black, Sino-Tibetan in lime, Cambodian Austronesian (Cham-CB-Bat and Cham-CB-Kam) in pink, Thai Austronesian in olive drab, Taiwanese Austronesian in blue, Philippine Austronesian in brown, Indonesian Austronesian in orange, and Malaysian Austronesian in turquoise. Haplogroup labels are in gray.

https://doi.org/10.1371/journal.pone.0304964.g004

sharing are observed in some groups, such as Malaysian Selatar (0.4619), Taiwanese Tao (0.2199), and Malaysian Bidayuh (0.1937). There is some sharing between Vietnamese AN groups and their AA-speaking neighbors, in particular between the Giarai and the Hre and the Bana, between the Ede-I and the Bana and the Ma and the Coho, between the Churu and the Mnong and the Ma and the Coho, and between the Cham-VN and the Ma and the Mnong. However, there is no sharing with other AN populations, with the exception of the two Giarai groups, who share haplotypes with some populations from Taiwan. Furthermore, there is sharing between VN AA-speaking groups Hre and Kinh and Cham from Cambodia.

To visualize the relationships among MSEA and ISEA populations, correspondence analysis (CA) was carried out, based on the haplogroup frequencies for both mtDNA and the Y chromosome (Fig 4). For mtDNA, the first dimension differentiated Mang at the top (and to a lesser extent, Pathen and Phula) from Malaysian Austronesian at the bottom (to a lesser extent with southern Thai Austronesian, Indonesian Austronesian, Lolo, and Lachi) (Fig 4A). The second dimension separated most of the Philippine and Taiwanese Austronesian groups at the left from the VN-AN and VN-AA groups at the right. Within this dimension, Cham-VN is pulled towards Taiwanese populations; Cham-CB is grouped with the Southern VN populations, with Cham-CB-Bat being close to the Cham-VN while Cham-CB-Kam is close to the VN-AN Giarai-II and the Churu. All other groups are more towards the middle. On the Y-chromosome CA plot, the first dimension separated most of the Philippine AN groups, including 6 traditional hunter-gatherer groups (Mamanwa, Aeta-Bataan, Aeta-Zambles, Agta-Iriga, Iraya, and Ati-Panay) [42]. The second dimension separated Philippine and Taiwanese AN groups at the left from VN-AN and VN-AA groups at the right, similar to the mtDNA plot, except that some VN-TK and VN-ST groups also cluster with the VN-AN and VN-AA groups (Fig 4B). Indonesian AN groups tend to be in the middle in both the mtDNA and Y plots.

To further investigate genetic differences between populations, pairwise Φ_{ST} distances based on mtDNA sequences were generated among Vietnamese populations and AN-speaking populations in other countries (S4 Fig). It was not possible to carry out a similar analysis for

the MSY as we only have haplogroup frequencies from the published data, not comparable sequence/SNP data. Among the VN populations, closer relationships (as demonstrated by non-significant Φ_{ST} values) are found between the Giarai-II and the AA-speaking Hre, between the Giarai-I and the Ede-I, between the Ede-I and the Ede-II, and between the Cham-VN and the TK-speaking Thai. Among VN and non-VN-AN populations, a non-significant Φ_{ST} value is found between the Cham-VN and the Cham-CB-Bat.

An MDS analysis was carried out, using the matrix of pairwise Φ_{ST} distances for mtDNA (S5 Fig), to further examine the genetic relationships of the AN groups from Vietnam to other populations from Vietnam and to AN populations from MSEA and ISEA. Because a high stress value was obtained with a two-dimensional analysis, we increased the dimensions to three. The results indicate that the Ede-I and Giarai-I are close to the Taiwanese Atayal and Saisiat in the first dimension, while the Cham-VN and Churu are separate from these and closer to the Cham-CB groups; Ede-II, Giarai-II, and Raglay show affinities with a variety of groups, including AA-speaking Bana and Hre, Taiwanese Amis, Indonesian Besemah, and the Manobo from the Philippines (S5 Fig). In the second dimension, the VN-AN groups are in a central cloud that includes various groups (both AN and non-AN), while the third dimension places the VN-AN groups close to the Cham-CB groups and the VN-AA groups, as well as some other groups from Vietnam.

Discussion

The five Austronesian-speaking groups in Vietnam (Cham, Raglay, Ede, Giarai, and Churu) share many cultural and social practices [52], yet their genetic relationships with one another, with other Vietnamese groups, and with other Austronesian groups have not been explored in detail. Here, we generated complete mtDNA genome sequences, and determined Y chromosome haplogroup frequencies based on genotyping 847 SNPs, for all five Vietnamese Austronesian-speaking groups, and from five neighboring Austroasiatic-speaking groups. We then compared these to published data from Vietnamese non-AN groups and non-Vietnamese AN groups, in order to assess the role of cultural vs. demic diffusion, and potential sex-biased migration, in the spread of Austronesian languages in Vietnam.

The haplogroup frequency plots for both mtDNA and the MSY (S2 and S3 Figs) indicate that VN-AN groups are, overall, more similar to their neighboring VN-AA groups than they are to AN groups from elsewhere. The correspondence analyses (Fig 4) further support this conclusion: for mtDNA, the VN-AN groups overlap with their neighboring VN-AA groups (and are close to Cham-CB groups); while for the MSY, the VN-AN and VN-AA groups are clustered together (and with two VN-TK groups, Thai and Colao) and apart from other groups. The clustering of the Thai and Colao with VN-AN and VN-AA populations reflects the high frequencies of haplogroup O-M95 in these groups (S3 Fig and S4 Table); however, it is unlikely to reflect recent interactions, since it is not evident in the MSY sequence data nor in the genome-wide data available for a subset of these populations [23,24].

The close genetic similarity between VN-AN groups and neighboring VN-AA groups would suggest a primary role for cultural diffusion in the spread of AN languages in Vietnam, in accordance with linguistic data suggesting language shifts from autochthonous populations beginning soon after the initial AN settlement [29]. Indeed, historical records indicate that the AN-speaking Champa maintained close relations with Mon-Khmer local groups, forming both social-economic partnerships and marital alliances [27]. The close genetic relationship between the Cham from Cambodia and the Cham from Vietnam that we see in the mtDNA data mirror their close linguistic relationship: the two groups speak the same language, although their dialects have by now become mutually unintelligible [53]. The split between the

two groups is probably recent, since the Cham are thought to have settled in Cambodia only after the fall of the southern Cham capital of Vijaya in 1471 CE [54]. That this migration occurred after the incorporation of AA-speaking populations into the Cham population is shown by the sharing of mtDNA haplotypes between the Hre and Kinh from Vietnam and the Cham from Cambodia.

A more detailed analysis of mtDNA haplotypes, based on Φ_{ST} distances (S4 Fig) and MDS analysis (S5 Fig), provides further support for close relationships between the VN-AN groups and neighboring VN-AA groups, with non-VN-AN groups generally showing more distant relationships. However, a striking result of the haplotype-based analyses is the sharing of mtDNA haplotypes between the Giarai (both groups) and the Bunun, Paiwan, and Rukai of Taiwan (Fig 3). The latter two are southern Taiwanese groups that, in genome-wide data, show closer relationships to AN groups outside Taiwan than do other Taiwanese groups [55]. This would suggest a role for demic diffusion in the introduction of AN languages to Vietnam. However, the direct connection between Taiwan and Vietnam that emerges from the genetic data is not supported by linguistics, which identifies a close relationship of the VN-AN languages with the AN languages of western Indonesia, and specifically northwest Borneo, as mentioned above [2,26]. Skeletal remains associated with jar burials at the Hoa Diem coastal site in central Vietnam show affinities both with populations from Taiwan and the Philippines and from Sumatra and Java [56], thus bridging both the genetic and the linguistic evidence. The discrepancies between genetic, archaeological and linguistic data point to a very complex history of the spread of AN languages to MSEA, which genome-wide studies might help disentangle.

VN-AN groups are characterized by matrilocal residence practice [52,57]: the husband often relocates from his home village to that of his wife. As such, Y-chromosome diversity is expected to be higher in matrilocal groups than in patrilocal groups, while the opposite is expected for mtDNA diversity [51]. While the 5 VN-AN populations did exhibit generally higher-than-average Y chromosome diversity values, especially compared to the neighboring AA-speaking populations (Fig 2 and S7 Table), the mtDNA diversity values were about the same as in other VN groups. Such departures from the strict predictions of the impact of patrilocality vs. matrilocality on patterns of genetic variation are not unexpected, given that many factors impact genetic variation, thereby complicating interpretations of mtDNA and MSY variation [58]. In the case of VN-AN groups, even though marriages are preferred within communities or between groups with similar cultural practices, intermarriages between neighboring groups often do occur [27,59]. The Vietnamese fundamental residence pattern is either patrilocal or ambilocal [52]; therefore, admixture between neighboring communities might lead to genetic features inconsistent with their cultural kinship systems. This deviation from the expected pattern of genetic variation with matrilocal residence thus provides further support for the substantial amount of contact with local populations–and hence, cultural diffusion–that was involved in the spread of AN languages in Vietnam.

In conclusion, our survey of the mtDNA and MSY relationships of all of the extant VN-AN groups, along with their AA neighbors, illustrates a complex history of migrations as well as cultural diffusion via language shifts and contact, resulting in the spread of AN languages across Vietnam. We see more or less the same picture for both mtDNA and the MSY, suggesting little sex bias in this process (consistent with a primary role for cultural diffusion); however, we caution that our inferences based on the MSY are more limited than for mtDNA, due to the underlying nature of the data. More in-depth studies of MSY variation, as well as genome-wide data, will provide a more holistic picture of the history of AN groups in Vietnam.

Supporting information

S1 Fig. Barplot showing major haplogroup frequencies of 10 newly sampled Vietnamese populations. (A) mtDNA, (B) MSY. Austronesian-speaking groups are in red font and Austroasiatic-speaking groups are in purple font.
(PDF)

S2 Fig. Barplot showing major mtDNA haplogroup frequencies of 57 Vietnamese and non-Vietnamese populations. Population labels are color coded by language family with Vietnamese Austroasiatic in purple, Vietnamese Austronesian in red, Vietnamese Tai-Kadai in yellow, Vietnamese Hmong-Mien in black, Vietnamese Sino-Tibetan in lime, Cambodian Austronesian (Cham-CB-Bat and Cham-CB-Kam) in pink, Thai Austronesian (South-Thailand) in olive drab, Taiwanese Austronesian in blue, Philippine Austronesian in brown, Indonesian Austronesian in orange, and Malaysian Austronesian in turquoise.
(PDF)

S3 Fig. Barplot showing major MSY haplogroup frequencies of 68 Vietnamese and non-Vietnamese populations. Population labels are color coded by language family with Vietnamese Austroasiatic in purple, Vietnamese Austronesian in red, Vietnamese Tai-Kadai in yellow, Vietnamese Hmong-Mien in black, Vietnamese Sino-Tibetan in lime, Taiwanese Austronesian in blue, Philippine Austronesian in brown, Indonesian Austronesian in orange, and Malaysian Austronesian in turquoise.
(PDF)

S4 Fig. Pairwise Φ_{ST} distances between populations for mtDNA. Distances that are not significantly different from zero are marked with a cross symbol ($p > 0.05$). Population labels are color coded by language family with Austroasiatic in purple, Vietnamese Austronesian in red, Tai-Kadai in yellow, Hmong-Mien in black, Sino-Tibetan in lime, Cambodian Austronesian in pink, Thai Austronesian in olive drab, Taiwanese Austronesian in blue, Philippine Austronesian in brown, Indonesian Austronesian in orange, and Malaysian Austronesian in turquoise.
(PDF)

S5 Fig. Three-dimensional MDS analysis based on the Φ_{ST} distance matrix of mtDNA sequences for 55 Vietnamese and non-Vietnamese populations. Results are shown as two-dimensional plots for each combination of the three dimensions, and the stress value is in percent. Population labels are color coded by language family with Vietnamese Austroasiatic in purple, Vietnamese Austronesian in red, Vietnamese Tai-Kadai in yellow, Vietnamese Hmong-Mien in black, Vietnamese Sino-Tibetan in lime, Cambodian Austronesian (Cham-CB-Bat and Cham-CB-Kam) in pink, Thai Austronesian in olive drab, Taiwanese Austronesian in blue, Philippine Austronesian in brown, Indonesian Austronesian in orange, and Malaysian Austronesian in turquoise.
(PDF)

S1 Table. Population information and haplogroup frequencies for mtDNA.
(XLSX)

S2 Table. Haplogroup frequencies of 369 complete mtDNA sequences from Vietnamese samples.
(XLSX)

S3 Table. Relative frequencies of Vietnamese mtDNA major haplogroups by population.
(XLSX)

S4 Table. Population information and haplogroup frequencies for Y chromosome.
(XLSX)

S5 Table. Y chromosome haplogroup frequencies based on SNP genotypes for 170 Vietnamese samples.
(XLSX)

S6 Table. Relative frequencies of Vietnamese Y chromosome major haplogroups by population.
(XLSX)

S7 Table. mtDNA and Y chromosome diversity values for Vietnamese populations.
(XLSX)

S1 Text. Ethical approval and sampling procedure.
(DOCX)

S1 Dataset. Chromosomal positions on hg19 for 2088 SNPs.
(PDF)

S2 Dataset. Genotypes of 2088 SNPs for 170 newly genotyped individuals.
(PDF)

S3 Dataset. Chromosomal positions on hg19 for 2079 SNPs.
(PDF)

S4 Dataset. Genotypes of 2079 SNPs for 768 individuals (170 newly genotyped here individuals and 598 previously published).
(PDF)

S5 Dataset. Chromosomal positions on hg19 for 847 SNPs.
(PDF)

S6 Dataset. Genotypes of 847 SNPs for 768 individuals (170 newly genotyped here individuals and 598 previously published).
(PDF)

Acknowledgments

We thank all sample donors for contributing to this research. We thank Bui Quang Thanh, Buon Krong Tuyet Nhung, Vo Thi Bich Thuy, and Do Hai Quynh for valuable advice and support. We thank the National Institute of Genetics, Japan, and the Max Planck Society for research support.

Author Contributions

Conceptualization: Brigitte Pakendorf, Mark Stoneking, Ituro Inoue, Nguyen Thuy Duong, Nong Van Hai.

Data curation: Tran Huu Dinh, La Duc Duy, Alexander Hübner.

Formal analysis: Dinh Huong Thao, Tran Huu Dinh.

Funding acquisition: Nguyen Thuy Duong.

Investigation: Nguyen Thanh Phuong, Nguyen Phuong Anh, Nguyen Tho Anh, Bui Minh Duc, Huynh Thi Thu Hue, Nguyen Hai Ha, Nguyen Dang Ton.

Methodology: Dinh Huong Thao, Tran Huu Dinh, Shigeki Mitsunaga, La Duc Duy, Nguyen Thanh Phuong, Bui Minh Duc, Nguyen Dang Ton, Alexander Hübner.

Project administration: Huynh Thi Thu Hue, Nguyen Hai Ha.

Supervision: Brigitte Pakendorf, Mark Stoneking, Ituro Inoue, Nguyen Thuy Duong, Nong Van Hai.

Validation: Shigeki Mitsunaga, Nguyen Phuong Anh, Nguyen Tho Anh.

Writing – original draft: Dinh Huong Thao, Nguyen Thuy Duong.

Writing – review & editing: Dinh Huong Thao, Tran Huu Dinh, Shigeki Mitsunaga, La Duc Duy, Nguyen Thanh Phuong, Nguyen Phuong Anh, Nguyen Tho Anh, Bui Minh Duc, Huynh Thi Thu Hue, Nguyen Hai Ha, Nguyen Dang Ton, Alexander Hübner, Brigitte Pakendorf, Mark Stoneking, Ituro Inoue, Nguyen Thuy Duong, Nong Van Hai.

References

1. Eberhard DM, Simons GF, Fennig CD. Ethnologue: Languages of the World. 26th ed. Dallas, Texas: SIL International; 2023.
2. Blust R. The Austronesian languages. Revised ed: Asia-Pacific Linguistics, The Australian National University; 2013.
3. Reid LA. Benedict's Austro-Tai Hypothesis—An Evaluation. Asian Perspectives. 1984; 26(1):19–34.
4. Sun J, Li YX, Ma PC, Yan S, Cheng HZ, Fan ZQ, et al. Shared paternal ancestry of Han, Tai-Kadai-speaking, and Austronesian-speaking populations as revealed by the high resolution phylogeny of O1a-M119 and distribution of its sub-lineages within China. Am J Phys Anthropol. 2021; 174(4):686–700. https://doi.org/10.1002/ajpa.24240 PMID: 33555039
5. Adelaar KA. Borneo as a Cross-Roads for Comparative Austronesian Linguistics. In: Bellwood P, Fox JJ, Tryon D, editors. The Austronesians. Historical and Comparative Perspectives: ANU Press; 2006. p. 81–102.
6. Arenas M, Gorostiza A, Baquero JM, Campoy E, Branco C, Rangel-Villalobos H, et al. The Early Peopling of the Philippines based on mtDNA. Sci Rep. 2020; 10(1):4901. https://doi.org/10.1038/s41598-020-61793-7 PMID: 32184451
7. Bellwood P. The Archaeological Record of Early Austronesian Communities. In: Bellwood P, editor. Prehistory of the Indo-Malaysian Archipelago. Revised ed: ANU Press; 2007. p. 201–54.
8. Blust R. The Austronesian Homeland and Dispersal. Annu Rev Linguist. 2019; 5(1):417–34. https://doi.org/10.1146/annurev-linguistics-011718-012440
9. Gomes SM, Bodner M, Souto L, Zimmermann B, Huber G, Strobl C, et al. Human settlement history between Sunda and Sahul: a focus on East Timor (Timor-Leste) and the Pleistocenic mtDNA diversity. BMC Genomics. 2015; 16(1):70. https://doi.org/10.1186/s12864-014-1201-x PMID: 25757516
10. Gunnarsdóttir ED, Li M, Bauchet M, Finstermeier K, Stoneking M. High-throughput sequencing of complete human mtDNA genomes from the Philippines. Genome Res. 2011; 21(1):1–11. https://doi.org/10.1101/gr.107615.110 PMID: 21147912
11. Horridge A. The Austronesian Conquest of the Sea—Upwind. In: Bellwood P, Fox JJ, Tryon D, editors. The Austronesians. Historical and Comparative Perspectives: ANU Press; 2006. p. 143–60.
12. Lim LS, Ang KC, Mahani MC, Shahrom AW, Md-Zain BM. Mitochondrial DNA Polymorphism and Phylogenetic Relationships of Proto Malays in Peninsular Malaysia. J Biol Sci. 2010; 10(2):71–83. https://doi.org/10.3923/jbs.2010.71.83
13. Mona S, Grunz KE, Brauer S, Pakendorf B, Castri L, Sudoyo H, et al. Genetic Admixture History of Eastern Indonesia as Revealed by Y-Chromosome and Mitochondrial DNA Analysis. Mol Biol Evol. 2009; 26(8):1865–77. https://doi.org/10.1093/molbev/msp097 PMID: 19414523
14. Sather C. Sea Nomads and Rainforest Hunter-Gatherers: Foraging Adaptations in the Indo-Malaysian Archipelago. In: Bellwood P, Fox JJ, Tryon D, editors. The Austronesians: Historical and Comparative Perspectives. 1st ed: ANU Press; 2006. p. 245–85.

15. Soares PA, Trejaut JA, Rito T, Cavadas B, Hill C, Eng KK, et al. Resolving the ancestry of Austronesian-speaking populations. Hum Genet. 2016; 135(3):309–26. https://doi.org/10.1007/s00439-015-1620-z PMID: 26781090

16. Tabbada KA, Trejaut J, Loo J-H, Chen Y-M, Lin M, Mirazón-Lahr M, et al. Philippine mitochondrial DNA diversity: a populated viaduct between Taiwan and Indonesia? Mol Biol Evol. 2010; 27(1):21–31. https://doi.org/10.1093/molbev/msp215 PMID: 19755666

17. Tumonggor MK, Karafet TM, Hallmark B, Lansing JS, Sudoyo H, Hammer MF, et al. The Indonesian archipelago: an ancient genetic highway linking Asia and the Pacific. J Hum Genet. 2013; 58(3):165–73. https://doi.org/10.1038/jhg.2012.154 PMID: 23344321

18. Duong NT, Macholdt E, Ton ND, Arias L, Schröder R, Van Phong N, et al. Complete human mtDNA genome sequences from Vietnam and the phylogeography of Mainland Southeast Asia. Sci Rep. 2018; 8(1):11651. https://doi.org/10.1038/s41598-018-29989-0 PMID: 30076323

19. Kloss-Brandstätter A, Summerer M, Horst D, Horst B, Streiter G, Raschenberger J, et al. An in-depth analysis of the mitochondrial phylogenetic landscape of Cambodia. Sci Rep. 2021; 11(1):10816. https://doi.org/10.1038/s41598-021-90145-2 PMID: 34031453

20. Peng MS, Quang HH, Dang KP, Trieu AV, Wang HW, Yao YG, et al. Tracing the Austronesian Footprint in Mainland Southeast Asia: A Perspective from Mitochondrial DNA. Mol Biol Evol. 2010; 27(10):2417–30. https://doi.org/10.1093/molbev/msq131 PMID: 20513740

21. Pham PD, Hoang TL, Le KT, Le PT, Nguyen NN, Tran HL, et al. The first data of allele frequencies for 23 autosomal STRs in the Ede ethnic group in Vietnam. Leg Med. 2022; 57:102072. https://doi.org/10.1016/j.legalmed.2022.102072 PMID: 35461037

22. Dancause KN, Chan CW, Arunotai NH, Lum JK. Origins of the Moken Sea Gypsies inferred from mitochondrial hypervariable region and whole genome sequences. J Hum Genet. 2009; 54(2):86–93. https://doi.org/10.1038/jhg.2008.12 PMID: 19158811

23. Liu D, Duong NT, Ton ND, Van Phong N, Pakendorf B, Van Hai N, et al. Extensive Ethnolinguistic Diversity in Vietnam Reflects Multiple Sources of Genetic Diversity. Mol Biol Evol. 2020; 37(9):2503–19. https://doi.org/10.1093/molbev/msaa099 PMID: 32344428

24. Macholdt E, Arias L, Duong NT, Ton ND, Van Phong N, Schröder R, et al. The paternal and maternal genetic history of Vietnamese populations. Eur J Hum Genet. 2020; 28(5):636–45. https://doi.org/10.1038/s41431-019-0557-4 PMID: 31827276

25. Zhang X, Qi X, Yang Z, Serey B, Sovannary T, Bunnath L, et al. Analysis of mitochondrial genome diversity identifies new and ancient maternal lineages in Cambodian aborigines. Nat Commun. 2013; 4(1):2599. https://doi.org/10.1038/ncomms3599 PMID: 24121720

26. Sidwell P. 33 Southeast Asian mainland: linguistic history. In: Ness I, editor. The Encyclopedia of Global Human Migration. 1 ed: Wiley; 2013. p. 1–10.

27. Lockhart BM, Tran KP. The Cham of Vietnam: history, society, and art: NUS Press; 2011.

28. Grant A. The effects of intimate multidirectional linguistic contact in Chamic. In: Grant A, Sidwell P, editors. Chamic and beyond studies in mainland Austronesian languages: Pacific Linguistics, The Australian National University; 2005. p. 37–104.

29. Sidwell P. Acehnese and the Aceh-Chamic language family. In: Grant A, Sidwell P, editors. Chamic and Beyond: Studies in mainland Austronesian languages: Pacific Linguistics, The Australian National University; 2005. p. 211–46.

30. Maricic T, Whitten M, Pääbo S. Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products. PLoS ONE. 2010; 5(11):e14004. https://doi.org/10.1371/journal.pone.0014004 PMID: 21103372

31. Arias L, Barbieri C, Barreto G, Stoneking M, Pakendorf B. High-resolution mitochondrial DNA analysis sheds light on human diversity, cultural interactions, and population mobility in Northwestern Amazonia. Am J Phys Anthropol. 2018; 165(2):238–55. https://doi.org/10.1002/ajpa.23345 PMID: 29076529

32. Behar Doron M, van Oven M, Rosset S, Metspalu M, Loogväli E-L, Silva Nuno M, et al. A "Copernican" Reassessment of the Human Mitochondrial DNA Tree from its Root. Am J Hum Genet. 2012; 90(4):675–84. https://doi.org/10.1016/j.ajhg.2012.03.002 PMID: 22482806

33. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013; 30(4):772–80. https://doi.org/10.1093/molbev/mst010 PMID: 23329690

34. Weissensteiner H, Pacher D, Kloss-Brandstätter A, Forer L, Specht G, Bandelt H-J, et al. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic Acids Res. 2016; 44(W1):W58–W63. https://doi.org/10.1093/nar/gkw233 PMID: 27084951

35. van Oven M, Kayser M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat. 2009; 30(2):E386–94. https://doi.org/10.1002/humu.20921 PMID: 18853457

36. Gunnarsdóttir ED, Nandineni MR, Li M, Myles S, Gil D, Pakendorf B, et al. Larger mitochondrial DNA than Y-chromosome differences between matrilocal and patrilocal groups from Sumatra. Nat Commun. 2011; 2(1):228. https://doi.org/10.1038/ncomms1235 PMID: 21407194

37. Jinam TA, Hong L-C, Phipps ME, Stoneking M, Ameen M, Edo J, et al. Evolutionary History of Continental Southeast Asians: "Early Train" Hypothesis Based on Genetic Analysis of Mitochondrial and Autosomal DNA Data. Mol Biol Evol. 2012; 29(11):3513–27. https://doi.org/10.1093/molbev/mss169 PMID: 22729749

38. Ko Albert M-S, Chen C-Y, Fu Q, Delfin F, Li M, Chiu H-L, et al. Early Austronesians: Into and Out Of Taiwan. Am J Hum Genet. 2014; 94(3):426–36. https://doi.org/10.1016/j.ajhg.2014.02.003 PMID: 24607387

39. Woravatin W, Stoneking M, Srikummool M, Kampuansai J, Arias L, Kutanan W. South Asian maternal and paternal lineages in southern Thailand and the role of sex-biased admixture. PLoS ONE. 2023; 18 (9):e0291547. https://doi.org/10.1371/journal.pone.0291547 PMID: 37708147

40. Delfin F, Min-Shan Ko A, Li M, Gunnarsdóttir ED, Tabbada KA, Salvador JM, et al. Complete mtDNA genomes of Filipino ethnolinguistic groups: a melting pot of recent and ancient lineages in the Asia-Pacific region. Eur J Hum Genet. 2014; 22(2):228–37. https://doi.org/10.1038/ejhg.2013.122 PMID: 23756438

41. Poznik GD. Identifying Y-chromosome haplogroups in arbitrarily large samples of sequenced or genotyped men. 2016. https://doi.org/10.1101/088716

42. Delfin F, Salvador JM, Calacal GC, Perdigon HB, Tabbada KA, Villamor LP, et al. The Y-chromosome landscape of the Philippines: extensive heterogeneity and varying genetic affinities of Negrito and non-Negrito groups. Eur J Hum Genet. 2011; 19(2):224–30. https://doi.org/10.1038/ejhg.2010.162 PMID: 20877414

43. Trejaut JA, Poloni ES, Yen J-C, Lai Y-H, Loo J-H, Lee C-L, et al. Taiwan Y-chromosomal DNA variation and its relationship with Island Southeast Asia. BMC Genet. 2014; 15(1):77. https://doi.org/10.1186/1471-2156-15-77 PMID: 24965575

44. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2022.

45. Paradis E. pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics. 2010; 26(3):419–20. https://doi.org/10.1093/bioinformatics/btp696 PMID: 20080509

46. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2019; 35(3):526–8. https://doi.org/10.1093/bioinformatics/bty633 PMID: 30016406

47. Excoffier L, Lischer HE. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour. 2010; 10(3):564–7. https://doi.org/10.1111/j.1755-0998.2010.02847.x PMID: 21565059

48. Venables W, Ripley B. Modern Applied Statistics with S. Fourth ed: Springer New York, NY; 2002.

49. Oksanen J, Simpson G, Blanchet FG, Kindt R, Legendre P, Minchin P, et al. vegan: Community Ecology Package. 2022.

50. Nenadic O, Greenacre M. Correspondence Analysis in R, with two- and three-dimensional graphics: The ca package. Journal of Statistical Software. 2007; 20(3):1–13.

51. Oota H, Settheetham-Ishida W, Tiwawech D, Ishida T, Stoneking M. Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. Nat Genet. 2001; 29(1):20–1. https://doi.org/10.1038/ng711 PMID: 11528385

52. Dang NV, Chu TS, Luu H. Ethnic Minorities in Vietnam: The Gioi; 2017.

53. Grant A, Sidwell P. Editor's preface. In: Grant A, Sidwell P, editors. Chamic and beyond: studies in mainland Austronesian languages: Pacific Linguistics, The Australian National University; 2005. p. ix–xvii.

54. Blust RA. From Ancient Cham to Modern Dialects: Two Thousand Years of Language Contact and Change (review). Ocean Linguist. 2000; 39(2):435–45. https://doi.org/10.1353/ol.2000.0014

55. Liu D, Ko AM-S, Stoneking M. The genomic diversity of Taiwanese Austronesian groups: implications for the "Into and Out of Taiwan" models. PNAS Nexus. 2023; 2(5):pgad122. https://doi.org/10.1093/pnasnexus/pgad122 PMID: 37200801

56. Yamagata M, Matsumura H. Austronesian Migration to Central Vietnam: Crossing over the Iron Age Southeast Asian Sea. In: Piper PJ, Matsumura H, Bulbeck D, editors. New Perspectives in Southeast Asian and Pacific Prehistory: ANU Press; 2017. p. 333–56.

57. Jordan FM, Gray RD, Greenhill SJ, Mace R. Matrilocal residence is ancestral in Austronesian societies. Proc Biol Sci. 2009; 276(1664):1957–64. https://doi.org/10.1098/rspb.2009.0088 PMID: 19324748

58. Wilkins JF. Unraveling male and female histories from human genetic data. Curr Opin Genet Dev. 2006; 16(6):611–7. https://doi.org/10.1016/j.gde.2006.10.004 PMID: 17067791

59. Hickey GC. Village in Vietnam. American Behavioral Scientist. 1964; 8(3):27–9. https://doi.org/10.1177/000276426400800309