Reconstructing the formation of Hmong-Mien genetic fine-structure

Zi-Yang Xia[6,8]*, Xingcai Chen[5], Chuan-Chao Wang[6,7,8,9,10,]*, Qiongying Deng[1,2,3,4,]*

1. Key Laboratory of Human Development and Disease Research (Guangxi Medical University), Education Department of Guangxi Zhuang Autonomous Region, China
2. Department of Human Anatomy, Guangxi Medical University, Nanning, Guangxi Zhuang Autonomous Region, China
3. Institute of Neuroscience and Guangxi Key Laboratory of Brain Science, Guangxi Medical University, Nanning, Guangxi, China
4. Key Laboratory of Longevity and Aging-related Diseases of Chinese Ministry of Education, Guangxi Medical University, Nanning, Guangxi, China
5. Department of Human Anatomy, Guangxi Medical University, Nanning, Guangxi Zhuang Autonomous Region, China
6. State Key Laboratory of Cellular Stress Biology, School of Life Sciences, Xiamen University, Xiamen 361102, China
7. Ministry of Education Key Laboratory of Contemporary Anthropology, Department of Anthropology and Human Genetics, School of Life Sciences, Fudan University, Shanghai 200433, China
8. Department of Anthropology and Ethnology, Institute of Anthropology, School of Sociology and Anthropology, Xiamen University, Xiamen 361005, China
9. Key Laboratory of Western China's Environmental Systems (Ministry of Education), College of Earth and Environmental Sciences, Lanzhou University, Lanzhou 730000, China
10. State Key Laboratory of Marine Environmental Science, Xiamen University, Xiamen 361102, China

*Corresponding authors: Qiongying Deng, yingzideng@sina.com; Chuan-Chao Wang, wang@xmu.edu.cn; Zi-Yang Xia, ziyang.xia.20@alumni.ucl.ac.uk

Abstract

The linguistic, historical, and subsistent uniqueness of Hmong-Mien (HM) speakers offers a wonderful opportunity to investigate how these factors impact the genetic structure. Nevertheless, the genetic differentiation among HM-speaking populations and the formation process behind are far from well characterized in previous studies. Here, we generated genome-wide data from 67 Yao ethnicity samples and analyzed them together with published data, particularly by leveraging haplotype-based methods. We identify that the fine-scale genetic substructure of HM-speaking populations corresponds better to linguistic classification than to geography, while the parallel of serial founder events and language differentiations can be found in West Hmongic speakers. Multiple lines of evidence indicate that ~500-year-old GaoHuaHua individuals are most closely related to West Hmongic-speaking Bunu. The excessive level of the genetic bottleneck of HM speakers, especially Bunu, is in agreement with their long-term practice of slash-and-burn agriculture. The inferred admixture dates in most of the HM-speaking populations overlap the reign of the Ming dynasty (1368 – 1644 CE). Besides the common genetic origin of HM speakers, their external ancestry majorly comes from neighboring Han Chinese and Kra-Dai speakers in South China. Conclusively, our analysis reveals the recent isolation and admixture events that contribute to the fine-scale genetic formation of present-day HM-speaking populations underrepresented in previous studies.

Introduction

Consisting of ~40 different languages and dialects (Hammarström *et al.*, 2022), the Hmong-Mien (HM) language family is currently spoken by ~6 million people living in the vast mountainous area of South China and Mainland Southeast Asia (Ratliff, 2010) *[Ratliff, p3]*. As suggested by historical linguistics, the age of the proto-HM language is ~500 BCE and is supposed to be originally spoken in the middle Yangtze Basin (Ratliff, 2021). According to historical records, the suggested precursors of the HM speakers, *Wǔxī Mán* (五溪蠻, literally "*Mán* of the Five Streams"), lived in the

Wǔlíng Mountains (武陵山脈) between Hunan and Guizhou of Southwest China ~100-500 CE (Mao and Meng, 1986) *[p2]*. During the Tang (618-907 CE) and Song (960-1279 CE) dynasties, ethnonyms identical to present-day exonyms of HM-speaking groups, like Shē (畲/畬) and Yáo (徭/瑶), started to appear in Chinese historical documents (Zeng, 2005, Litzinger, 1995). Uprisings in Guizhou and Guangxi of Southwest China led by HM-speaking Miao and Yao people became more frequent during the reign of Ming (1368-1644 CE) and Qing (1644-1911 CE) dynasties (Scott, 2009) *[p137-140]*, which was followed by the migration of some HM-speaking groups to Mainland Southeast Asia ~1800 CE (Kutanan *et al.*, 2021).

Previous genomic studies have provided many intriguing insights into the population history of HM speakers, such as their shared genetic origin and admixture events during their past (Huang *et al.*, 2022, Xia *et al.*, 2019, Liu *et al.*, 2020, Kutanan *et al.*, 2021, Wang *et al.*, 2021b, Yang *et al.*, 2022). However, these studies focus more on the external genetic relationship of HM speakers to other groups (e.g., Sino-Tibetan and Kra-Dai speakers), while the formation history of genetic substructure within the HM speakers is obscure. For example, although the ~500-year-old GaoHuaHua population shares a similar genetic profile with the modern HM-speaking population (Wang *et al.*, 2021b), it still remains unclear whether and with which modern HM-speaking populations they are more closely genetically related. Besides, haplotype-based methods, especially chromosome painting-based ones, have hardly ever been applied in previous studies, which have been shown to have larger power to identify fine-scale genetic history than allele frequency-based ones (Lawson *et al.*, 2012, Leslie *et al.*, 2015). The haplotype-based methods could enable us to investigate previously unresolved aspects of the genetic history of HM-speaking populations, such as exact sources of non-HM ancestries and the relationship between fine-scale genetic structure and language classifications.

In cultural anthropology, HM-speaking communities have often been collectively studied together with other ethnolinguistic groups in Southeast Asian mainland massif,

also known as *Zomia* (Scott, 2009) *[preface, ix]*. Historically, the population in Zomia, including the majority of HM speakers, fell outside of the direct administration of the Sinicized and Indianized monarchies centralized at low elevations, where taxes and corvée labors were usually obligatory for the subjects (Scott, 2009) *[p13, p19, p116]*. In contrast to concentrated grain production (e.g., wet rice farming) practiced in the lowlands (Scott, 2009) *[p13]*, HM-speaking Miao and Yao people historically practiced shifting and slash-and-burn agriculture in mountains regions *[p194-195]* (Scott, 2009). In fact, the exonyms of HM-speaking She (literally "slash-and-burn") and Yao (originally *Mòyáo*/莫徭, "no corvée labor") come from their historical practice of slash-and-burn agriculture and stateless, respectively (Zeng, 2005, Ozawa, 2000). Many recent genomic studies have focused on the populations historically marginalized from the state system in Europe (e.g., Roma) (Font-Porterias *et al.*, 2019) and Africa (e.g., Ari Blacksmiths) (Van Dorp *et al.*, 2015). In comparison, the genetic consequence of the stateless social practices of the population in Zomia, especially HM speakers, has never been investigated previously. Besides the historically documented Sinicization of HM speakers, anthropologists have also proposed an undocumented transition of Han Chinese into HM-speaking communities (Scott, 2009) *[p125-126]*. Such theories are also promising to be assessed by using genetic data.

It is indisputable that the HM language family is formed by two major branches: Hmongic and Mienic (Ratliff, 2021). In China, all the HM speakers are categorized into one of the following three officially recognized ethnicities: Miao, Yao, and She (Wang, 1983b, Mao *et al.*, 1982, Mao and Meng, 1986). Yao is linguistically heterogeneous, including all the Mienic-speaking groups in China (except for Kin Mun in Hainan), Hmongic-speaking Bunu, Pahng, and Hmnai, Kra-Dai-speaking Lakkja, and currently Chinese-speaking Pingdi Yao (literally 'Yao in plains') (Simons and Fennig, 2017, Mao *et al.*, 1982) *[p5-9]*. Guangxi has the largest population and richest linguistic diversity of the Yao ethnicity (Mao *et al.*, 1982), but Yao populations in Guangxi have never been covered in previous genome-wide studies.

To overcome the aforementioned limitations in previous studies, we carried out this comprehensive investigation on the genetic differentiation within the HM speakers and the demographic and ancestral factors behind its formation, i.e., population isolation and admixture. We generated genome-wide SNP genotyping data from 67 Yao individuals from five Yao autonomous counties in Guangxi, whose geographic (Figure 1A) and linguistic (Figure 1B) information are shown in Figure 1. We addressed the three major issues throughout our study: What is the internal genetic substructure within the HM speakers, and how does it relate to ethnic identities (i.e., ethnic labels based on autonyms) and linguistic classifications? How does the social and cultural practice (e.g., shifting farming) in the HM history influence the pattern of population sizes across the HM-speaking populations? What is genetic interaction between HM and non-HM groups suggested by the inference of historical admixture events?

Results

An overview of population structure

We initially performed principal component analysis (PCA) to investigate the genetic variability of HM speakers in the contexts of southern East Asian populations (Figure 2), where ancient DNA (aDNA) samples were projected. In agreement with the previously reported language family-associated genetic structure of southern East Asians (Huang *et al.*, 2022), HM speakers, including the newly reported individuals, form a distinct gradient apart from other southern East Asian populations (Figure 2). Likewise, ~500-year-old GaoHuaHua individuals are positioned together with present-day HM speakers, separating from other ancient genomes from South China. We also confirmed this genetic structure in ADMIXTURE, as HM speakers and GaoHuaHua share a distinct genetic component when K = 10 (SI Figure 1).

Haplotype-based clustering

To further characterize the genetic substructure within the HM speakers, we applied

haplotype-based ChromoPainter and fineSTRUCTURE (Figure 3) exclusively for HM-speaking individuals (Lawson *et al.*, 2012). ChromoPainter infers each phased genotype as a mosaic of other most closely related genotypes, while fineSTRUCTURE leverages this haplotype-sharing pattern to cluster individual genotypes into a dendrogram. In general, individuals with the same autonym and linguistic affiliation tend to cluster together, suggesting that the marital practice might be more frequent in individuals with shared ethnic identities. Particularly, Yao from Du'an and Bama respectively group into two distinct genetic clusters ('Bunu' and 'Mienic3' for Yao_Duan, 'Mienic1' and 'Bunu' for Yao_Bama, respectively, Figure 3), consistent with the fact that both Hmongic (Bunu) and Mienic (Iu Mien and Kim Mun, respectively) languages are distributed in both localities (Figure 1B).

The hierarchical structure of genetic clusters partially reflects their linguistic classifications (Figure 3), which supports the parallel between language differentiation and population differentiation in the HM history to some extent. The highest level of the hierarchical tree separates Bunu and Hmong on the left from all the other HM speakers on the right (Figure 3). This indicates that Bunu and Hmong, both of whose languages fall under the classification of 'West Hmongic' or 'Chuanqiandian' with broad consensus (Strecker, 1987, Ratliff, 2010, Ratliff, 2021, Wang, 1983a) *[Ratliff 2010, p3]*, share an extra founder event in relation to other HM speakers, in addition to the founder event shared by all the HM speakers. Given the fact that Bunu and Hmong belong to different officially recognized ethnicities (Yao and Miao, respectively), linguistic affinity is more consistent with the shared genetic history than the classification of ethnicities for both groups. In the right meta-cluster, PaThen from Vietnam (Pahng) splits first from the others (Figure 3), which is in accordance with the fact that they are linguistically most distantly related to other Hmongic languages (Ratliff, 2010, Ratliff, 2021) *[Ratliff 2010, p3]*. Congruent with the fact that the Mienic Kim Mun language is spoken by both Yao_Bama (Figure 1B) and Dao from Vietnam (Scott, 2009) *[Scott 2010, p410]*, the majority of both groups are sister clusters in fineSTRUCTURE ('Mienic1', Figure 3). By contrast, geography

explains less for the genetic structure, as most sister clusters do not share close proximity (e.g., Hmong and Bunu, Dao and Yao_Bama1).

Nevertheless, instead of a sharp split between Hmongic- and Mienic-speaking populations, the genetic clusters with both linguistic affiliations intervene with each other. For example, Mienic-speaking Iu Mien from Thailand forms a sister cluster with North Hmongic-speaking (Ratliff, 2021) Miao (Qo Xong). Therefore, besides genetic isolations, admixture from external groups may also play an extra role in the genetic differentiation among HM-speaking populations. To increase the power of ancestral history inference, we separated all the HM-speaking individuals with the same ethnic labels but different fineSTRUCTURE clusters into distinct populations for all the following analyses, following López *et al.* (2021) (see Method).

Genetic origin of the ancient GaoHuaHua population

Since the ancient HM-related GaoHuaHua genomes are pseudo-haploid, we performed allele-frequency-based analyses to determine their genetic affinity to all the present-day HM-speaking populations (Figure 4).

In the PCA exclusively for GaoHuaHua and present-day HM-speaking populations (Figure 4A), PC1 separates West Hmongic-speaking Bunu and Hmong from all the other HM-speaking populations, while PC2 further makes PaThen stands out. GaoHuaHua individuals are projected together with West Hmongic speakers, particularly Bunu-speaking Yao_Duan1, Yao_Dahua, and Yao_Bama2. A consistent pattern can be observed in outgroup-f_3 (Figure 4B), where two of the three Bunu-speaking populations share the most genetic drift with GaoHuaHua (Yao_Dahua and Yao_Bama2). Then, we applied TreeMix to generate a maximum likelihood tree with no admixture event and rooted in Onge (Figure 4C). According to the phylogeny, Hmong speakers (Hmong, Hmong Daw, Hmong Njua) and Bunu speakers (Yao_Dahua, Yao_Duan1, Yao_Bama3) are sister clades, whereas GaoHuaHua belongs to the clade of Bunu speakers. Therefore, we conclude that

GaoHuaHua is genetically closer related to Bunu speakers than to any other present-day HM-speaking populations. The inferred genetic connection between GaoHuaHua and Bunu speakers is congruent with the fact that the funeral practices in GaoHuaHua sites are similar to present-day Baiku Yao (Wang *et al.*, 2021b, Zhang *et al.*, 1986), a Bunu-speaking group (Meng, 2001) [*p3*]. GaoHuaHua also provides a time calibration for serial founder events in the population history of HM speakers, as the divergence of the Hmong clade and the Bunu clade (Figure 4C) must precede the date of GaoHuaHua (1437–1656 CE) (Wang *et al.*, 2021b).

Demographic dynamics of HM population history

The distribution of shared haploblock in different lengths can provide pivotal insights into the pattern of effective population size (N_e) in population history, i.e., demographic history (Ceballos *et al.*, 2018, Ralph and Coop, 2013, Browning and Browning, 2013a, Palamara *et al.*, 2012, Ringbauer *et al.*, 2021). We sought to characterize the demographic history of HM speakers by using the pattern of shared haploblocks in three different levels: (1) parental relatedness within each individual; (2) individuals within a population; (3) in pairwise populations.

Runs of homozygosity (ROH) is a commonly used measurement for how the parents of an individual are genetically related to each other (Ceballos *et al.*, 2018, Ringbauer *et al.*, 2021). Theoretically, consanguineous marriage between relatives who share a common ancestor within a few generations tends to result in an excessive number of long ROHs. By contrast, a rapid decline of N_e (i.e., genetic bottleneck) tends to result in the sharing of relatively few ancestors in long-term history and an excessive number of short ROHs (Ceballos *et al.*, 2018). As instructed by Ceballos *et al.* (2018), we computed the average ROH of HM-speaking and other South Chinese populations (Figure 5A).

Compared with non-HM-speaking populations in South China, especially Kra-Dai-speaking sedentary lowland wet-rice farmers (e.g., Zhuang and Dai), most of

the HM-speaking populations have an excessive level of the total length of ROH (Figure 5A). The majority of HM-speaking populations, except for Dao_o and Pa Then, have an enrichment of short ROHs (2–10 cM) than long ROHs (> 10 cM), which indicates that their excessive ROH level is primarily a result of strong genetic bottlenecks in their population history rather than recent consanguineous marriage. Notably, the three Bunu-speaking populations (Yao_Dahua, Yao_Duan1, and Yao_Bama2) have the strongest level of ROH among all the Southern Chinese populations, which is even stronger than their sister clade, Hmong-speaking populations (Hmong, Hmong Daw, Hmong Njua).

We compared the average ROH of Bunu-speaking populations (126.1 cM, dash line, Figure 5B) in relation to chosen ancient and modern global populations with diverse modes of subsistence and marital practice (Figure 5B). In modern populations, the ROH level of Bunu speakers is comparable to Kalash, who is known for their high degree of genetic isolation (Ayub *et al.*, 2015, Hellenthal *et al.*, 2016), and even stronger than many hunter-gatherer populations in Africa (Mbuti, Biaka, and Jul'hoan) and Asia (Chukchi, Eskimo, Kusunda, Nivh, and Ulchi). As for the ancient genomes, the total ROH of ancient hunter-gatherers is ~ 0.6 – 2.1 folds as Bunu speakers, while one of Bunu speakers is 2–16-fold higher than ancient plant farmers (Anatolia_N, LBK_EN) and pastoralists (Steppe_EMBA). In conclusion, the extent of genetic isolation and bottleneck in Bunu speakers is more similar to the one of ancient and modern hunter-gatherers rather than neighboring sedentary lowland wet-rice farmers, which is likely due to their practice of slash-and-burn agriculture during their history.

We applied *IBD-Ne* to infer the change of Ne in HM-speaking populations over time from the distribution of identity-by-descent (IBD) (Figure 6A) (Browning and Browning, 2015), and we observed bottlenecks taking place in most of the HM-speaking populations. Except for Hmong Njua (1303 CE), the bottleneck time estimates for all the Bunu-speaking populations [Yao_Duan1, 1037 CE; Yao_Dahua, 1079 CE] and Hmong-speaking populations [Hmong, 1107 CE; Hmong Daw, 1107

CE] are very close to each other. Since such a time range of bottlenecks obviously predates the time of Bunu-related GaoHuaHua (1437–1656 CE) (Wang *et al.*, 2021b), this might reflect a shared bottleneck by all the West Hmongic speakers. Bottleneck events occurring around the South Song Dynasty (1127–1279 CE) are inferred in Miao (1275 CE) and Yao_Bama1 (1247 CE). The bottleneck time estimate for She (1471 CE) postdates the first occurrence of She in Fujian (1236 CE) (Chan, 2006), suggesting that the bottleneck of She might occur after their initial settlement in Fujian. More recent estimates of bottleneck events are observed in HM-speaking populations in Vietnam, i.e., Pa Then (1639 CE), Dao (1667 CE), as well as a second bottleneck for Hmong (1779 CE), which indicates the occurrence of founder events along with their migration from South China to Vietnam.

We then focused on the intra-population sharing of IBD segments in HM speakers (Figure 6B). Following the instruction by Ralph and Coop (2013), we classified the IBD segments shared by pairwise populations into three categories: 1–5 cM, 5–10 cM, and > 10 cM, which approximately correspond to the time range of 500 BCE – 500 CE, 500 CE – 1500 CE, and > 1500 CE, respectively. In ~500 BCE – 500 CE, each HM-speaking population tends to share common ancestors with all the other HM speakers. In ~500 CE – 1500 CE, HM-speaking populations tend to share more common ancestors with closely related clusters identified in fineSTRUCTURE, which roughly overlaps the frequent emergence of various ethnic groups related to present-day HM speakers (e.g., She and Yao) (Zeng, 2005, Scott, 2009, Litzinger, 1995). Since ~1500 CE, the majority of shared common ancestors are within the sister clusters in fineSTRUCTURE.

Admixture scenario of HM population history

Besides the variation of N_e over time, admixture with external populations is another factor that can contribute to the genetic differentiation among the HM-speaking populations. To this end, we verified and inferred the temporal scenario of admixture events in HM population history using the haplotype-based fastGLOBETROTTER

(Hellenthal *et al.*, 2014, Wangkumhang *et al.*, 2022). We used all three populations in the 'Hmong' cluster defined by fineSTRUCTURE (i.e., Hmong, Hmong Daw, and Hmong Njua, Figure 3, referred to as 'Hmong_all') to surrogate the shared ancestry by HM speakers, along with 120 ancient and modern global populations to surrogate other ancestries potentially contributing to present-day HM speakers genetically (SI Table 2). In addition to the other 14 HM-speaking populations, we also included 10 non-HM populations who likely received recent HM-related gene flow (SI Table 2) with the HM-related component > 5% in ADMIXTURE when K = 10 (SI Figure 1).

In all the 24 target populations, 16 of them show sufficiently strong statistical evidence for 'one-date' admixture between two primary ancestral sources (fit quality for one event ≥ 0.990, fit event quality for two events ≥ 0.999, goodness-of-fit r^2 = 0.316–0.859 for single admixture, Figure 6 and SI Table 3). In contrast, HM-speaking Dao_o, Yao_Duan2, Yao_Fuchuan, and Yao_Gongcheng do not show significant statistical evidence for recent admixture ('best guess conclusion' inferred as 'unclear signal', SI Table 3). Regarding the admixing sources, one of the primary sources is always inferred to be HM-like (Hmong Daw), while the other source is either Kra-Dai- (Hlai) or Southern Han Chinese-related (Han_Guangdong, Han_Fujian, or Han_Hubei, Figure 7B & SI Table 3). This can be explained by the fact that all the non-HM-speaking speakers inferred with this HM-related admixture are either Kra-Dai speakers affiliating to Kra (Gelao/Cờ Lao) or Kam-Sui (Dong_Hunan, Dong_Guizhou, and Maonan) or Tujia, who is historically Tibeto-Burman-speaking but largely Sinicized. The inferred average admixture dates for most of the populations (Figure 7A) overlap the more frequent presence of Miao and Yao in Chinese historical documents since the Ming Dynasty (1368–1644 CE) (Diamond, 1995, Faure, 2006). The inferred dates of admixture events in Bunu-speaking Yao_Dahua (997 CE, 95% confidence interval (CI) 791–1223 CE), Yao_Duan1 (1108 CE, 95% CI 832–1345 CE), and Yao_Bama2 (1334 CE, 95% CI 1089 – 1594 CE, Figure 7A) suggest that the genetic profile of Bunu speakers was largely fixed prior to or at the time of GaoHuaHua (1437–1656 CE). The inferred dates of IuMien (1826

CE, 95% CI 1538–1947 CE) and IuMien_o (1745 CE, 95% CI 1644 – 1866 CE) roughly coincide with their migration from South China to Southeast Asia.

To further characterize a global picture of the fine-scale ancestry composition in HM speakers, we applied SOURCEFIND (Chacón-Duque *et al.*, 2018) with the same target and surrogate sets as fastGLOBETROTTER analysis (SI Table 2). The variation in the proportion of the HM-like ancestry ('Hmong_all', Figure 8A) in HM-speaking populations majorly corresponds to their genetic clustering in fineSTRUCTURE (Figure 3): IuMien_o (81.5%), Bunu-speaking Yao_Dahua (67.9%), Yao_Bama2 (66.7%), and Yao_Duan1 (58.3%) harbour the largest proportion of this HM-like source, which is followed by PaThen (26.0%). Notably, Kra-speaking Cờ Lao (16.6%) and Gelao (9.7%), as well as Kam-Sui-speaking Dong_Guizhou (9.6%) and Dong_Hunan (11.6%), possess a comparable level of this HM-like ancestry as many HM speakers. Rather than a term specific HM speakers, the exonym 'Miao' is more likely to be a collective term for indigenous groups in Southwest China (or more exactly, in 'Miao territory'/Miao Jiang (苗疆) of Guizhou and western Hunan) who were out of the direct administration during Ming and Qing dynasties, which also includes Kra-Dai-speaking populations in this region. (Diamond, 1995, Scott, 2009) *[p110, p121]*. The inferred admixture dates of these populations (1346 – 1606 CE, Figure 7A) roughly fall within the reign of the Ming dynasty (1368 – 1644 CE), which suggests that massive admixture forming the 'HM Cline' described previously (Huang *et al.*, 2022) were among ethnic groups who were linguistically heterogeneous but with similar ethnic identity during this period.

Consistent with the pattern inferred by fastGLOBETROTTER, the majority of non-HM-like ancestry in HM speakers comes from either southern Han Chinese-like sources (Figure 8A), primarily surrogated by Han_Guangdong (Figure 8B), or Southeast Asian-like sources (Figure 8A), primarily surrogated by Kinh, Hlai, and Nung (Figure 8C). However, the distribution of these non-HM-like ancestries in HM-speaking populations, especially those in South China, varies geographically

(Figure 8A). All the Yao populations from localities in western Guangxi (i.e., Duan, Dahua, and Bama, Figure 1A) derive more of their non-HM-like ancestry from Southeast Asian-like sources than southern Han Chinese-like ones (Figure 8A). By contrast, for HM-speaking populations in eastern Guangxi (Yao_Gongcheng and Yao_Fuchuan, Figure 1A), eastern Guizhou (Miao, 28N, 109E), and northeast Fujian (She, 27N, 119E) (Cann *et al.*, 2002), their non-HM-like ancestry is primarily surrogated by southern Han Chinese-like sources (Figure 7A & B). Miao, Iu Mien, Gelao, and Cờ Lao derive additional Tibeto-Burman-like ancestry (> 5%, Figure 8A) primarily surrogated by Loloish-speaking Yi, Phù Lá, and Lahu.

We also observed special admixture histories in some HM-speaking populations. Congruent with their admixture date interval (1304 CE, 95% CI 963 – 1591 CE, Figure 7A), the largest proportion of ancestry surrogated by Han_Fujian (79.0%, Figure 8B) in She confirms that they acquired the majority of their southern Han Chinese-related ancestry only after their settlement in Fujian. Likewise, the extra genetic influx surrogated by Laos (36.0%, Figure 8C) is observed in Iu Mien currently living in Northeast Thailand (Kutanan *et al.*, 2021), which is compatible with the scenario of admixture during their migration out of South China. Conforming the results from fastGLOBETROTTER (SI Table 3), Yao_Fuchuan and Yao_Gongcheng – both of whom include both Iu Mien speakers and Pingdi Yao (Figure 1B) – have a negligible amount of HM-like ancestry (0.2% and 0.6%, Figure 8A) and a small amount of Southeast Asian-like ancestry (9.9% and 7.4%, Figure 8A), but they derive the majority of their ancestry surrogated by Han_Guangdong (88.9% and 90.9%, Figure 8B). Since SOURCEFIND infers a considerable amount of Southeast Asian-like ancestry in all the Kra-Dai-speaking populations (18.9 – 95.4%, Figure 8A), it is unlikely that all of the Han_Guangdong-related ancestry in both groups derive from Kra-Dai speakers, while at least some come from southern Han Chinese. Our result is compatible with the theory that some of the Yao people may have originated from Han Chinese who avoided tax and corvée and adopted the ethnic identity of 'Yao' (Scott, 2009) *[p121, p125]*.

Discussion

By leveraging newly reported genome-wide data and haplotype-based methods, our study extends the understanding of how historical isolation and admixture events shape the genetic structure of HM speakers to an unprecedented level of resolution. For example, we are able to identify the closely related genetic clusters in HM speakers (e.g., Yao_Bama1 and Dao, Figure 3) and the external gene flow from Han Chinese in a province level (e.g., Han_Guangdong vs Han_Fujian, Figure 8). The genetic structure in such a fine scale enables us to discuss further the genetic impact of the unique mode of subsistence that makes HM speakers distinctive from most of the other southern East Asian groups in history: slash-and-burn and shifting agriculture.

The practice of slash-and-burn agriculture tends to be associated with reduced population sizes. In South China and Mainland Southeast Asia, wet-rice agriculture is majorly practiced in lowland valleys where most of the easily cultivable land is distributed, so it promises high return per unit of land and high population density (Scott, 2009) *[p13, p41, p74]*. By contrast, slash-and-burn agriculture is more commonly practiced in mountainous areas because there is much sparser arable land, and the slash-and-burn technique transforms forests into nutrient-rich soil suitable for farming (Scott, 2009) *[p18]*. Consequently, the population density for slash-and-burn farmers is much lower than for wet-rice farmers, approximately an order of magnitude (Bellwood, 1993). Therefore, our observation of the strong genetic bottleneck in HM speakers – especially Bunu, Hmong, and Pahng/Pa Then (Figures 5 & 6) – supports the theoretically expected demographic patterns of slash-and-burn farmers.

Since slash-and-burn soils lose fertility quickly, slash-and-burn farmers must regularly move to a new place and reclaim a new land using the slash-and-burn technique, i.e., shifting farming. Consequently, slash-and-burn farmers, e.g., HM speakers, tend to be

more nomadic than sedentary wet-rice farmers. This explains why the genetic substructure of HM speakers is less geographically associated (Figures 2 & 3) in contrast to an overall geographically associated genetic structure in East Asia (Wang *et al.*, 2021a). By contrast, the distribution of external ancestries in HM speakers is more geographically associated (Figure 8), suggesting the pattern of admixture after settlement. The long-term genetic bottleneck and nomadic lifestyle also led to serial founder events and population differentiation that parallel language differentiation in West Hmongic speakers (Figure 3). Our analysis highlights that the practice of slash-and-burn and shifting agriculture is a crucial factor associated with the genetic and demographic history of HM speakers, which has rarely been addressed in previous genetic studies about HM-speaking populations.

On a worldwide scale, previous genetic studies have revealed the impact of geographic isolation, modes of subsistence, and marital practice on effective population size (Ceballos *et al.*, 2018, Ringbauer *et al.*, 2021, Tournebize *et al.*, 2022). Regarding modes of subsistence, plant farmers tend to have a lower degree of a founder effect, while a nomadic lifestyle leads to enhanced genetic bottleneck (Tournebize *et al.*, 2022). However, it remains unclear whether plant agriculture or nomadism is the more decisive factor in terms of the effective population size in previous studies. Therefore, the evidence from HM speakers – who are farmers but historically nomadic – supports that nomadism plays a more crucial role in demographic patterns.

In conclusion, our study reveals the recent isolation and admixture events within the recent ~2,500 years that contribute to the fine-scale genetic formation of present-day HM-speaking populations and investigates the impact of sociological/anthropological factors – especially slash-and-burn and shifting agriculture – on the genetic pattern of HM speakers. Given the power of GaoHuaHua in the calibration of HM history, we predict that more ancient genomes from Southwest China in the recent ~2,500 years will shed new light on the origin and differentiation of HM-speaking populations. Our

study also highlights the importance of incorporating currently underrepresented groups in future genetic studies, e.g., genome-wide association studies (GWAS).

Methods

Sample collection and genomic data curation

We collected saliva and blood samples from 67 Yao individuals from five autonomous counties (Du'an, Dahua, Bama, Gongcheng, Fuchuan) in Guangxi, China (SI Table 1). All sample donors read and signed informed content, and this research was approved by the Ethical Committee of Xiamen University (approval number: XDYX2019009). All of the processes were in accordance with the corresponding ethical principles. Then, we obtained the genotyped data of these samples using an Infinium Global Screening Array covering 699,537 genome-wide SNPs. We imputed the genotype using Eagle v2.4.1 (Loh *et al.*, 2016) and Minimac4 (Das *et al.*, 2016) with default parameters (with a chunk size of 10 Mb and step size of 3 Mb) against the 1000 Genomes project Phase3 v5 reference haplotypes (Consortium *et al.*, 2015). We then removed SNPs with imputation quality < 0.3, MAF < 1% or missing rate > 2%. To identify presumably related individuals, we used PLINK v1.9 (Purcell *et al.*, 2007) with the parameter '-genome' to estimate PI_HAT values for each pair of newly sampled individuals and removed all three individuals with PI_HAT > 0.15 in all the subsequent analyses. After that, we merged our data with '1240k+HO' dataset in Allen Ancient DNA Resource (AADR) V50.0 (Reich, 2022) and recently published modern (Liu *et al.*, 2020, Kutanan *et al.*, 2021) and ancient genomes (Wang *et al.*, 2021b, Liu *et al.*, 2022) for co-analysis, resulting in 362,468 autosomal SNPs used in all the following analyses. We only kept diploid modern and ancient genomes for haplotype-based analyses and added ancient diploid genomes (Devil's Gate, Yana, and Kolyma) from Sikora *et al.* (2019). We identified outliers given the results from PCA (Patterson *et al.*, 2006) and fineSTRUCTURE (Lawson *et al.*, 2012). Non-HM outliers were removed in subsequent analysis, while HM outliers were marked with labels ('_o' or '_2').

Allele frequency-based analyses

The program *smartpca* packed in EIGENSOFT (Patterson *et al.*, 2006) was used to perform PCA with default parameters and lsqproject: YES, numoutlieriter: 0. Only modern samples were used to construct principal components, whereas ancient samples were projected. For ADMIXTURE analysis (Alexander *et al.*, 2009), we first used PLINK v1.9 (Purcell *et al.*, 2007) to prune linkage disequilibrium (-indep-pairwise 200 20 0.4). ADMIXTURE v1.3.0 was then performed from K=2 – 20 (SI Figure 1) with unsupervised mode and default parameters. K = 10 was reported for its lowest cross-validation error. We used ADMIXTOOLS (Patterson *et al.*, 2012) to compute outgroup f_3-statistics with Mbuti as the outgroup for East Asians. We used *TreeMix* (Pickrell and Pritchard, 2012) to generate allele frequency-based phylogeny with no migration (-m 0), a window size of every 500 SNPs (-k 500), and Onge as the root.

Haplotype-based analyses

We used SHAPEIT v2 (O'Connell *et al.*, 2014) to phase diploid ancient and modern genomes in our data. According to López *et al.* (2021), we used the following procedures to generate a copying-vector profile using ChromoPainter (Lawson *et al.*, 2012). We first estimated the parameters for mutation/emission (*Mut*, '-M') and switch rate (N_e, '-n') with ten steps of the Expectation-Maximization (E-M) algorithm for chromosomes 1, 8, 15, and 22 for the first 10 individuals. Then, we used the fixed *Mut* and N_e values estimated in the previous procedure to run ChromoPainter for all the individuals. For the copying-vector profile used for fineSTRUCTURE (Lawson *et al.*, 2018) analysis, we used only HM-speaking individuals as both donors and recipients with the fineSTRUCTURE normalization parameter 'c' estimated to be 0.395. Then, we used fineSTRUCTURE to perform 200,000 sample iterations under Markov chain Monte Carlo (MCMC) with the first 100,000 samples as burn-ins, and the inferred trees were sampled every 1,000 iterations. After that, we carried out 100,000 hill-climbing steps, and fineSTRUCTURE summarized the most likely tree.

For SOURCEFIND (Chacón-Duque *et al.*, 2018) and fastGLOBETROTTER (Hellenthal *et al.*, 2014, Wangkumhang *et al.*, 2022) analyses, we used all the three populations in the 'Hmong' cluster defined by fineSTRUCTURE (i.e., Hmong, Hmong Daw, and Hmong Njua, Figure 3, referred as 'Hmong_all') to surrogate the shared ancestry by HM speakers, along with 120 ancient and modern global populations to surrogate other ancestries potentially contributing to present-day HM speakers genetically (SI Table 2). In addition to the other 14 HM-speaking populations, we also included 10 non-HM populations who likely received recent HM-related gene flow (SI Table 2) with the HM-related component > 5% in ADMIXTURE when K = 10 (SI Figure 1). We generated the copying-vector profiles for both analyses following the suggestions by Wangkumhang *et al.* (2022). All the other parameters of both methods were kept by default. For ROH analysis, we used PLINK v1.9 (Purcell *et al.*, 2007) with the parameters following the suggestions by Ceballos *et al.* (2018). We used Refine IBD to obtain pairwise IBD of HM-speaking individuals with default parameters (Browning and Browning, 2013b). For intra-population IBDs, we used all the IBD segments > 1.0 cM to estimate the change of N_e over time using IBDNe (Browning and Browning, 2015) for the recent ~2,000 years. Following Ralph and Coop (2013), we grouped inter-population IBDs into three categories: 1–5 cM, 5–10 cM, and > 10 cM, which approximately correspond to the time range of 500 BCE – 500 CE, 500 CE – 1500 CE, and > 1500 CE, respectively. For fastGLOBETROTTER and IBDNe, we used 1975 CE as the average birthdate for all the modern individuals and 28 years per generation to covert generation into year following López *et al.* (2021).

ACKNOWLEDGEMENTS

The work was funded by Guangxi First-class Discipline Project for Basic medicine Sciences (No. GXFCDP-BMS-2018), the National Natural Science Foundation of China (32270667), Basic Medical Science and Technology Innovation Education Fund of Guangxi Medical University (No.: GXMUBMSTCF-G10), the "Double First Class University Plan" key construction project of Xiamen University

(0310/X2106027), Nanqiang Outstanding Young Talents Program of Xiamen University (X2123302), the Major Project of National Social Science Foundation of China granted to Chuan-Chao Wang (21&ZD285), Major Special Project of Philosophy and Social Sciences Research of the Ministry of Education (2022JZDZ023), and European Research Council (ERC) grant (ERC-2019-ADG-883700-TRAM).

AUTHOR CONTRIBUTIONS

C.C Wang and Z.Y Xia conducted the project and conceived the idea. Q.Y Deng and X.C Chen collected the samples and carried out the experiments. Z.Y Xia analyzed the data and wrote the paper. C.C Wang edited the draft. All the authors revised the paper.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Figure Legends

Figure 1. Geographic and linguistic information of newly reported samples. (A) Locations of all the five Yao autonomous counties where the individuals were sampled. (B) Languages spoken in all the five counties by people with Yao ethnicity and corresponding autonyms, according to Mao *et al.* (1982) and Meng (2001).

Figure 2. PCA of HM-speaking populations in the context of southern East Asians.

Figure 3. FineSTRUCTURE dendrogram of HM-speaking populations.

Figure 4. Genetic relationship of GaoHuaHua ancient samples with present-day HM speakers. (A) PCA exclusively for GaoHuaHua and modern HM speakers. (B)

Outgroup f_3(Mbuti; X, GaoHuaHua) where X are all the East Asian populations. (C) TreeMix phylogeny with GaoHuaHua and modern HM speakers.

Figure 5. ROH analysis. (A) ROH of HM speakers and other South Chinese populations. (B) ROH of selected worldwide ancient and present-day populations. Dash line shows the average total sum of ROH > 2 cM for all the three Bunu populations (Yao_Dahua, Yao_Duan1, and Yao_Bama2).

Figure 6. IBD analysis. (A) The trajectories of N_e of HM-speaking populations estimated from intra-population IBD. (B) Inter-population IBD sharing of HM speakers.

Figure 7. Admixture dates and most likely contributors inferred by fastGLOBETROTTER. (A) Point estimates and 95% CIs of inferred admixture dates in HM speakers and neighboring populations. (B) Contributions of two-way admixture and single best-matching proxies for each ancestry source.

Figure 8. Finescale ancestry composition inferred by SOURCEFIND. (A) A summary of language family-level ancestry contribution. Ancestry contribution in detail from proxies of (B) Han Chinese, (C) Southeast Asians (Kra-Dai, Austronesian, and Austroasiatic speakers), and (D) Tibeto-Burman speakers.

Reference

Alexander, D. H., Novembre, J. & Lange, K. (2009). 'Fast model-based estimation of ancestry in unrelated individuals'. *Genome research,* 19**,** 1655-1664.

Ayub, Q., Mezzavilla, M., Pagani, L., Haber, M., Mohyuddin, A., Khaliq, S., Mehdi, S. Q. & Tyler-Smith, C. (2015). 'The Kalash genetic isolate: ancient divergence, drift, and selection'. *The American Journal of Human Genetics,* 96**,** 775-783.

Bellwood, P. (1993). Southeast Asia before History. *In:* TARLING, N. (ed.) *The*

Cambridge History of Southeast Asia: Volume 1: From Early Times to 1800. Cambridge: Cambridge University Press.

Browning, B. L. & Browning, S. R. (2013a). 'Detecting identity by descent and estimating genotype error rates in sequence data'. *The American journal of human genetics*, 93**,** 840-851.

Browning, B. L. & Browning, S. R. (2013b). 'Improving the accuracy and efficiency of identity-by-descent detection in population data'. *Genetics*, 194**,** 459-471.

Browning, S. R. & Browning, B. L. (2015). 'Accurate non-parametric estimation of recent effective population size from segments of identity by descent'. *The American Journal of Human Genetics*, 97**,** 404-418.

Cann, H. M., de Toma, C., Cazes, L., Legrand, M. F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W. F., Bonne-Tamir, B., Cambon-Thomsen, A., Chen, Z., Chu, J. Y., Carcassi, C., Contu, L., Du, R. F., Excoffier, L., Ferrara, G. B., Friedlaender, J. S., Groot, H., Gurwitz, D., Jenkins, T., Herrera, R. J., Huang, X. Y., Kidd, J., Kidd, K. K., Langaney, A., Lin, A. A., Mehdi, S. Q., Parham, P., Piazza, A., Pistillo, M. P., Qian, Y. P., Shu, Q. F., Xu, J. J., Zhu, S., Weber, J. L., Greely, H. T., Feldman, M. W., Thomas, G., Dausset, J. & Cavalli-Sforza, L. L. (2002). 'A human genome diversity cell line panel'. *Science*, 296**,** 261-262.

Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M. & Wilson, J. F. (2018). 'Runs of homozygosity: windows into population history and trait architecture'. *Nature Reviews Genetics*, 19**,** 220-234.

Chacón-Duque, J.-C., Adhikari, K., Fuentes-Guajardo, M., Mendoza-Revilla, J., Acuña-Alonzo, V., Barquera, R., Quinto-Sánchez, M., Gómez-Valdés, J., Everardo Martínez, P., Villamil-Ramírez, H., Hünemeier, T., Ramallo, V., Silva de Cerqueira, C. C., Hurtado, M., Villegas, V., Granja, V., Villena, M., Vásquez, R., Llop, E., Sandoval, J. R., Salazar-Granara, A. A., Parolin, M.-L., Sandoval, K., Peñaloza-Espinosa, R. I., Rangel-Villalobos, H., Winkler, C. A., Klitz, W., Bravi, C., Molina, J., Corach, D., Barrantes, R., Gomes, V., Resende, C., Gusmão, L., Amorim, A., Xue, Y., Dugoujon, J.-M., Moral, P., González-José, R., Schuler-Faccini, L., Salzano, F. M., Bortolini, M.-C.,

Canizales-Quinteros, S., Poletti, G., Gallo, C., Bedoya, G., Rothhammer, F., Balding, D., Hellenthal, G. & Ruiz-Linares, A. (2018). 'Latin Americans show wide-spread Converso ancestry and imprint of local Native ancestry on physical appearance'. *Nature communications*, 9**,** 5388.

Chan, W.-h. (2006). 9. Ethnic Labels in a Mountainous Region: The Case of She "Bandits". *In:* PAMELA KYLE, C., HELEN, F. S. & DONALD, S. S. (eds.) *Empire at the Margins.* Berkeley: University of California Press.

Consortium, G. P., Auton, A., Brooks, L., Durbin, R., Garrison, E. & Kang, H. (2015). 'A global reference for human genetic variation'. *Nature*, 526**,** 68-74.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S. & McGue, M. (2016). 'Next-generation genotype imputation service and methods'. *Nature genetics*, 48**,** 1284-1287.

Diamond, N. (1995). Defining the Miao: Ming, Qing, and Contemporary Views. *In:* HARRELL, S. (ed.) *Cultural Encounters on China's Ethnic Frontiers.* University of Washington Press.

Faure, D. (2006). The Yao Wars in the Mid-Ming and their Impact on Yao Ethnicity. *In:* PAMELA KYLE, C., HELEN, F. S. & DONALD, S. S. (eds.) *Empire at the Margins.* Berkeley: University of California Press.

Font-Porterias, N., Arauna, L. R., Poveda, A., Bianco, E., Rebato, E., Prata, M. J., Calafell, F. & Comas, D. (2019). 'European Roma groups show complex West Eurasian admixture footprints and a common South Asian genetic origin'. *PLoS genetics*, 15**,** e1008417.

Hammarström, H., Forkel, R., Haspelmath, M. & Bank, S. (2022). Glottolog 4.6. *In:* ANTHROPOLOGY, L. M. P. I. F. E. (ed.).

Hellenthal, G., Busby, G. B., Band, G., Wilson, J. F., Capelli, C., Falush, D. & Myers, S. (2014). 'A genetic atlas of human admixture history'. *Science*, 343**,** 747-751.

Hellenthal, G., Falush, D., Myers, S., Reich, D., Busby, G. B., Lipson, M., Capelli, C. & Patterson, N. (2016). 'The Kalash genetic isolate? the evidence for recent admixture'. *American journal of human genetics*, 98**,** 396.

Huang, X., Xia, Z.-Y., Bin, X., He, G., Guo, J., Adnan, A., Yin, L., Huang, Y., Zhao, J.

& Yang, Y. (2022). 'Genomic Insights Into the Demographic History of the Southern Chinese'. *Frontiers in Ecology and Evolution*, 556.

Kutanan, W., Liu, D., Kampuansai, J., Srikummool, M., Srithawong, S., Shoocongdej, R., Sangkhano, S., Ruangchai, S., Pittayaporn, P. & Arias, L. (2021). 'Reconstructing the human genetic history of mainland Southeast Asia: insights from genome-wide data from Thailand and Laos'. *Molecular biology and evolution*, 38, 3459-3477.

Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. (2012). 'Inference of population structure using dense haplotype data'. *PLoS genetics*, 8, e1002453.

Lawson, D. J., Van Dorp, L. & Falush, D. (2018). 'A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots'. *Nature communications*, 9, 3258.

Leslie, S., Winney, B., Hellenthal, G., Davison, D., Boumertit, A., Day, T., Hutnik, K., Royrvik, E. C., Cunliffe, B. & Lawson, D. J. (2015). 'The fine-scale genetic structure of the British population'. *Nature*, 519, 309.

Litzinger, R. A. (1995). Making Histories: Contending Conceptions of the Yao Past. *In:* HARRELL, S. (ed.) *Cultural Encounters on China's Ethnic Frontiers.* University of Washington Press.

Liu, C.-C., Witonsky, D., Gosling, A., Lee, J. H., Ringbauer, H., Hagan, R., Patel, N., Stahl, R., Novembre, J. & Aldenderfer, M. (2022). 'Ancient genomes from the Himalayas illuminate the genetic history of Tibetans and their Tibeto-Burman speaking neighbors'. *Nature communications*, 13, 1-14.

Liu, D., Duong, N. T., Ton, N. D., Van Phong, N., Pakendorf, B., Van Hai, N. & Stoneking, M. (2020). 'Extensive ethnolinguistic diversity in Vietnam reflects multiple sources of genetic diversity'. *Molecular biology and evolution*, 37, 2503-2519.

Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., A Reshef, Y., K Finucane, H., Schoenherr, S., Forer, L., McCarthy, S. & Abecasis, G. R. (2016). 'Reference-based phasing using the Haplotype Reference Consortium panel'. *Nature genetics*, 48, 1443-1448.

López, S., Tarekegn, A., Band, G., van Dorp, L., Bird, N., Morris, S., Oljira, T., Mekonnen, E., Bekele, E. & Blench, R. (2021). 'Evidence of the interplay of genetics and culture in Ethiopia'. *Nature communications,* 12**,** 1-15.

Mao, Z. & Meng, C. (1986). *A sketch of the She language (in Chinese)*, Beijing, Nationalities Press.

Mao, Z., Meng, C. & Zheng, Z. (1982). *A Sketch of Yao People's Languages (in Chinese)*.

Meng, C. (2001). *A Study of the Yao Bunu Dialects [in Chinese]*, Beijing, Publishing House of Minority Nationalities.

O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., Traglia, M., Huang, J., Huffman, J. E. & Rudan, I. (2014). 'A general approach for haplotype phasing across the full spectrum of relatedness'. *PLoS Genet,* 10**,** e1004234.

Ozawa, M. (2000). 'On Slash-and-Burn Agriculture in Tang-Song China (originally in Japanese, translated into Chinese)'. *Journal of Chinese Historical Geography***,** 223-249.

Palamara, P. F., Lencz, T., Darvasi, A. & Pe'er, I. (2012). 'Length distributions of identity by descent reveal fine-scale demographic history'. *The American journal of human genetics,* 91**,** 809-822.

Patterson, N., Price, A. L. & Reich, D. (2006). 'Population structure and eigenanalysis'. *PLoS genetics,* 2**,** e190.

Patterson, N. J., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T. & Reich, D. (2012). 'Ancient admixture in human history'. *Genetics,* 192(3)**,** 1065-93.

Pickrell, J. K. & Pritchard, J. K. (2012). 'Inference of population splits and mixtures from genome-wide allele frequency data'. *PLoS genetics,* 8**,** e1002967.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. & Daly, M. J. (2007). 'PLINK: a tool set for whole-genome association and population-based linkage analyses'. *The American journal of human genetics,* 81**,** 559-575.

Ralph, P. & Coop, G. (2013). 'The geography of recent genetic ancestry across Europe'. *PLoS biology,* 11**,** e1001555.

Ratliff, M. (2021). 14 Classification and historical overview of Hmong-Mien languages. *The Languages and Linguistics of Mainland Southeast Asia.* De Gruyter Mouton.

Ratliff, M. S. (2010). *Hmong-Mien language history,* Camberra, Research School of Pacific and Asian Studies, The Australian National University.

Reich, D. (2022). *Allen Ancient DNA Resource (version 50.0)* [Online]. Available: https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data [Accessed 2022].

Ringbauer, H., Novembre, J. & Steinrücken, M. (2021). 'Parental relatedness through time revealed by runs of homozygosity in ancient DNA'. *Nature communications,* 12**,** 1-11.

Scott, J. C. (2009). *The Art of Not Being Governed: An Anarchist History of Upland Southeast Asia,* Yale University Press.

Sikora, M., Pitulko, V. V., Sousa, V. C., Allentoft, M. E., Vinner, L., Rasmussen, S., Margaryan, A., de Barros Damgaard, P., de la Fuente, C., Renaud, G., Yang, M. A., Fu, Q., Dupanloup, I., Giampoudakis, K., Nogués-Bravo, D., Rahbek, C., Kroonen, G., Peyrot, M., McColl, H., Vasilyev, S. V., Veselovskaya, E., Gerasimova, M., Pavlova, E. Y., Chasnyk, V. G., Nikolskiy, P. A., Gromov, A. V., Khartanovich, V. I., Moiseyev, V., Grebenyuk, P. S., Fedorchenko, A. Y., Lebedintsev, A. I., Slobodin, S. B., Malyarchuk, B. A., Martiniano, R., Meldgaard, M., Arppe, L., Palo, J. U., Sundell, T., Mannermaa, K., Putkonen, M., Alexandersen, V., Primeau, C., Baimukhanov, N., Malhi, R. S., Sjögren, K.-G., Kristiansen, K., Wessman, A., Sajantila, A., Lahr, M. M., Durbin, R., Nielsen, R., Meltzer, D. J., Excoffier, L. & Willerslev, E. (2019). 'The population history of northeastern Siberia since the Pleistocene'. *Nature,* 570(7760)**,** 182-188.

Simons, G. F. & Fennig, C. D. (2017). *Ethnologue: Languages of the World. Online version* [Online]. Dallas, Texas: SIL International. Available:

http://www.ethnologue.com [Accessed 2017].

Strecker, D. (1987). 'The Hmong-Mien languages'. *Linguistics of the Tibeto-Burman area,* 10**,** 1-11.

Tournebize, R., Chu, G. & Moorjani, P. (2022). 'Reconstructing the history of founder events using genome-wide patterns of allele sharing across individuals'. *PLoS Genetics,* 18**,** e1010243.

Van Dorp, L., Balding, D., Myers, S., Pagani, L., Tyler-Smith, C., Bekele, E., Tarekegn, A., Thomas, M. G., Bradman, N. & Hellenthal, G. (2015). 'Evidence for a common origin of blacksmiths and cultivators in the Ethiopian Ari within the last 4500 years: lessons for clustering-based inference'. *PLoS genetics,* 11**,** e1005397.

Wang, C.-C., Yeh, H.-Y., Popov, A. N., Zhang, H.-Q., Matsumura, H., Sirak, K., Cheronet, O., Kovalev, A., Rohland, N., Kim, A. M., Mallick, S., Bernardos, R., Tumen, D., Zhao, J., Liu, Y.-C., Liu, J.-Y., Mah, M., Wang, K., Zhang, Z., Adamski, N., Broomandkhoshbacht, N., Callan, K., Candilio, F., Carlson, K. S. D., Culleton, B. J., Eccles, L., Freilich, S., Keating, D., Lawson, A. M., Mandl, K., Michel, M., Oppenheimer, J., Özdoğan, K. T., Stewardson, K., Wen, S., Yan, S., Zalzala, F., Chuang, R., Huang, C.-J., Looh, H., Shiung, C.-C., Nikitin, Y. G., Tabarev, A. V., Tishkin, A. A., Lin, S., Sun, Z.-Y., Wu, X.-M., Yang, T.-L., Hu, X., Chen, L., Du, H., Bayarsaikhan, J., Mijiddorj, E., Erdenebaatar, D., Iderkhangai, T.-O., Myagmar, E., Kanzawa-Kiriyama, H., Nishino, M., Shinoda, K.-i., Shubina, O. A., Guo, J., Cai, W., Deng, Q., Kang, L., Li, D., Li, D., Lin, R., Nini, Shrestha, R., Wang, L.-X., Wei, L., Xie, G., Yao, H., Zhang, M., He, G., Yang, X., Hu, R., Robbeets, M., Schiffels, S., Kennett, D. J., Jin, L., Li, H., Krause, J., Pinhasi, R. & Reich, D. (2021a). 'Genomic Insights into the Formation of Human Populations in East Asia'. *Nature,* 591(7850)**,** 413–419.

Wang, F. (1983a). 'On the dialect divisions of the Miao language'. *Minzu Yuwen,* 5**,** 1-22.

Wang, F. (1983b). *A Sketch of the Miao Language (in Chinese)*, Beijing, Nationalities

Press.

Wang, T., Wang, W., Xie, G., Li, Z., Fan, X., Yang, Q., Wu, X., Cao, P., Liu, Y. & Yang, R. (2021b). 'Human population history at the crossroads of East and Southeast Asia since 11,000 years ago'. *Cell,* 184**,** 3829-3841.

Wangkumhang, P., Greenfield, M. & Hellenthal, G. (2022). 'An efficient method to identify, date, and describe admixture events using haplotype information'. *Genome research,* 32**,** 1553-1564.

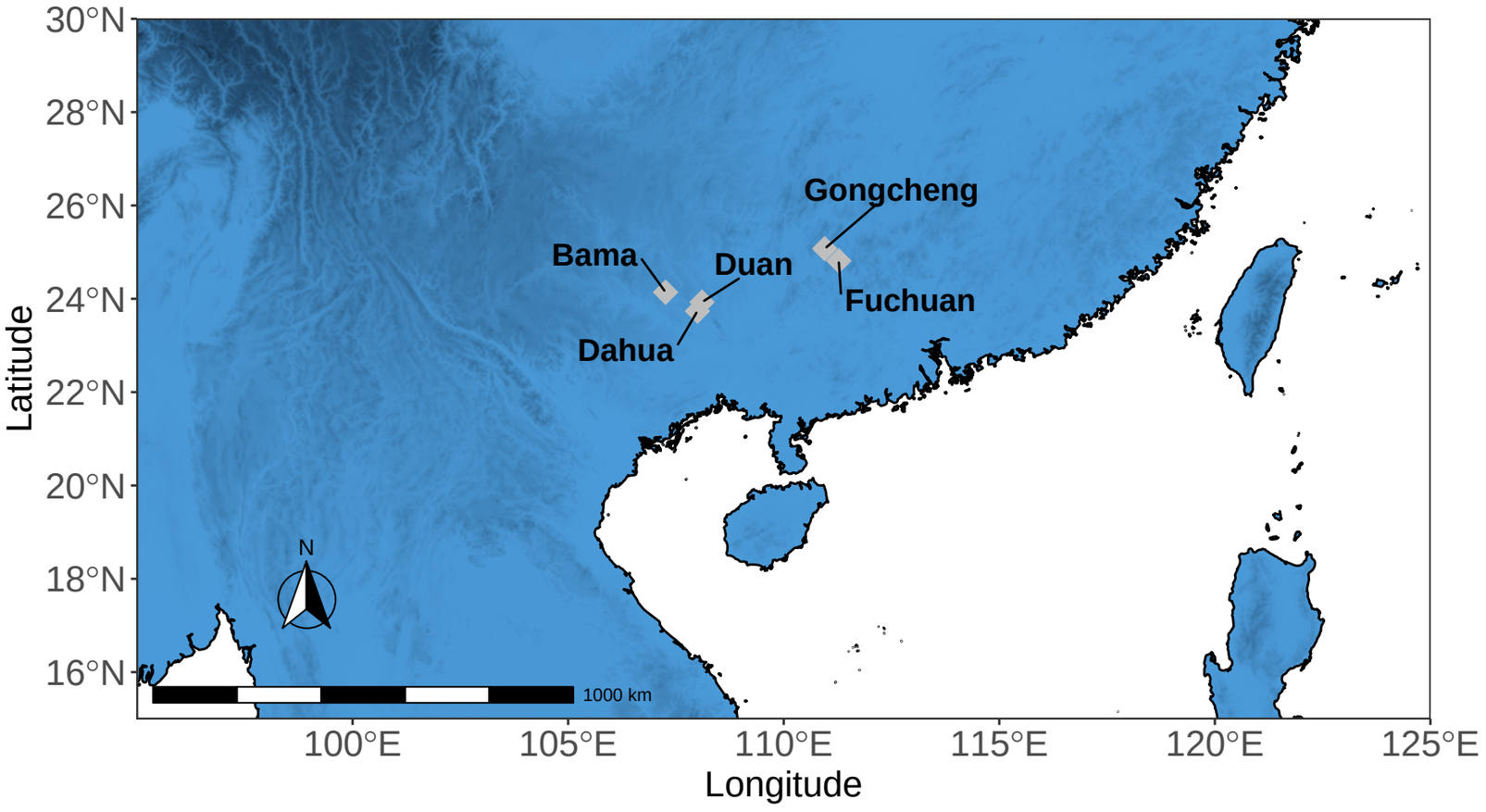
Xia, Z.-Y., Yan, S., Wang, C.-C., Zheng, H.-X., Zhang, F., Liu, Y.-C., Yu, G., Yu, B.-X., Shu, L.-L. & Jin, L. (2019). 'Inland-coastal bifurcation of southern East Asians revealed by Hmong-Mien genomic history'. *Unpublished work, available at https://www.biorxiv.org/content/10.1101/730903v1***,** 730903.

Yang, M., He, G., Ren, Z., Wang, Q., Liu, Y., Zhang, H., Zhang, H., Chen, J., Ji, J. & Zhao, J. (2022). 'Genomic insights into the unique demographic history and genetic structure of five Hmong-Mien speaking Miao and Yao populations in Southwest China'. *Frontiers in Ecology and Evolution***,** 498.

Zeng, X. (2005). 'On She-Tian and Its Related Nationalities in Tang and Song Dynasties (in Chinese)'. *Ancient and Modern Agriculture*.

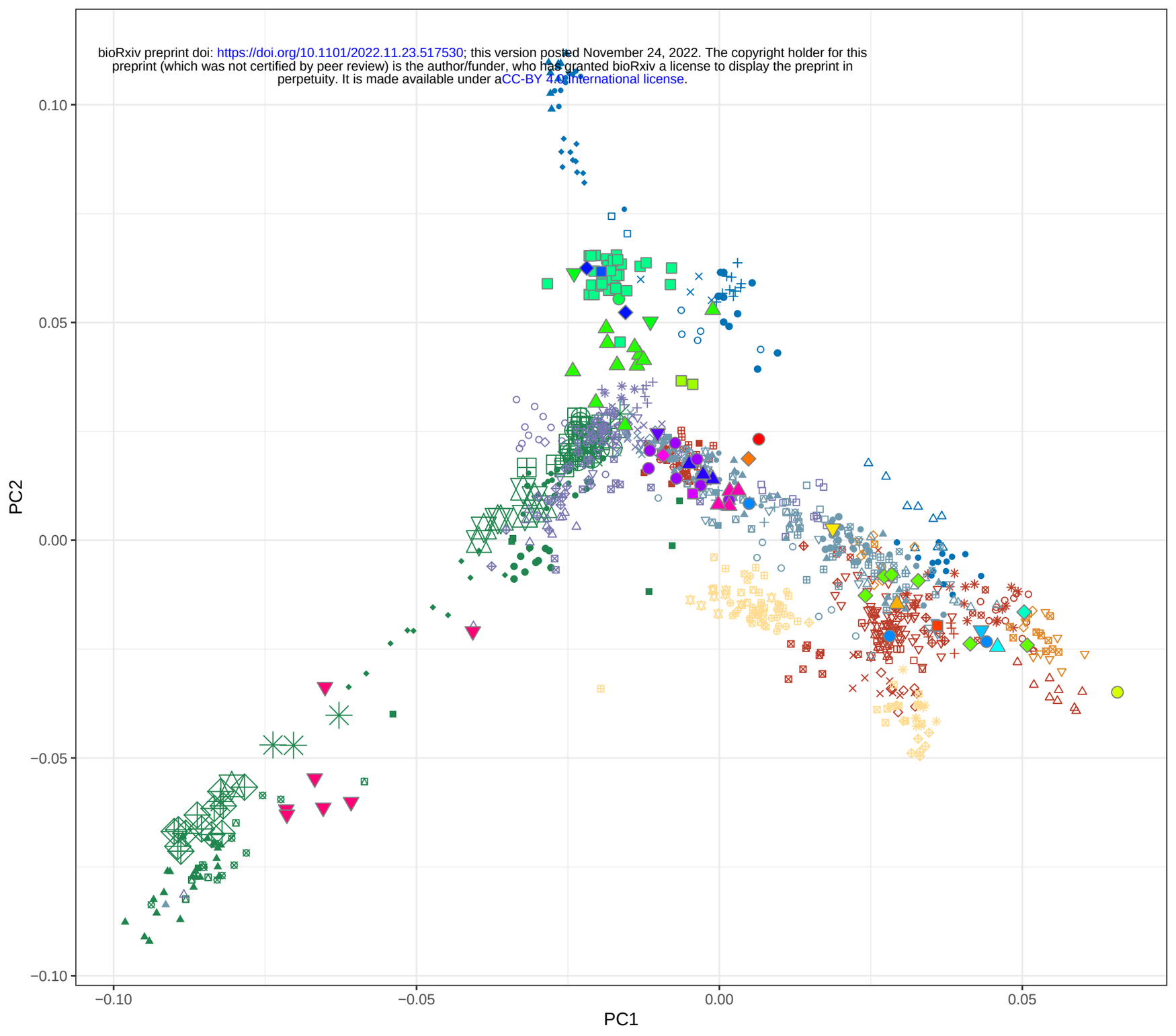
Zhang, S., Peng, S., Zhou, S. & Wu, W. (1986). 'Survey and Research on Cliff Burials of Lihu Cliff Cave, Nandan County, Guangxi [in Chinese]'. *Chinese Cultural Relics***,** 65-75.

A

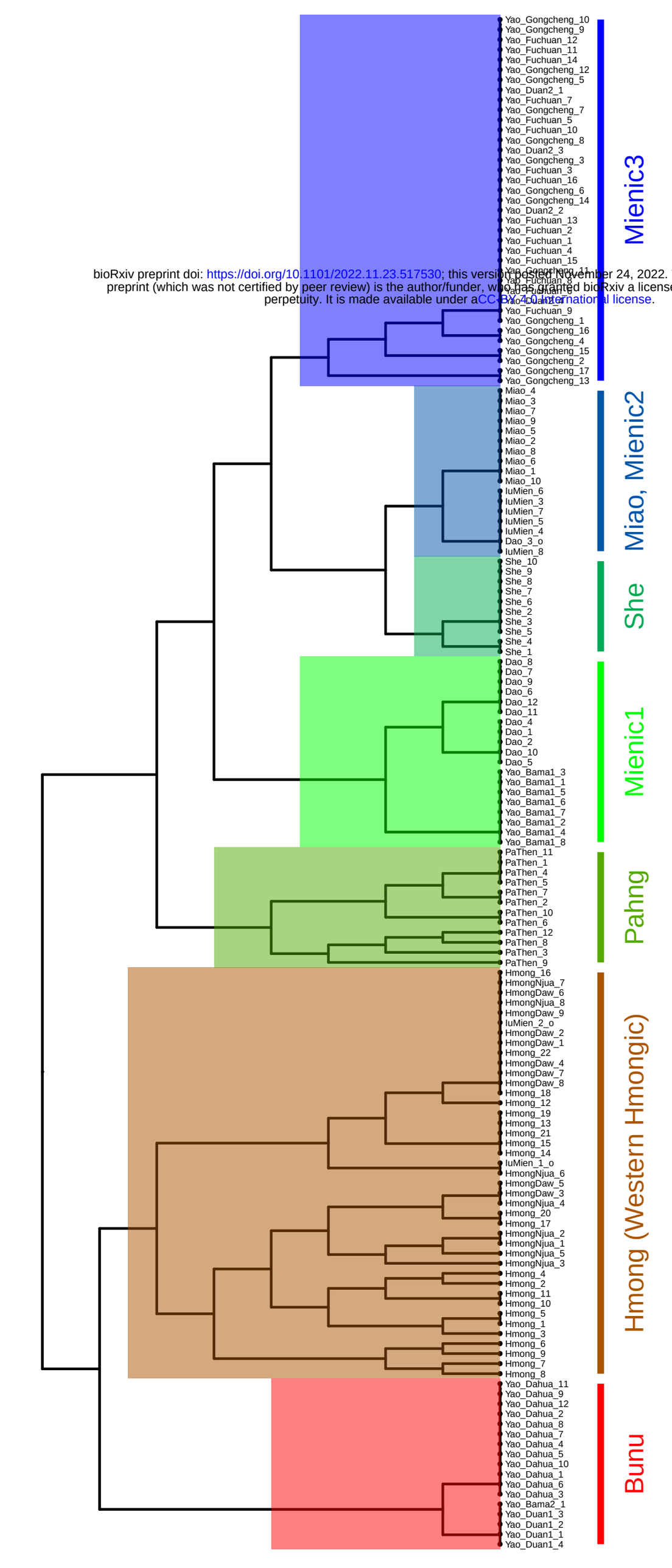


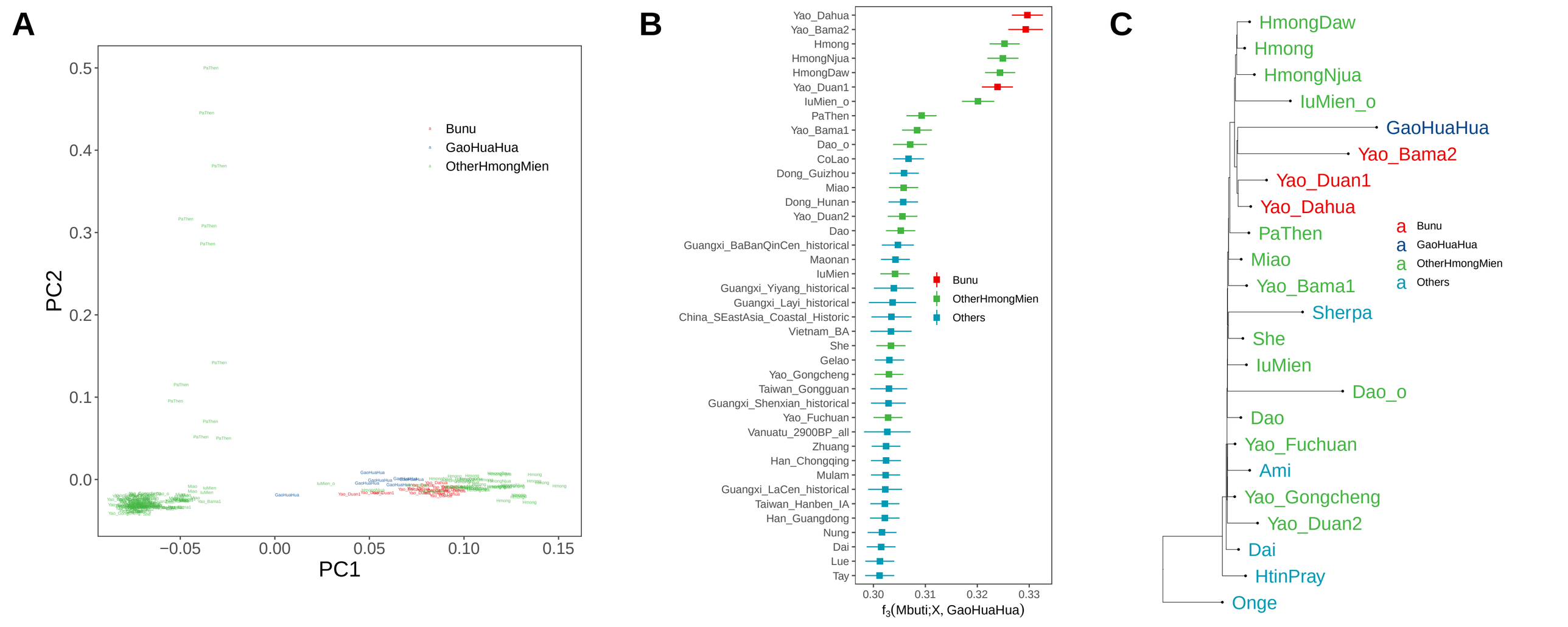
B

Linguistic.classification	Autonym	Distribution
HM, Hmongic, West Hmongic	Bunu	Dahua, Duan, Bama
HM, Mienic	Iu Mien	Duan, Fuchuan, Gongcheng
HM, Mienic	Biao Min	Gongcheng
HM, Mienic	Kim Mun	Bama
ST, Sinitic	Pingdi Yao	Fuchuan, Gongcheng

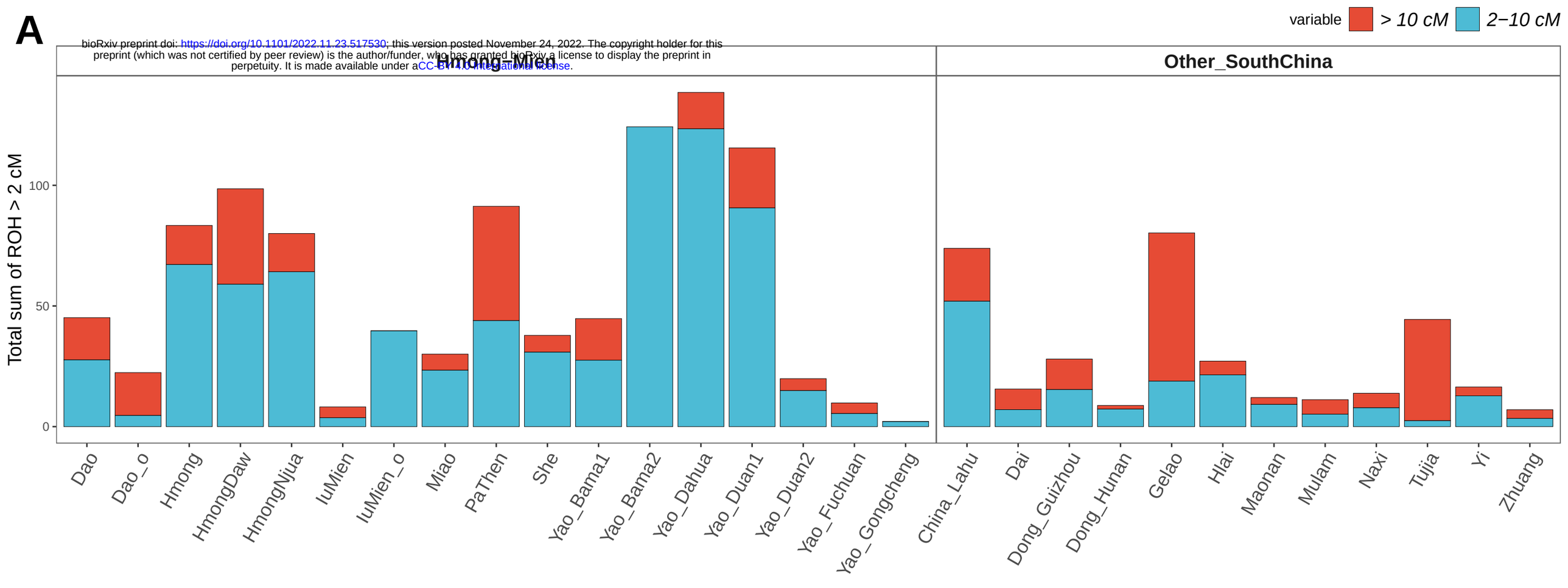


Blang
Bru
HtinPray
Khmu
Lawa_Eastern
Lawa_Western
Mon
Palaung
Soa
KhoMu
Kinh_Vietnam
Mang
Muong
Cambodian
Kinh
Vietnamese
SouthernThai_AN
Atayal
Ami
Dusun
Kankanaey
Ilocano
Tagalog
Malay
Murut
Visayan
Cham
Ede
Giarai
Yao_Duan
Yao_Dauhua
Yao_Fuchuan
Yao_Bama
Yao_Gongcheng
HmongDaw
HmongNjua
IuMien
Dao
Hmong
PaThen
Miao
She
Saek
BoY
CoLao
LaChi
Nung
Maonan
Zhuang
Gelao
Hlai
Dong
Mulam
BlackTai
CentralThai
Kalueang
Khonmueang
Khuen
LaoIsan
Laotian
Lue
Nyaw
Phuan
Phutai
Shan
SouthernThai_TK
Yuan
Tay
Dai
Thai
KarenPadaung
KarenPwo
KarenSkaw
Lahu
Lisu
Cong
Vietnam_Lahu
China_Lahu
Fujian_Qihe3_EP
Guangxi_Longlin_EP
China_SEastAsia_Coastal_EN
Guangxi_Baojianshan_MN
Guangxi_Dushan_EN
Laos_Hoabinhian.SG
China_SEastAsia_Island_EN
Vietnam_N
China_SEastAsia_Coastal_LN
China_SEastAsia_Island_LN
Vanuatu_2900BP_all
Taiwan_Hanben_IA
Indonesia_LN_BA_IA.SG
Laos_LN_BA.SG
Malaysia_LN.SG
Vietnam_LN.SG
Vanuatu_3000BP
Guam_LateUnai_Ritidian
Guangxi_LaCen_historical
Guangxi_Yiyang_historical
Guangxi_BaBanQinCen_historical
Guangxi_Shenxian_historical
Guangxi_Layi_historical
Vietnam_BA_DongSonCulture.SG
Guangxi_GaoHuaHua_historical

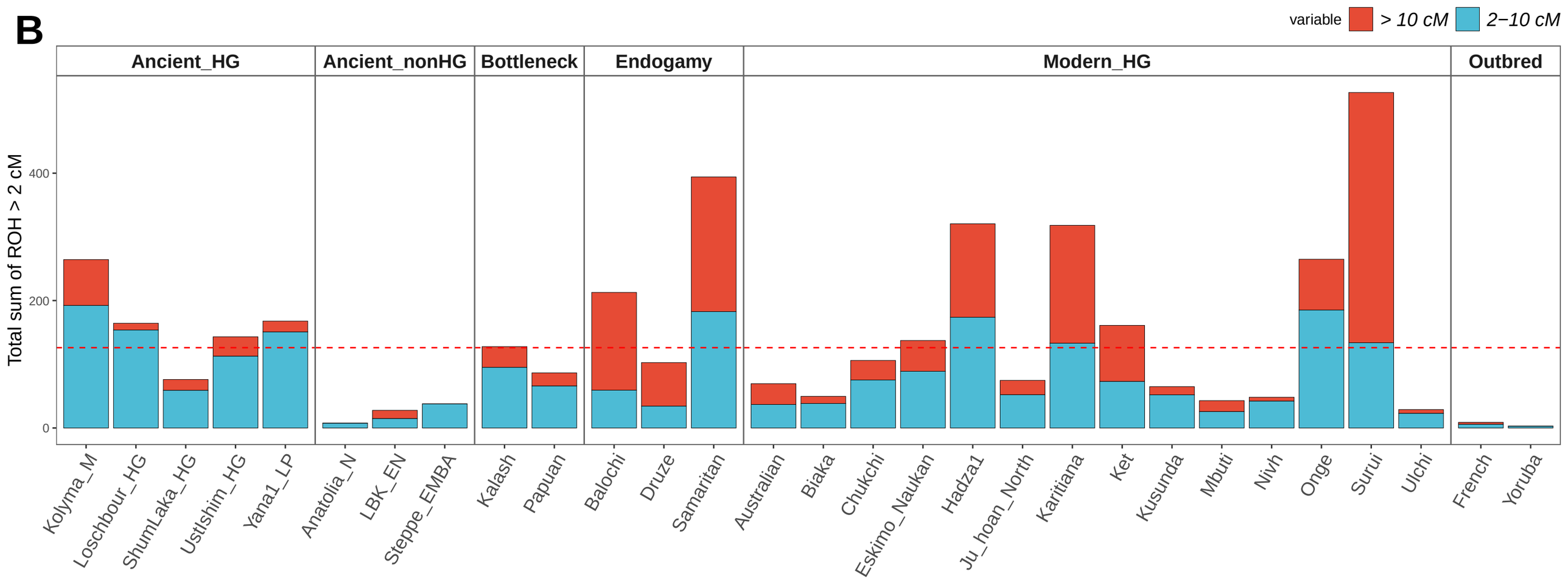


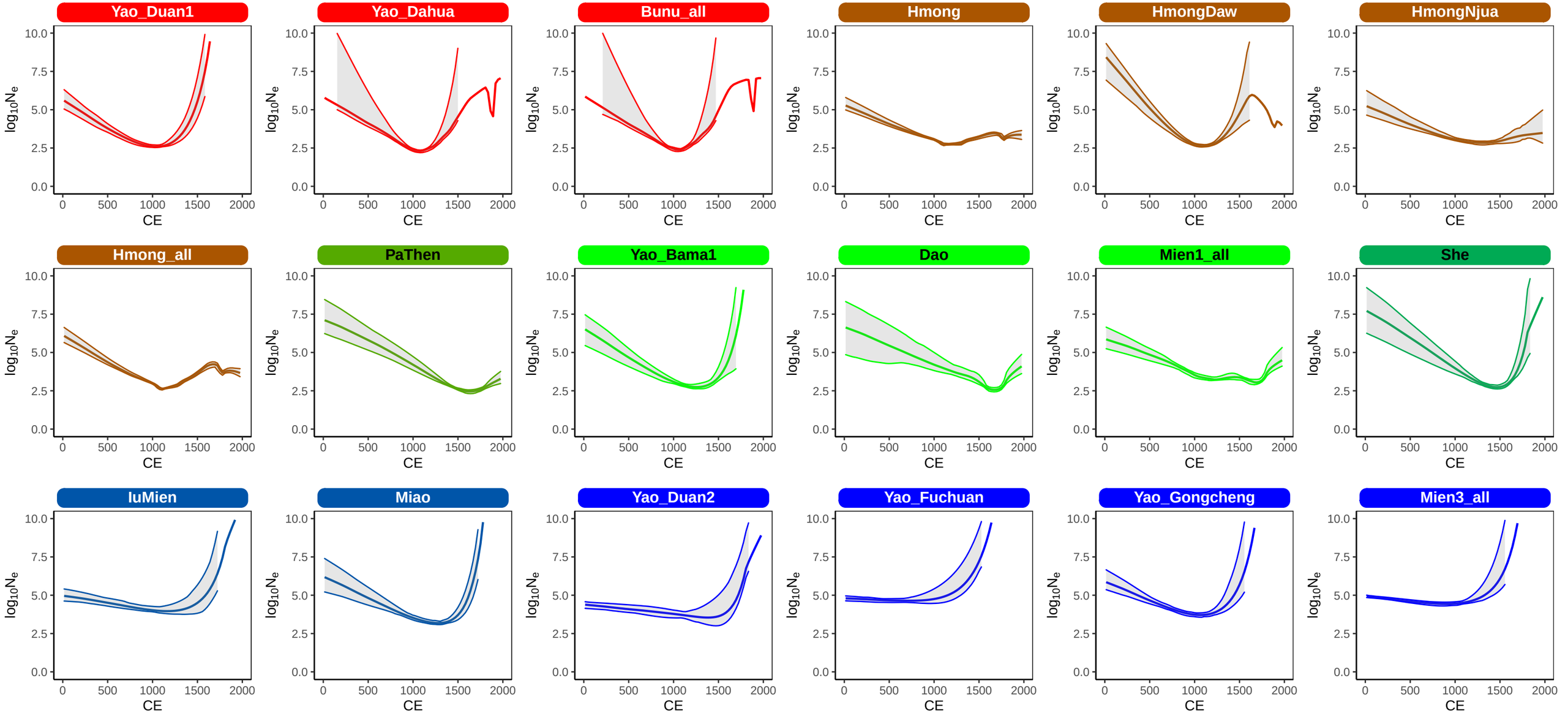

A
PC2
PC1
0.5
0.4
0.3
0.2
0.1
0.0
−0.05
0.00
0.05
0.10
0.15
Bunu
GaoHuaHua
OtherHmongMien
B
Yao_Dahua
Yao_Bama2
Hmong
HmongNjua
HmongDaw
Yao_Duan1
IuMien_o
PaThen
Yao_Bama1
Dao_o
CoLao
Dong_Guizhou
Miao
Dong_Hunan
Yao_Duan2
Dao
Guangxi_BaBanQinCen_historical
Maonan
IuMien
Guangxi_Yiyang_historical
Guangxi_Layi_historical
China_SEastAsia_Coastal_Historic
Vietnam_BA
She
Gelao
Yao_Gongcheng
Taiwan_Gongguan
Guangxi_Shenxian_historical
Yao_Fuchuan
Vanuatu_2900BP_all
Zhuang
Han_Chongqing
Mulam
Guangxi_LaCen_historical
Taiwan_Hanben_IA
Han_Guangdong
Nung
Dai
Lue
Tay
Bunu
OtherHmongMien
Others
0.30
0.31
0.32
0.33
f3(Mbuti;X, GaoHuaHua)
C
HmongDaw
Hmong
HmongNjua
IuMien_o
GaoHuaHua
Yao_Bama2
Yao_Duan1
Yao_Dahua
PaThen
Miao
Yao_Bama1
Sherpa
She
IuMien
Dao_o
Dao
Yao_Fuchuan
Ami
Yao_Gongcheng
Yao_Duan2
Dai
HtinPray
Onge
a Bunu
a GaoHuaHua
a OtherHmongMien
a Others

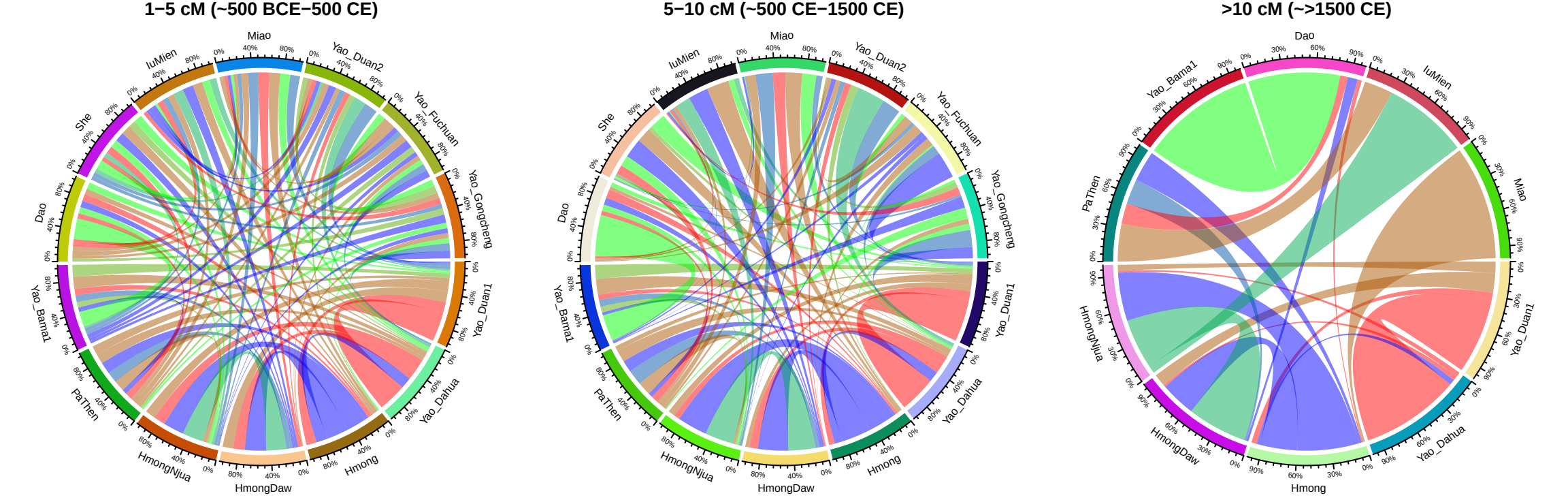
A

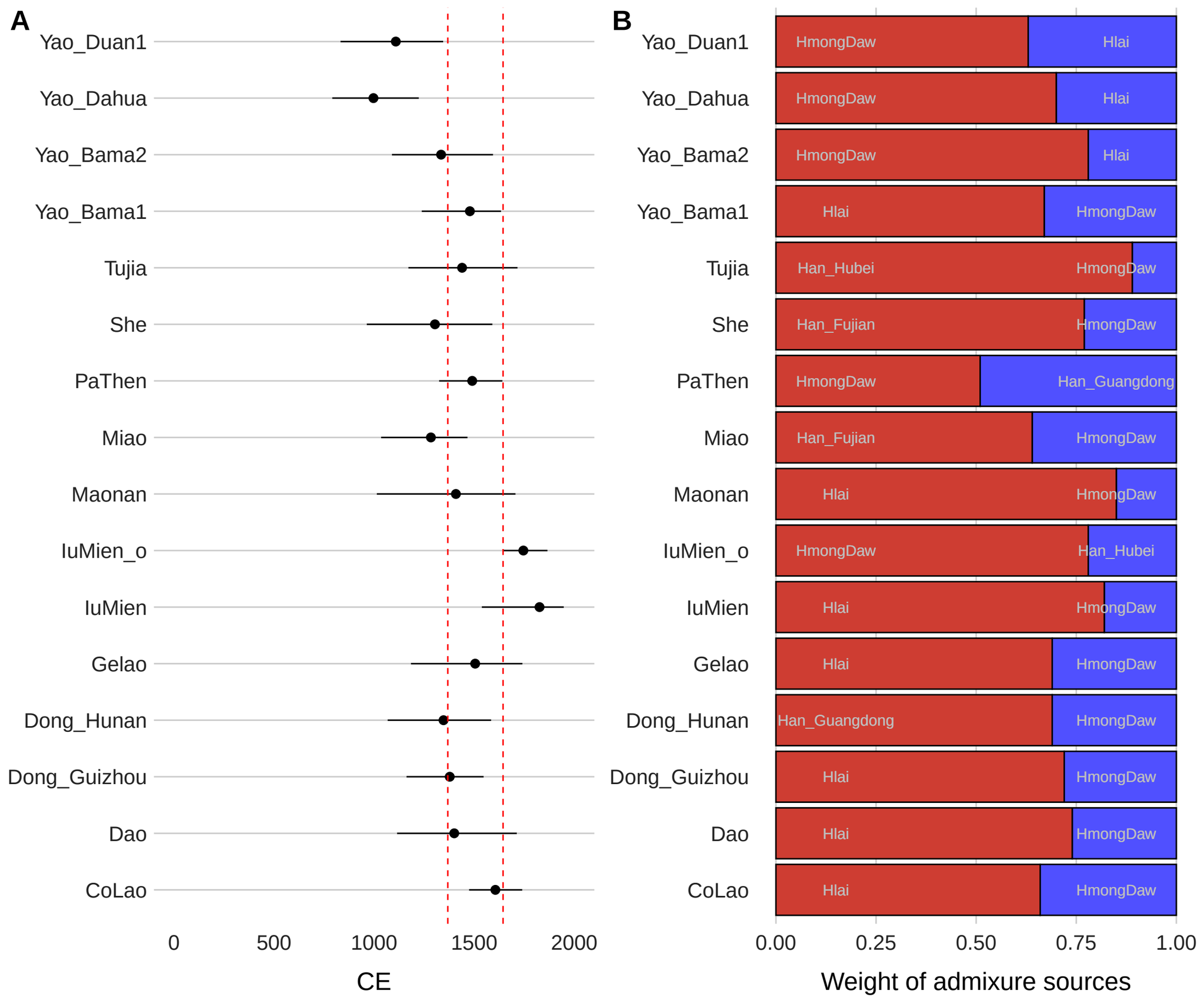


B




Yao_Duan1
Yao_Dahua
Bunu_all
Hmong
HmongDaw
HmongNjua
Hmong_all
PaThen
Yao_Bama1
Dao
Mien1_all
She
IuMien
Miao
Yao_Duan2
Yao_Fuchuan
Yao_Gongcheng
Mien3_all
log10Ne
CE
0.0
2.5
5.0
7.5
10.0
0
500
1000
1500
2000


1−5 cM (~500 BCE−500 CE)
5−10 cM (~500 CE−1500 CE)
>10 cM (~>1500 CE)
Miao
IuMien
Yao_Duan2
Yao_Fuchuan
Yao_Gongcheng
Yao_Duan1
Yao_Dahua
Hmong
HmongDaw
HmongNjua
PaThen
Yao_Bama1
She
Dao


A
Yao_Duan1
Yao_Dahua
Yao_Bama2
Yao_Bama1
Tujia
She
PaThen
Miao
Maonan
IuMien_o
IuMien
Gelao
Dong_Hunan
Dong_Guizhou
Dao
CoLao
0
500
1000
1500
2000
CE
B
HmongDaw
Hlai
Han_Hubei
Han_Fujian
Han_Guangdong
0.00
0.25
0.50
0.75
1.00
Weight of admixure sources