European Journal of Human Genetics
https://doi.org/10.1038/s41431-019-0557-4

ARTICLE

The paternal and maternal genetic history of Vietnamese populations

Enrico Macholdt[1] · Leonardo Arias[1] · Nguyen Thuy Duong[2] · Nguyen Dang Ton[2] · Nguyen Van Phong[2] · Roland Schröder[1] · Brigitte Pakendorf[3] · Nong Van Hai[2] · Mark Stoneking[1]

Received: 13 June 2019 / Revised: 14 October 2019 / Accepted: 17 November 2019

Abstract

Vietnam exhibits great cultural and linguistic diversity, yet the genetic history of Vietnamese populations remains poorly understood. Previous studies focused mostly on the majority Kinh group, and thus the genetic diversity of the many other groups has not yet been investigated. Here we analyze complete mtDNA genome sequences and ~2.3 Mb sequences of the male-specific portion of the Y chromosome from the Kinh and 16 minority populations, encompassing all five language families present in Vietnam. We find highly variable levels of diversity within and between groups that do not correlate with either geography or language family. In particular, the Mang and Sila have undergone recent, independent bottlenecks, while the majority group, Kinh, exhibits low levels of differentiation with other groups. The two Austronesian-speaking groups, Giarai and Ede, show a potential impact of matrilocality on their patterns of variation. Overall, we find that isolation, coupled with limited contact involving some groups, has been the major factor influencing the genetic structure of Vietnamese populations, and that there is substantial genetic diversity that is not represented by the Kinh.

Introduction

Southeast Asia (SEA) is a melting pot of ethnolinguistic diversity shaped by many demographic events, beginning with the initial arrival of anatomically modern humans at least 65 kya [1, 2], and including migrations accompanying the spread of agriculture, in particular rice and millet farming, the expansion of the Austronesian (AN) language family, and movements of Tai-Kadai (TK) and Hmong-Mien (HM) speakers [3]. The languages spoken in SEA today belong to five language families: Austro-Asiatic (AA), AN, HM, Sino-Tibetan (ST), and TK. Geographically SEA is divided into two subregions, Island SEA (ISEA) and Mainland SEA (MSEA).

Vietnam is a multiethnic country that occupies a key position within MSEA and exhibits both geographic and ethnolinguistic diversity. The northern part of the country consists of highlands and the Red River delta; the central part also comprises highlands, while the southern part encompasses mostly coastal lowlands and the Mekong River delta. There are 54 official ethnicities in Vietnam, and a total of 109 different languages are spoken in the country. These belong to all of the five major language families present in SEA [4]. Groups speaking AA languages are distributed throughout the country, while those speaking TK, HM, or ST languages historically were found mainly in the north but are now also living in other areas; AN-speaking groups are located in the south part of central Vietnam and the Tay Nguyen highlands. The AA language family is considered the oldest within the area; AA languages are scattered across MSEA and South Asia, and the location of the AA homeland is under debate [5]. The AA languages are associated with a major occupation of MSEA after the introduction of agriculture [6].

Supplementary information The online version of this article (https://doi.org/10.1038/s41431-019-0557-4) contains supplementary material, which is available to authorized users.

✉ Nong Van Hai
vhnong@igr.ac.vn

✉ Mark Stoneking
stoneking@eva.mpg.de

[1] Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, D04103 Leipzig, Germany

[2] Institute of Genome Research, Vietnam Academy of Science and Technology, 18 Hoang Quoc Viet, Cau Giay, Hanoi, Vietnam

[3] Laboratoire Dynamique du Langage, UMR5596, CNRS & Université de Lyon, Lyon, France

Published online: 11 December 2019

AN speakers are found all over ISEA and Oceania, and trace at least a part of their ancestry to aboriginal Taiwanese AN-speaking populations, supporting a start of the AN expansion out of Taiwan about 4 kya [3, 7]. The genetic composition of modern AN speakers in ISEA is heterogeneous; AN speakers in western Indonesia have substantial AA-related ancestry, caused most likely by a movement of AN speakers through MSEA mixing with AA speakers in Vietnam or peninsular Malaysia [3], while AN speakers in eastern Indonesia harbor both Papuan and AN-related ancestry [8]. AN speakers in MSEA include the Cham, Chru, Raglai, Giarai, and Ede from the south part of central Vietnam and Tay Nguyen highlands, Cham of Cambodia, and Malay groups in Malaysia and Thailand. In contrast to the predominant patrilocal residence pattern of other groups, AN groups are thought to have had an ancestral matrilocal residence pattern [9]. The TK and HM languages likely originated in the area of present-day southern China and north Vietnam then spread by multiple migrations southward to what is now Thailand, Laos, and other parts of Vietnam hundreds to thousands of years ago [10, 11]. Whether the current distribution of these languages and the farming culture across MSEA was a result of human migration events (demic diffusion) or happened without the major movement of people (cultural diffusion) is still highly disputed. MtDNA variation in Thailand supports a model of demic diffusion of TK speakers [12], while recent studies based on ancient DNA provide further evidence for Neolithic and Bronze age migrations from East Asia [13], and explain present-day SEA populations as the result of admixture of early mainland Hòabìnhian hunter-gatherers and several migrant groups from ancient East Asia associated with speakers of the AN, AA, and TK languages [14]. ST languages are assumed to have diverged 5.9 kya in Northern China, and ST speakers are thought to have migrated southward into the area of MSEA about 3 kya [15, 16].

While several genetic studies have focused on SEA, research on the ethnic groups in Vietnam remains rather limited [17–23]. Most of these studies either focused solely on the majority group in Vietnam, the Kinh, as representative of the entire country, or are based on a restricted number of SNPs, microsatellites, or only partial sequencing of mtDNA. Because the Kinh comprise 86% of the population, sampling individuals from this group is a promising way to capture the main signal of genetic diversity in Vietnam. But the complicated history of SEA indicates that there might be hidden complexity and genetic structure in the minority populations. We have therefore initiated a comprehensive study of the genetic history of Vietnamese ethnolinguistic groups. Here, we analyze sequences of full mtDNA genomes and ~2.3 million bases of the male-specific portion of the Y chromosome (MSY) of the Kinh and 16 minority groups, encompassing all five language families, to investigate their maternal and paternal genetic structure. We use the genetic results based on our extensive sampling to investigate whether the genetic composition of the Kinh is a valid representation of all populations living in Vietnam today, and we assess the impact of geographic, linguistic, and cultural factors (i.e., postmarital residence pattern) on the genetic structure of Vietnamese populations.

Material and methods

Sample information

We analyzed DNA from 600 male Vietnamese individuals (Supplementary Material Table S1) belonging to 17 ethnic groups that speak languages belonging to the five major language families in Vietnam. In detail the data set consists of two AA speaking groups (Kinh and Mang), five TK speaking groups (Tay, Thai, Nung, Colao, and Lachi), two AN speaking groups (Giarai and Ede), three HM speaking groups (Pathen, Hmong, and Dao), and five ST speaking groups (Lahu, Hanhi, Phula, Lolo, and Sila). The average sampling locations per population are shown in Fig. 1. Ethnic groups sampled for this project, name, language affiliation, and census size were based on the General Statistics Office of Vietnam (www.gso.gov.vn and the 2009 Vietnam Population and Housing census, accessed September 2018) and the Ethnologue [4]. All sample donors gave written informed consent, and this research received ethical clearance from the Institutional Review Board of the Institute of Genome Research, Vietnam Academy of Science and Technology (no. 4-2015/NCHG-HDDD), and from the Ethics Commission of the University of Leipzig Medical Faculty.

MtDNA sequencing

DNA was extracted from blood using the Qiagen Tissue and Blood extraction kit. Double-stranded, double bar-coded Illumina sequencing libraries were constructed as described previously [24]. The libraries were enriched for mtDNA sequences via in-solution capture [25]. The mtDNA haplogroups were defined using haplogrep2 [26], and the phylogeny and terminology of Phylotree 17 [27], http://www.phylotree.org. For further experimental details see [28]. The complete mtDNA data set can be found at GenBank MH448947–MH449555.

Y-chromosome sequencing

We enriched for ~2.3 million bases of the MSY from the same libraries used for mtDNA capture enrichment [28].

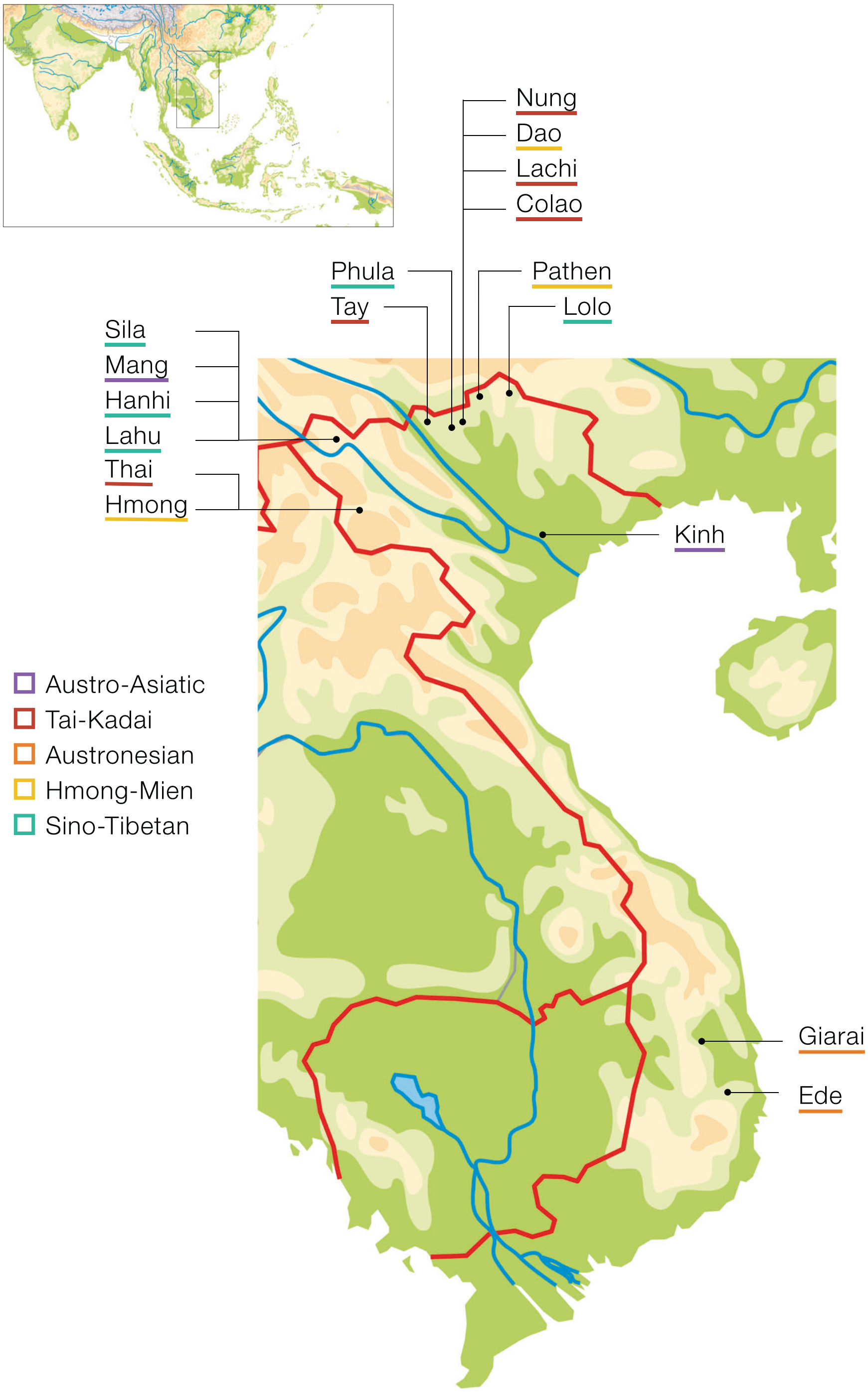

Fig. 1 Map of sampling locations. Dots show average sampling locations per population. Population labels are color coded by language family with Austro-Asiatic in purple, Tai-Kadai in red, Austronesian in orange, Hmong-Mien in yellow, and Sino-Tibetan in turquoise.

The MSY sequence processing pipeline is described elsewhere [29]. To increase the quality of the data set, we removed 41 positions with more than 16.6% missing information across the Vietnam MSY sequences and then imputed the remaining missing genotypes with BEAGLE 4.0 [30] using published reference sequences (Supplementary Material Table S2) from South Asia, East Asia, SEA, and Oceania [20, 31, 32]. After making initial haplogroup calls (using the procedure described below), we then went back and added additional reference samples from haplogroups C2-M217 and N1-M2291 for imputation, as these were present in 28 and 22 individuals respectively in the Vietnamese sample set but not present in the initial set of reference sequences used for imputation. Further, a merged A00 sequence [31] was added as an outgroup. From the combined data set we additionally excluded 53 positions not covered by more than 75% of the samples. The aligned MSY reads are deposited in the European Nucleotide

Archive (PRJEB33028). Final SNP genotypes and their chromosomal positions on hg19 are provided in Supplementary Material Data 1 and Data 2. MSY haplogroups were called with yhaplo [33] using a stopping condition parameter "ancStopThresh" = 10. Haplogroups were typed to the maximum depth possible given the phylogeny of ISOGG version 11.04 (http://www.isogg.org/) and the available genetic markers in our target region. Labels denoted with an asterisk in the text and figures are paragroups that do not include subgroups.

Sequence analysis

For both markers, we calculated the mean number of pairwise differences by averaging over the sum of nucleotide differences for each pair of sequences within a population (R function: dist.dna package: ape) divided by the total number of pairs. The nucleotide diversity (π) and its variance were computed using the R function nuc.div (package:pegas). Because the MSY information is based on a set of linked SNPs without recombination, we use the term haplotype throughout this paper to refer to the MSY sequences, and not to STR profiles as has been done in the past. We calculated the number of unique haplotypes for each population and obtained the haplotype diversity (*H*) values using Arlequin version 3.5.2.2 [34]. To visualize π and *H* we calculated the percentage difference from the mean for each population. Arlequin version 3.5.2.2 [34] was additionally used to calculate the pairwise genetic distances (Φ_{ST} distances) among the populations and the analyses of molecular variance (AMOVA) for both markers. The *p* values of the genetic distances were corrected for multiple testing by applying the Benjamini–Hochberg procedure. The Φ_{ST} distances were used to compute nonmetric multidimensional scaling (MDS) plots. We created two-dimensional projections (R function: isoMDS package: MASS) and calculated heatplots with five dimensions, showing per-dimension standardized values between 0 and 1. We calculated Mantel matrix correlation tests between genetic distances and great circle distances of the average geographical location per population using Pearson's correlation with 10,000 times random resampling. The correspondence analyses were computed in R using the libraries "vegan", "fields," and "ca". The haplotype sharing analysis was based on sequence haplotypes via string matching. We excluded Ns and indels for the mtDNA sequences; this step was not necessary for the MSY sequences because indels were not called and there were no Ns after imputation.

We performed mtDNA and MSY Bayesian analysis with BEAST 1.8 [35]. The software jmodeltest2 [36] was used to determine that the HKY + I + G and GTR models were the best substitution models for the mtDNA and MSY sequences, respectively. We partitioned the mtDNA genomes into the coding and noncoding sections and applied previously published and widely used mutation rates [37] of 1.708×10^{-8} and 9.883×10^{-8} mutations/site/year, respectively. This partitioning was supported by PartitionFinder2 [38]. For all MSY analyses the MSY mutation rate of 0.871×10^{-9}, based on an Icelandic pedigree [39], was applied. Because of the uncertainty in MSY mutation rates [40] and to provide a comprehensive comparison, we additionally applied an ancient DNA calibrated MSY mutation rate (0.76×10^{-9} substitutions/bp/year [41]) to MSY results where date estimates are relevant, as reported in the Supplementary Material Text. A Bayes factor analysis including marginal likelihood estimations [42] was used to test different clock models. We applied the Bayesian skyline piecewise linear tree prior for the dating and Bayesian skyline generation, so as to allow for population size changes over time. To ensure successful Bayesian estimation and to reach ESS values above 200, we combined multiple MCMC runs with 100 million steps using the BEAST logcombiner with a resampling up to ~40,000 trees.

We constructed MP trees for both markers and counted the mutations from the outgroup per sample. The mutation counts were used to compare the average distance of macrohaplogroups to the base of the trees as a measurement of branch length heterogeneity. We tested for significant differences in branch length distributions of major haplogroups with the Mann–Whitney *U* test.

Results

MSY sequences

We sequenced 2,346,049 bases of the MSY of 600 Vietnamese from 17 populations to a mean coverage of 30.2 × (minimum: 5×, maximum: 72×). After filtering, there were 3932 SNP positions, including 1908 novel sites that have not been described previously (dbSNP Build 153, accessed October 9, 2019, http://www.ncbi.nlm.nih.gov/SNP/). Fifty-seven different haplogroups were present in the 17 populations (Table S3). A detailed analysis of the phylogeography of the MSY data set will be presented elsewhere; the focus of the present study is the comparison of patterns of mtDNA and MSY variation in the sampled Vietnamese populations, and so we only briefly mention some interesting features about the MSY haplogroup distribution (Supplementary Material Text 1, Fig. S1, Tables S3, S4).

Genetic diversity within populations

The nucleotide diversity (π) and haplotype diversity (*H*) for mtDNA and MSY sequences varied substantially among

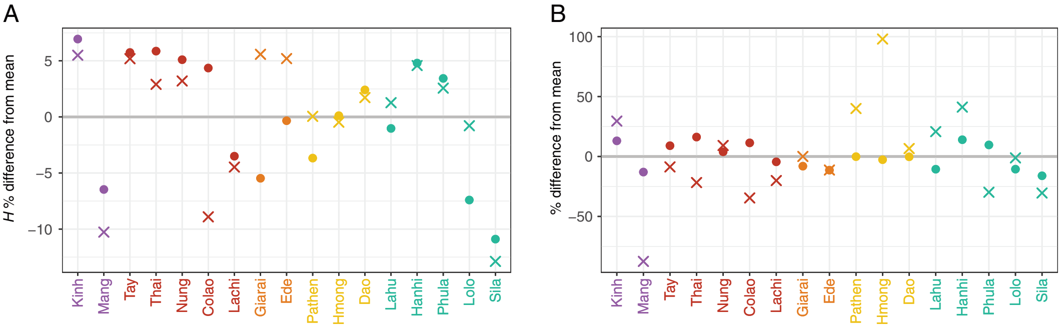

Fig. 2 Diversity statistics shown as the percent difference from the mean. a The haplotype diversity (H) and **b** the nucleotide diversity (π). Crosses and dots denote the MSY and mtDNA values, respectively. Population labels are color coded by language family with Austro-Asiatic in purple, Tai-Kadai in red, Austronesian in orange, Hmong-Mien in yellow, and Sino-Tibetan in turquoise. The gray line shows the mean across populations.

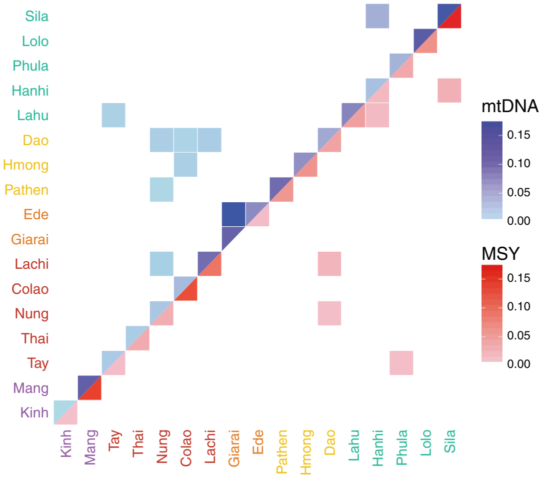

Fig. 3 Frequency of shared haplotypes between populations. mtDNA (upper triangle) and MSY (lower triangle) shared haplotype frequencies are represented by the blue and red color scale, respectively. White squares indicate no sharing. Population labels are color coded by language family with Austro-Asiatic in purple, Tai-Kadai in red, Austronesian in orange, Hmong-Mien in yellow, and Sino-Tibetan in turquoise.

populations (Fig. 2, Table S5). Kinh had high values of both π and H for both genetic markers, compared with the mean across populations. Sila, Lachi, and Mang had lower than average H values for both markers. The Mang MSY π-value was particularly low (5.91×10^{-6}) compared with the mean (4.72×10^{-5}), reflecting the unusual MSY haplogroup composition of this group, dominated by haplogroup O1b-B426, which had a frequency of 97%. The HM groups Pathen and Hmong had higher than average π values for the MSY, which reflects the higher frequency of C and D haplogroup sequences in these two groups (Table S3). The two AN groups were notable in having substantially higher than average MSY H values but average or below average mtDNA H values. Overall, the variation in H and π values for both markers was not consistent within language families. We found high frequencies of both mtDNA and MSY haplotype sharing, up to 17% (Fig. 3), within all 17 populations. In total, 28.5% of the mtDNA and 24.7% of the MSY haplotypes were shared within at least one population. All groups shared mtDNA types within the population, and only the Giarai lacked any shared MSY haplotypes within the population. The highest frequencies of within-population MSY haplotype sharing were present among the Mang (0.15), Sila (0.17), and Colao (0.14), in keeping with their very low H values (Fig. 2).

Population relationships

We examined haplotype sharing between populations as an indication of recent genetic contact or shared ancestry (Fig. 3). In general, there were more occurrences of sharing between populations in the mtDNA sequences compared with the MSY sequences, but only 5.5% of the mtDNA and 2.8% of the MSY haplotypes were shared between populations. The two AN populations (Giarai and Ede) had the highest frequency of mtDNA haplotype sharing between them (0.16), which was even higher than their within-population sharing (0.10 and 0.07). However, in notable contrast to the high degree of mtDNA sharing, there was no MSY haplotype sharing between the AN groups. The three HM populations each shared mtDNA haplotypes with some TK populations, but not with other populations of their own language family. The Dao (HM) shared MSY haplotypes with the TK groups Nung and Lachi; they are located geographically close to one another and also shared mtDNA haplotypes. All other occurrences of MSY haplotype sharing involved populations that are settled in close geographical proximity, namely Sila-Hanhi, Lahu-Hanhi, and Phula-Tay. To visualize the relationships among Vietnamese populations, we generated MDS plots (Fig. 4) from the matrices of pairwise Φ_{ST} distances (Supplementary Material Fig. S2) for both mtDNA and the MSY. Sila and Mang exhibited large distances to all populations, which explains their position

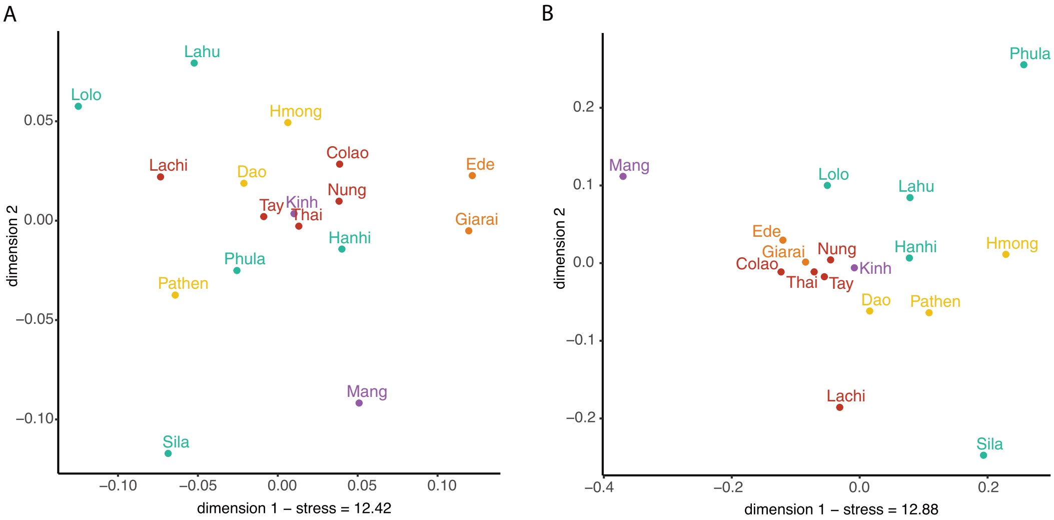

Fig. 4 MDS plots based on Φ_{ST} distances. a mtDNA and **b** MSY. Stress values are in percent. Population labels are color coded by language family with Austro-Asiatic in purple, Tai-Kadai in red, Austronesian in orange, Hmong-Mien in yellow, and Sino-Tibetan in turquoise.

in the periphery of the MDS plots (Fig. 4a, b). In addition, Phula (ST) stands out in the MSY MDS plot. Giarai and Ede showed large Φ_{ST} distances for mtDNA but not for the MSY (Fig. 4), and larger Φ_{ST} distances to most ST and HM populations than to Kinh (AA) and TK (except Lachi) groups. The Kinh group showed overall low genetic distances with other groups (Supplementary Material Fig. S2) and a central position in both MDS plots (Fig. 4). Because the rather high stress values of the two-dimensional MDS plots (Fig. 4a, b) indicated potentially more complex structure, we calculated a five-dimensional MDS and depicted the results in a heat plot (Supplementary Material Fig. S3A and B). The Kinh, Thai, and Tay remained centrally located across all five dimensions for both markers (Supplementary Material Fig. S3A and B), while the Mang remain an outlier in most dimensions in the MSY plot (Supplementary Material Fig. S3B).

We additionally explored population relationships based on haplogroup frequencies via a correspondence analysis (Supplementary Material Fig. S4). The results were similar to the MDS results for the AN groups, in that the Giarai and Ede were outliers for mtDNA but not for the MSY. Their mtDNA separation, in the first dimension of the plot, was driven mainly by the high frequencies of haplogroup M71 + C151T (37–42%), as well as by other exclusive haplogroups such as M68a1a, F1a4a, M21b, M24b, M68a1a, M73b, M74b1, M7b1a1f, and R9b1a1a. The second dimension separated the Mang (haplogroup A, D4, M71) and Pathen (A14, F1d, F2a). For the MSY, the high frequencies of F-M89 separated Phula (74%) and Lahu (32%) from the rest in the first dimension, while the second dimension spreads the populations between Mang and Hmong (Fig. S4).

Factors influencing the genetic structure

To test for correspondence between linguistic affiliation or geographic location with genetic structure, we analyzed three groupings (linguistic affiliation and two levels of geographical proximity) via AMOVA (Table 1). With respect to geography, we grouped populations on a broad scale by regions (political units), and on a finer scale by their origin in the same or neighboring districts. All three tested grouping patterns (language family, district, and region) revealed that ~90% of the total mtDNA variation and ~77% of the total MSY variation is explained by the differences within populations. Although the among-group component was significant for language family (1.8%) and districts (2.6%) for the mtDNA sequences, and for language family (4.4%) for the MSY sequences, in all of these cases the within-group component was considerably larger, indicating that differences between populations assigned to the same group were bigger than differences between populations assigned to different groups.

To further assess the impact of geography on the genetic structure of Vietnamese populations, we tested for correlations between the geographic and genetic distances for both mtDNA and MSY sequences (Table 2). This analysis was carried out for the entire data set, for a subset excluding the Kinh, and for a subset excluding the Ede and Giarai. We excluded the Kinh to control for the influence of their geographically widespread distribution and potentially mixed gene pool, as they are the majority group in Vietnam. Giarai and Ede, the only two groups from the Central Highlands of Vietnam, were excluded to test for a potential bias caused by their unique geographic position, as including these groups results in a bimodal distribution of

Table 1 AMOVA results.

Grouping	Number of groups	Marker	Percent variation explained		
			Among groups	Within groups among populations	Within populations
–		mtDNA	–	9.95***	90.05***
Language family	5	mtDNA	1.2	8.9***	89.9***
District	7	mtDNA	2.5*	7.8***	89.7***
Region	4	mtDNA	0.02	9.9***	90.1***
–		MSY	–	22.8***	77.2***
Language family	5	MSY	4.4*	18.8***	76.8***
District	7	MSY	0.5	22.3***	77.2***
Region	4	MSY	−1.0	23.5***	77.5***

Populations included in each group for each classification are indicated below the table. Language family: (Kinh, Mang) (Thai, Tay, Nung, Lachi, Colao) (Ede, Giarai) (Hmong, Dao, Pathen) (Sila, Lolo, Phula, Lahu, Hanhi). Region: (Mang, Gelao, Lachi, Lolo, Nung, Pathen, Dao) (Tay, Thai, Hanhi, Hmong, Lahu, Phula, Sila) (Kinh) (Giarai, Ede). Districts: (Lolo) (Giarai, Ede) (Colao, Dao, Lachi, Nung, Tay) (Kinh) (Hanhi, Lahu, Mang, Sila) (Hmong, Thai) (Pathen, Phula)

*p value < 0.05; ***p value < 0.001

Table 2 Correlation coefficients obtained in the Mantel correlation tests.

	All populations	Excluding Kinh	Excluding Ede and Giarai
mtDNA—geography	0.26*	0.24*	−0.22
MSY—geography	−0.07	−0.10	−0.24
mtDNA—MSY	0.17*	0.11	0.21*

*p value < 0.05

geographical distances. We found a significant correlation between the geographic distance and the mtDNA genetic distance matrices when analyzing the entire data set and the population subset excluding the Kinh (Table 2). However, the correlation between mtDNA distances and geography became nonsignificant when excluding the two Central Highlands groups (Ede and Giarai), suggesting that their large geographic distance (Fig. 1) and high mtDNA genetic distances (Fig. S2) from the other groups was driving the significant correlation. Furthermore, no significant correlation was detected for any comparisons of MSY genetic and geographic distances. However, there was a significant correlation between the genetic distance matrices of the two uniparental markers (Table 2) when the Kinh were included.

We additionally carried out maximum parsimony and Bayesian analyses of the MSY sequences and tested for branch length heterogeneity; the results of these analyses (which indicate substantial branch length heterogeneity in the MSY tree, but not the mtDNA tree) are discussed in Supplementary Material Texts 2 and 3.

Discussion

Previous genetic studies of Vietnam have focused largely on the Kinh majority as representative of the country [14, 18, 19, 21–23, 43, 44]. In contrast, we have investigated the patterns of genetic variation in a large sample of ethnolinguistic groups from Vietnam that speak languages encompassing all of the five language families present in the country. We found varying levels of genetic diversity within and among groups, not all of which is represented in the Kinh. For example, the high genetic diversity observed within the Kinh and several TK speaking groups differed substantially from the much lower levels of diversity found in small populations like the Sila and the Mang (Fig. 2). The predominant sharing of haplotypes within populations (Fig. 3), as well as this reduced diversity (Fig. 2), suggests that many of the Vietnamese groups are relatively isolated from one another. The higher number of population pairs that share mtDNA haplotypes, compared with MSY sharing, likely reflects recent contact and is in accordance with the expectation of more female exchange due to patrilocality in all non-AN groups. However, lower levels of MSY sharing could also reflect the greater resolution of MSY haplotypes based on 2.3 Mb of sequence, compared with the 16.5 Kb mtDNA genome sequences. Considering the difference in mutation rates for the different molecules [37, 39], a new mutation is expected in an MSY haplotype of our target size every 489.4 years, and in the whole mtDNA genome every 3624 years; thus, new mutations erase MSY sharing faster than mtDNA sharing. The proportion of MSY haplotypes that are shared within Vietnamese populations (24.2%) is much higher than observed in previous studies of the same MSY regions in other populations (7.1% for Thai/Lao populations [45], 13.7% for northwest Amazonian populations [29], and 6.9% for Angolan populations [46] (Supplementary Material Fig. S7).

There was no correspondence between the genetic structure and geography, as indicated by the absence of a significant correlation between genetic and geographic distances (Table 2), and the lack of any significant influence of geographical clusterings in the AMOVA (Table 1). While the mtDNA genetic structure is slightly influenced by geography, a significant correlation between mtDNA genetic distances and geographic distances disappears when the AN groups are excluded (Table 2). Because there are other factors that differ between AN groups and non-AN groups, such as the postmarital residence pattern (discussed in more detail in Supplementary Material Text 4), which might influence the genetic structure, geography is not necessarily directly related to genetic structure. We also did not find any evidence for an association between genetic structure and language family affiliation (Table 1). A consistent finding across Vietnamese populations was a higher female than male effective population size (Supplementary Material Text 2, Fig. S5), and more genetic structure on the MSY (MSY $F_{ST} = 23\%$ vs. mtDNA $F_{ST} = 10\%$) (Table 1). These are common patterns in human populations [31, 44, 47], and likely reflect a predominant patrilocal residence pattern and higher levels of female migration [48–50], a greater variance in reproductive success for males than for females [49, 51], and male-specific cultural inheritance of fitness [31]. A recent study of MSEA populations supports the negative influence of clan fission and extinction dynamics on Y-chromosome effective population size in patrilineal societies [52]. Strikingly, the Pathen (HM) mtDNA and MSY effective population sizes were about the same. Why this is the case is not known, but we speculate that this could reflect greater homogeneity in male reproductive success for the Pathen, compared with other Vietnamese groups. On a global scale, previous analyses have revealed consistently higher female than male population sizes, and increases in both at ~40–60 kya for all non-African continents [31]. While this signal of female population size increase was present in our study, the MYS lineages of all of the minority groups coalesced more recently than 40 kya. (Fig. S5). Previous results [31] further indicated a reduction in male Ne for continental populations between 8 and 4 kya, linked to the spread of Neolithic cultures and adoption of farming and changes in social structures, leading to an increase in the variance of male reproductive success and sex-specific demographic events. This trend was not found in the trajectories of most of our Vietnamese population, which may reflect the substantially lower structure present on the local compared with the continental scale.

Overall, it appears that the genetic structure of Vietnam has been primarily influenced by two main factors. The first is isolation and genetic drift, leading to high levels of genetic differentiation between groups and variable levels of genetic diversity within groups. Second, there has been limited recent contact between some groups, leading to some haplotype sharing. The levels of genetic differentiation among groups of 10% based on mtDNA and 23% based on the MSY (Table 1) are similar to what was found previously for populations from Northwestern Amazonia (13% mtDNA and 27% MSY; [29]) and higher than those found for Thai populations (8.5% mtDNA and 11% MSY; [45]). We also found particularly low levels of diversity in some specific groups, like the Mang and Sila (Fig. 2, Supplementary Material Table S5). The low levels of haplotype sharing for both markers (Fig. 3) are further evidence of isolation and limited contact between geographically close populations. The observed level of mtDNA haplotype sharing between Vietnamese groups (5.5%) is lower than that observed in most other studies of complete mtDNA genome sequences (Supplementary Material Fig. S6), while the MSY haplotype sharing between Vietnamese groups (2.2%) is also lower than what was observed in studies that sequenced the same regions of the MSY (Supplementary Material Fig. S7).

In addition to the general aspects of Vietnamese genetic diversity discussed above, our results provide some insights into the genetic profile and history of specific groups. These are discussed in detail in Supplementary Text Material Text 4, and include: the higher male than female isolation of HM groups; recent expansions and diversification of TK groups, along with some contact between them and HM groups; the impact of matrilocality on patterns of genetic variation in the AN groups; evidence for the probable incorporation of other groups into the Kinh during their initial spread; and the pronounced bottleneck (especially in the MSY sequences) in the Mang and Sila.

Conclusion

The sequencing of 2.34 Mb of the MSY chromosome and the complete mtDNA genome of 17 different ethnic groups (encompassing all five language families) enabled us to carry out the first comprehensive analysis of the genetic diversity within Vietnam. Overall, isolation leading to genetic drift has had an important impact on Vietnamese groups, with recent contact limited to some TK and HM groups. There are several differences between the maternal and paternal genetic history of some populations; in particular, a matrilocal vs. patrilocal residence pattern appears to be one of the major drivers of differences between the MSY and mtDNA signals in AN vs. other groups. There is also a profound impact of genetic drift for the Mang and Sila, especially in the MSY lineages, suggesting male-specific bottlenecks or founder events. And although we find genetic evidences for a central position of the Kinh as the majority group within the country, there is substantial genetic diversity in the other ethnic groups that is not represented in the Kinh. Genetic studies of the remaining ethnic groups in Vietnam, and expansion of the genetic data to include genome-wide variation will provide further insights into the genetic history of this key region of MSEA.

Acknowledgements We thank Alexander Hübner for computational advice, and Wibhu Kutanan and Dang Liu for helpful discussions. This research was supported by funds from the Max Planck Society and the Ministry of Science and Technology, Vietnam (grant DTDL. CN-05/15). BP is grateful to the LABEX ASLAN (ANR-10-LABX-0081) of Université de Lyon for its financial support within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) of the French government operated by the National Research Agency (ANR).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Westaway KE, Louys J, Awe RD, Morwood MJ, Price GJ, Jx Zhao, et al. An early modern human presence in Sumatra 73,000–63,000 years ago. Nature. 2017;548:322.
2. Demeter F, Shackelford L, Westaway K, Barnes L, Duringer P, Ponche J-L, et al. Early modern humans from Tam Pà Ling, Laos: fossil review and perspectives. Curr Anthropol. 2017;58:S527–38.
3. Lipson M, Loh P-R, Patterson N, Moorjani P, Ko Y-C, Stoneking M, et al. Reconstructing Austronesian population history in island Southeast Asia. Nat Commun. 2014;5:4689.
4. Eberhard DM, Gary FS, Charles DF. Ethnologue: languages of the world, twenty-second edition Dallas. Texas: SIL International; 2019. http://www.ethnologue.com.
5. Sidwell P, Blench R. The Austroasiatic urheimat: the Southeastern riverine hypothesis. In: Dynamics of human diversity. 2011. p. 315. Canberra: Pacific Linguistics.
6. Bellwood PS. First farmers: the origins of agricultural societies. 2005. Malden, MA: Blackwell.
7. 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;94:426–36.
8. Hudjashov G, Karafet TM, Lawson DJ, Downey S, Savina O, Sudoyo H, et al. Complex patterns of admixture across the Indonesian archipelago. Mol Biol Evol. 2017;34:2439–52.
9. Jordan FM, Gray RD, Greenhill SJ, Mace R. Matrilocal residence is ancestral in Austronesian societies. Proc R Soc B. 2009;276:1957–64.
10. Kutanan W, Kampuansai J, Brunelli A, Ghirotto S, Pittayaporn P, Ruangchai S, et al. New insights from Thailand into the maternal genetic history of Mainland Southeast Asia. Eur J Hum Genet. 2018;26:898–911.
11. Peng M-S, Quang HH, Dang KP, Trieu AV, Wang H-W, Yao Y-G, et al. Tracing the Austronesian footprint in Mainland Southeast Asia: a perspective from mitochondrial DNA. Mol Biol Evol. 2010;27:2417–30.
12. Kutanan W, Kampuansai J, Srikummool M, Kangwanpong D, Ghirotto S, Brunelli A, et al. Complete mitochondrial genomes of Thai and Lao populations indicate an ancient origin of Austroasiatic groups and demic diffusion in the spread of Tai-Kadai languages. Hum Genet. 2017;136:85–98.
13. Lipson M, Cheronet O, Mallick S, Rohland N, Oxenham M, Pietrusewsky M, et al. Ancient genomes document multiple waves of migration in Southeast Asian prehistory. Science. 2018;361:92–5.
14. McColl H, Racimo F, Vinner L, Demeter F, Gakuhari T, Moreno-Mayar JV, et al. The prehistoric peopling of Southeast Asia. Science. 2018;361:88–92.
15. Zhang M, Yan S, Pan W, Jin L. Phylogenetic evidence for Sino-Tibetan origin in northern China in the late Neolithic. Nature. 2019;569:112–5.
16. Bellwood P, Ness I. The Global Prehistory of Human Migration. Hoboken: Wiley; 2014.
17. The HUGO Pan-Asian SNP Consortium. Mapping human genetic diversity in Asia. Science. 2009;326:1541–5.
18. He J-D, Peng M-S, Quang HH, Dang KP, Trieu AV, Wu S-F, et al. Patrilineal perspective on the Austronesian diffusion in Mainland Southeast Asia. PLoS ONE. 2012;7:e36437.
19. Irwin JA, Saunier JL, Strouss KM, Diegoli TM, Sturk KA, O'Callaghan JE, et al. Mitochondrial control region sequences from a Vietnamese population sample. Int J Leg Med. 2008;122:257–9.
20. Poznik GD, Xue Y, Mendez FL, Willems TF, Massaia A, Wilson Sayres MA, et al. Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. Nat Genet. 2016;48:593–9.
21. Simonson TS, Xing J, Barrett R, Jerah E, Loa P, Zhang Y, et al. Ancestry of the Iban is predominantly Southeast Asian: genetic evidence from autosomal, mitochondrial, and Y chromosomes. PLoS ONE. 2011;6:e16338.
22. Soares PA, Trejaut JA, Rito T, Cavadas B, Hill C, Eng KK, et al. Resolving the ancestry of Austronesian-speaking populations. Hum Genet. 2016;135:309–26.
23. Li H, Cai X, Winograd-Cort ER, Wen B, Cheng X, Qin Z, et al. Mitochondrial DNA diversity and population differentiation in southern East Asia. Am J Phys Anthropol. 2007;134:481–8.
24. Meyer M, Kircher M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc. 2010;2010:prot5448. pdb

25. Maricic T, Whitten M, Paabo S. Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. PLoS ONE. 2010;5:e14004.
26. Weissensteiner H, Pacher D, Kloss-Brandstatter A, Forer L, Specht G, Bandelt HJ, et al. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic Acids Res. 2016;44:W58–63.
27. van Oven M, Kayser M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat. 2009;30:E386–94.
28. Nguyen TD, Macholdt E, Nguyen DT, Arias L, Schröder R, Phong NV, et al. Complete human mtDNA genome sequences from Vietnam and the phylogeography of Mainland Southeast Asia. Sci Rep. 2018;8:11651.
29. Arias L, Schröder R, Hübner A, Barreto G, Stoneking M, Pakendorf B. Cultural innovations influence patterns of genetic diversity in Northwestern Amazonia. Mol Biol Evol. 2018;35: 2719–35.
30. Browning BL, Yu Z. Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. Am J Hum Genet. 2009;85:847–61.
31. Karmin M, Saag L, Vicente M, Sayres MAW, Järve M, Talas UG, et al. A recent bottleneck of Y chromosome diversity coincides with a global change in culture. Genome Res. 2015;25: 459–66.
32. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature. 2016;538:201.
33. Poznik GD. Identifying Y-chromosome haplogroups in arbitrarily large samples of sequenced or genotyped men. 2016. https://doi.org/10.1101/088716.
34. Excoffier L, Lischer HEL. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour. 2010;10:564–7.
35. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012;29:1969–73.
36. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods. 2012;9:772.
37. Soares P, Ermini L, Thomson N, Mormina M, Rito T, Röhl A, et al. Correcting for purifying selection: an improved human mitochondrial molecular clock. Am J Hum Genet. 2009;84:740–59.
38. Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. Mol Biol Evol. 2016;34:772–3.
39. Helgason A, Einarsson AW, Guðmundsdóttir VB, Sigurðsson Á, Gunnarsdóttir ED, Jagadeesan A, et al. The Y-chromosome point mutation rate in humans. Nat Genet. 2015;47:453.
40. Jobling MA, Tyler-Smith C. Human Y-chromosome variation in the genome-sequencing era. Nat Rev Genet. 2017;18:485.
41. Fu Q, Mittnik A, Johnson Philip LF, Bos K, Lari M, Bollongino R, et al. A revised timescale for human evolution based on ancient mitochondrial genomes. Curr Biol. 2013;23:553–9.
42. Baele G, Li WLS, Drummond AJ, Suchard MA, Lemey P. Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. Mol Biol Evol. 2013;30:239–43.
43. Karafet TM, Mendez FL, Sudoyo H, Lansing JS, Hammer MF. Improved phylogenetic resolution and rapid diversification of Y-chromosome haplogroup K-M526 in Southeast Asia. Eur J Hum Genet. 2014;23:369.
44. Kayser M, Brauer S, Weiss G, Schiefenhövel W, Underhill P, Shen P, et al. Reduced Y-chromosome, but not mitochondrial DNA, diversity in human populations from West New Guinea. Am J Hum Genet. 2003;72:281–302.
45. Kutanan W, Hübner A, Macholdt E, Arias L, Schröder R, Stoneking M, et al. Contrasting paternal and maternal genetic histories of Thai and Lao populations. Mol Biol Evol. 2019;36:1490–1506.
46. Oliveira S, Hübner A, Fehn A-M, Aço T, Lages F, Pakendorf B, et al. The role of matrilineality in shaping patterns of Y chromosome and mtDNA sequence variation in southwestern Angola. Eur J Hum Genet. 2019;27:475–83.
47. Seielstad MT, Minch E, Cavalli-Sforza LL. Genetic evidence for a higher female migration rate in humans. Nat Genet. 1998;20:278.
48. 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;2:228.
49. Heyer E, Chaix R, Pavard S, Austerlitz F. Sex-specific demographic behaviours that shape human genomic variation. Mol Ecol. 2012;21:597–612.
50. Verdu P, Heyer E, Austerlitz F, Georges M, Becker NSA, Bahuchet S, et al. Sociocultural behavior, sex-biased admixture, and effective population sizes in Central African Pygmies and non-Pygmies. Mol Biol Evol. 2013;30:918–37.
51. Wilder JA, Kingan SB, Mobasher Z, Pilkington MM, Hammer MF. Global patterns of human mitochondrial DNA and Y-chromosome structure are not influenced by higher migration rates of females versus males. Nat Genet. 2004;36:1122.
52. Ly G, Alard B, Laurent R, Lafosse S, Toupance B, Monidarin C, et al. Residence rule flexibility and descent groups dynamics shape uniparental genetic diversities in South East Asia. Am J Phys Anthropol. 2018;165:480–91.