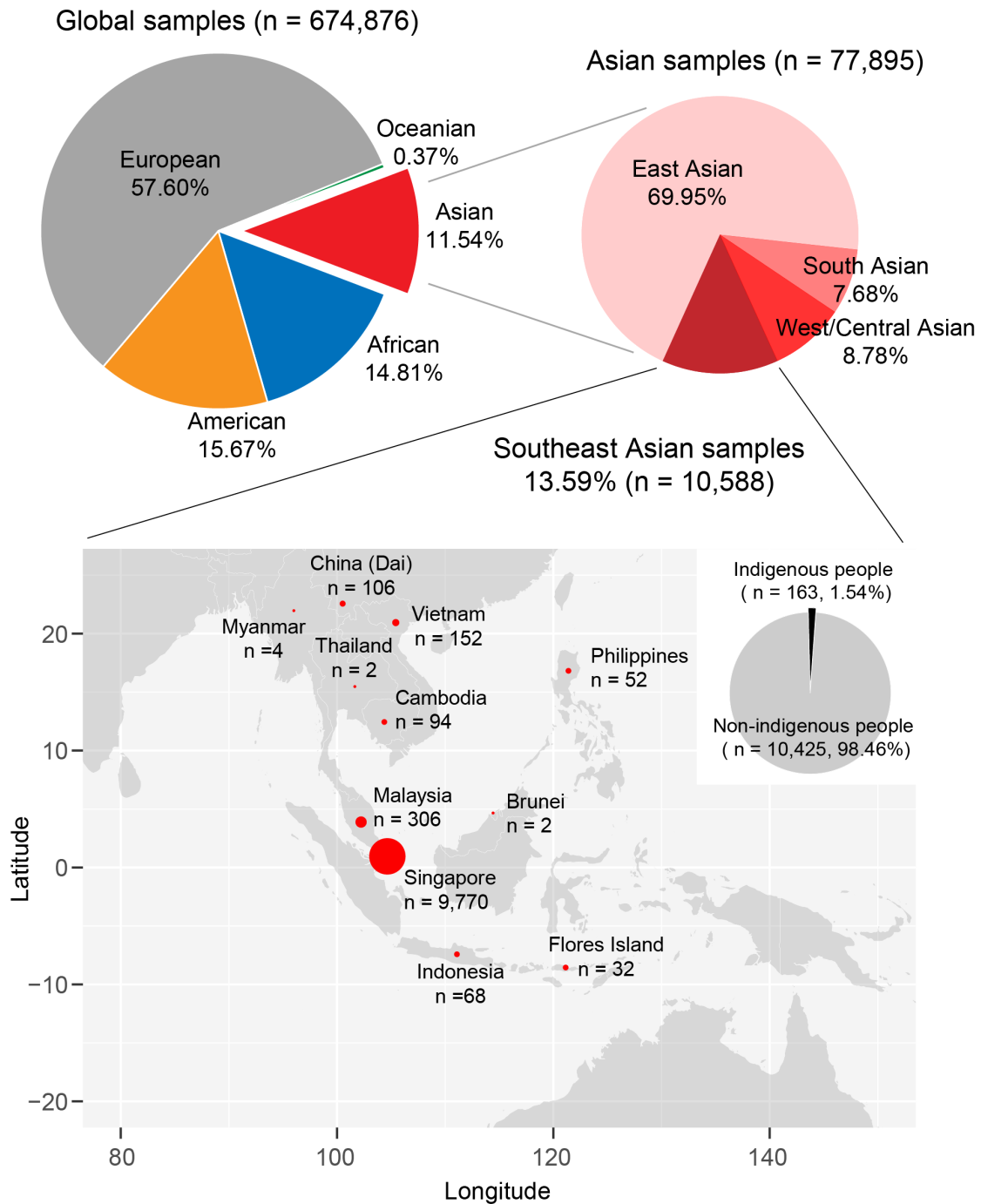
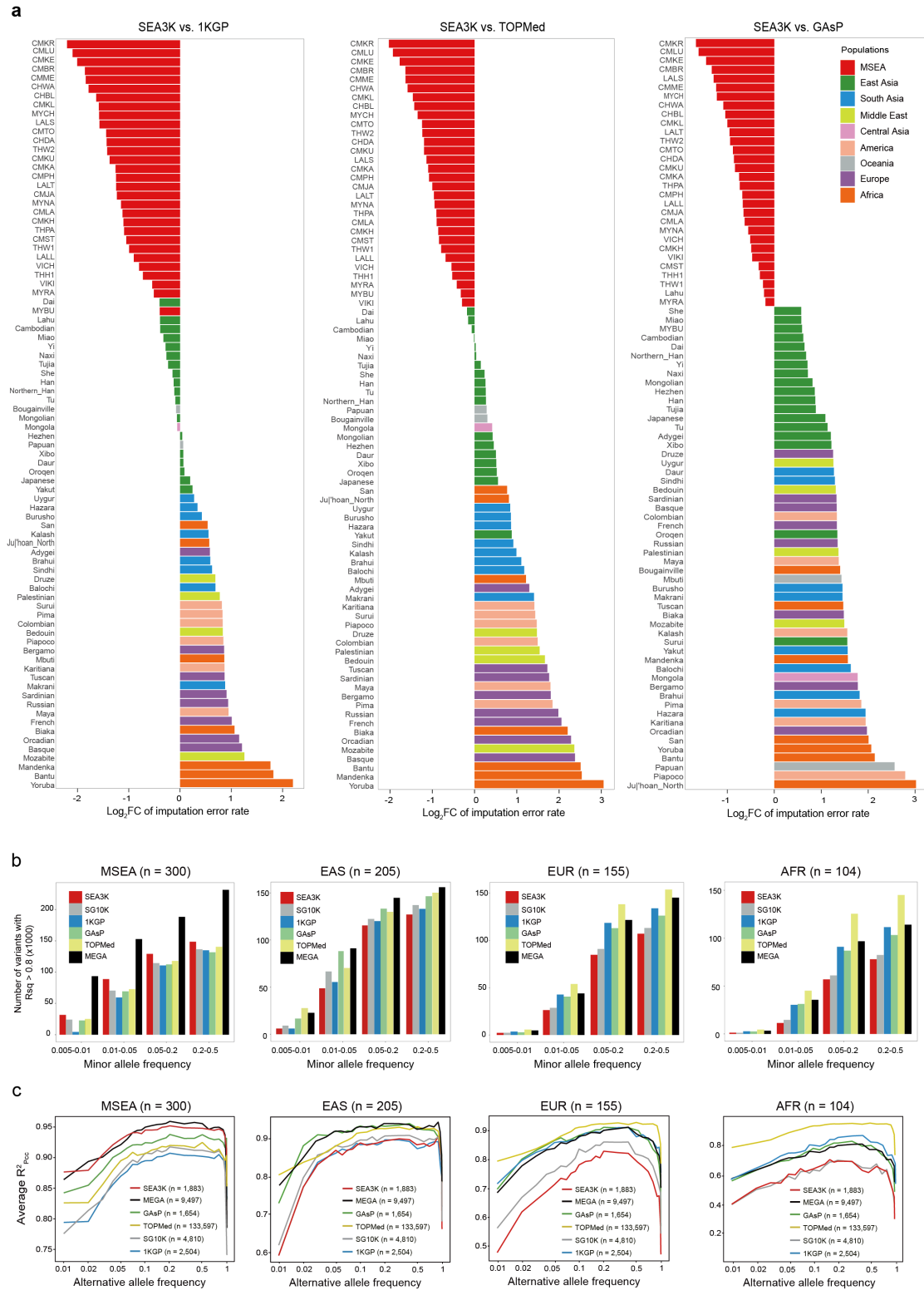

Supplementary information

Genome diversity and signatures of natural selection in mainland Southeast Asia

In the format provided by the
authors and unedited

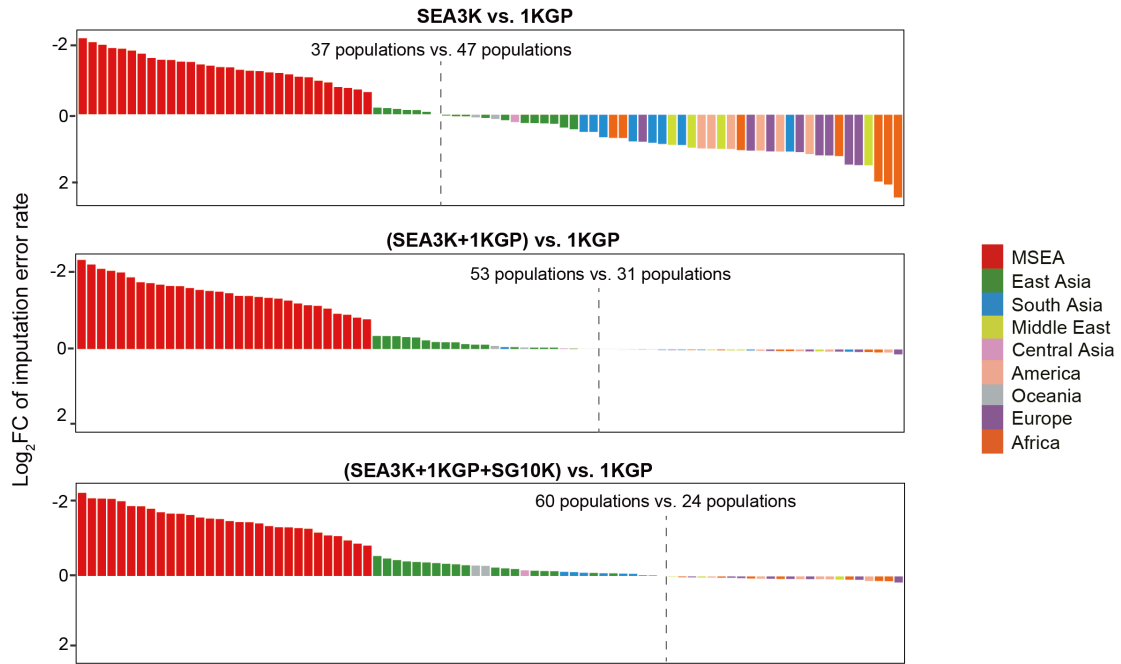


Supplementary Fig. 1 | Counts of human individuals with whole-genome sequencing data in global populations, Asia, and Southeast Asia. The data were retrieved from the published literatures (see Supplementary Table 1).

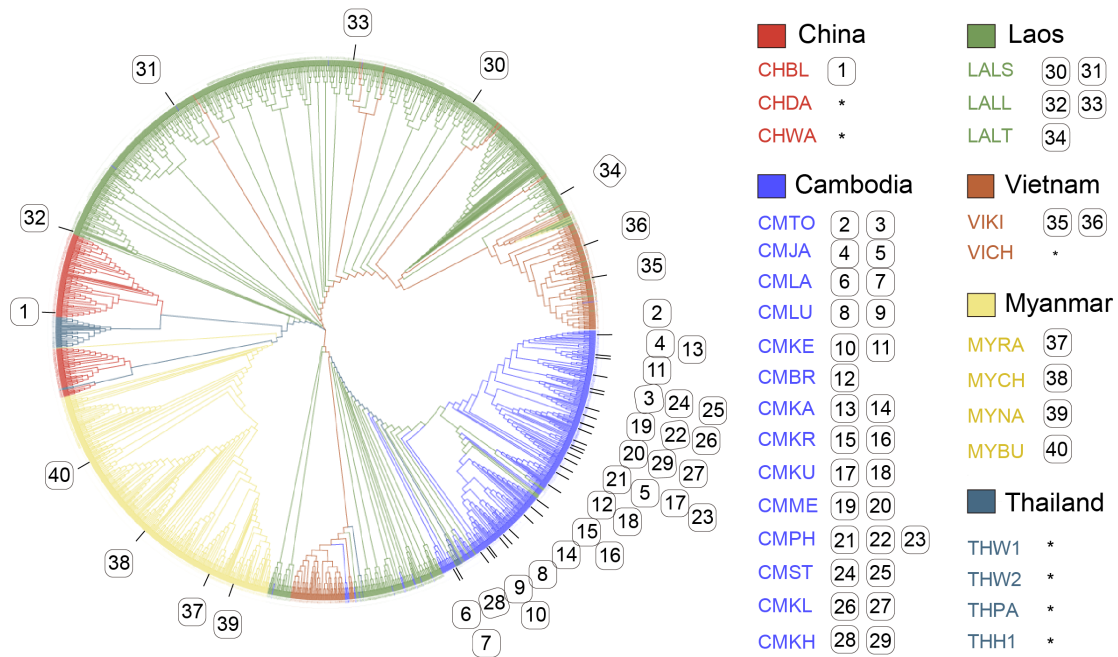


Supplementary Fig. 2 | Evaluation of imputation performance using the SEA3K reference panel. a, Fold change ($\log_2FC$) of imputation error rates for MSEA populations ($n = 300$, randomly selected from the SEA3K dataset, see Methods) and other worldwide populations (from HGDP dataset) based on the SEA3K and other reference panels. Each bar denotes a population and the colors refer to super-population groups. **b**, Number of high-quality imputed

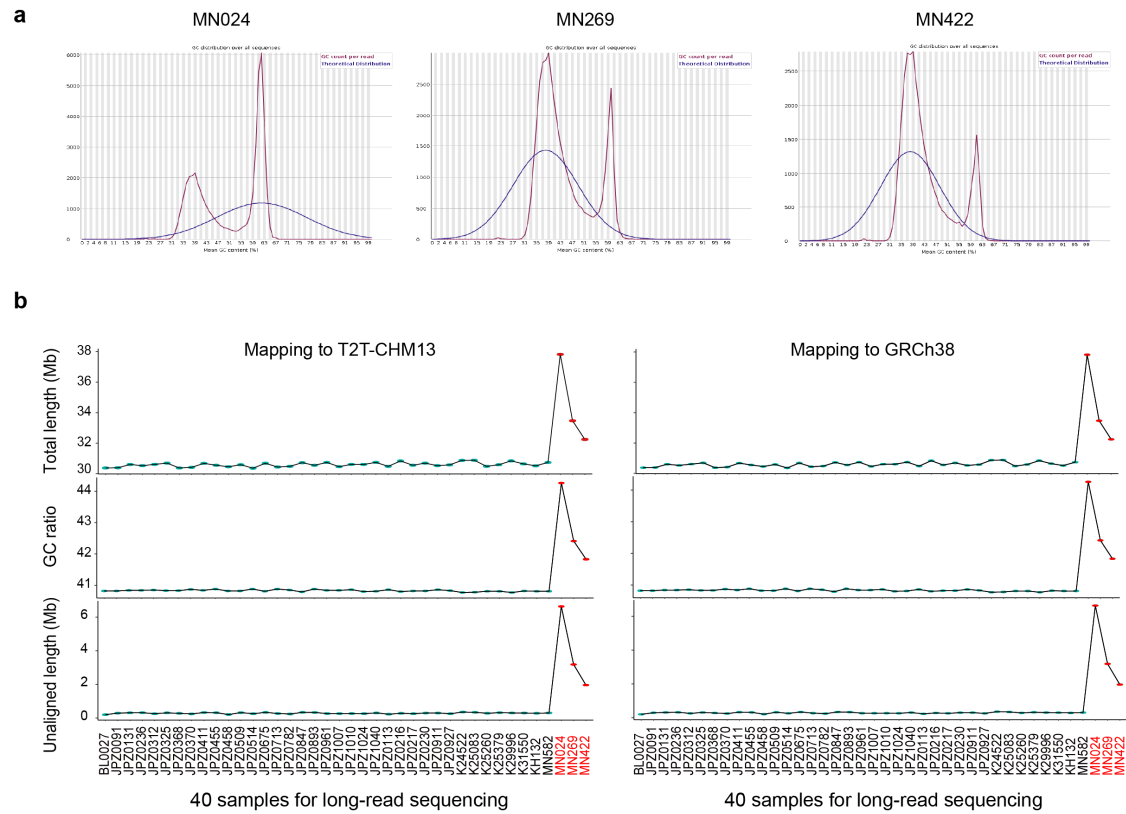
variants (based on the estimated imputation accuracy from Minimac4: $R^2_{\text{Est}} > 0.8$) in different minor allele frequency (MAF) bins for MSEA and global populations. The sample size of the target population is indicated, including 300 MSEA samples from SEA3K, 205 East Asian (EAS) individuals, 155 European (EUR) individuals and 104 African (AFR) individuals from HGDP dataset. The MEGA panel is the combination of the SEA3K, 1KGP and SG10K datasets (Methods). **c**, The imputation accuracy evaluation using squared Pearson correlation coefficient (R^2_{Pcc}) between genotypes called from the sequence reads and the imputed genotypes as a function of alternative allele frequency for six reference panels.



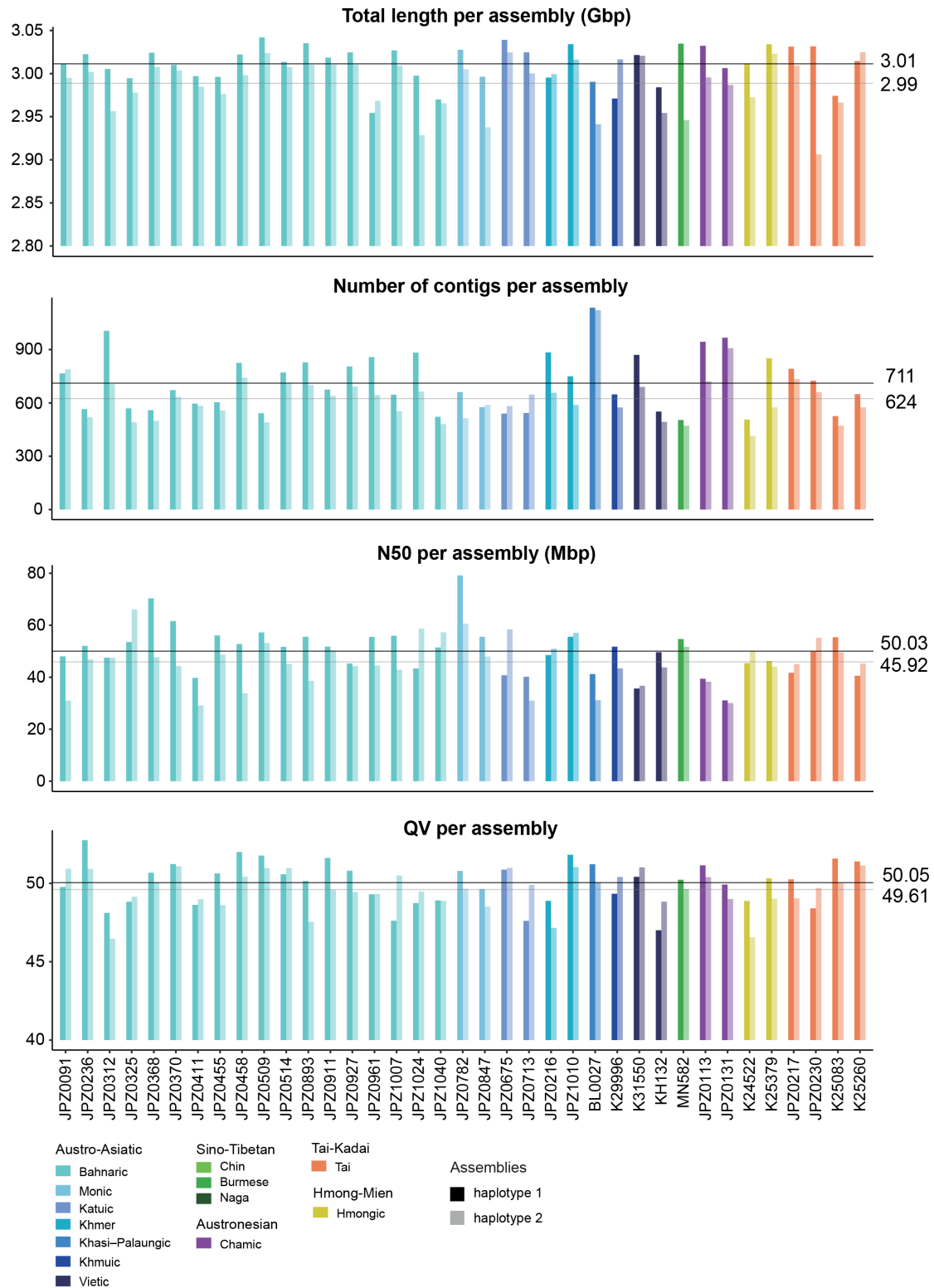
Supplementary Fig. 3 | Evaluation of imputation performance using different combinations of reference panels. Comparison of the fold changes ($\log_2\text{FC}$) of the imputation error rates for the worldwide populations using either the SEA3K, SEA3K+1KGP, and the MEGA panel (SEA3K+1KGP+SG10K) or the 1KGP panel. The population number with better imputation performance by SEA3K-included panels (SEA3K, SEA3K+1KGP, and MEGA) or 1KGP are indicated, respectively.



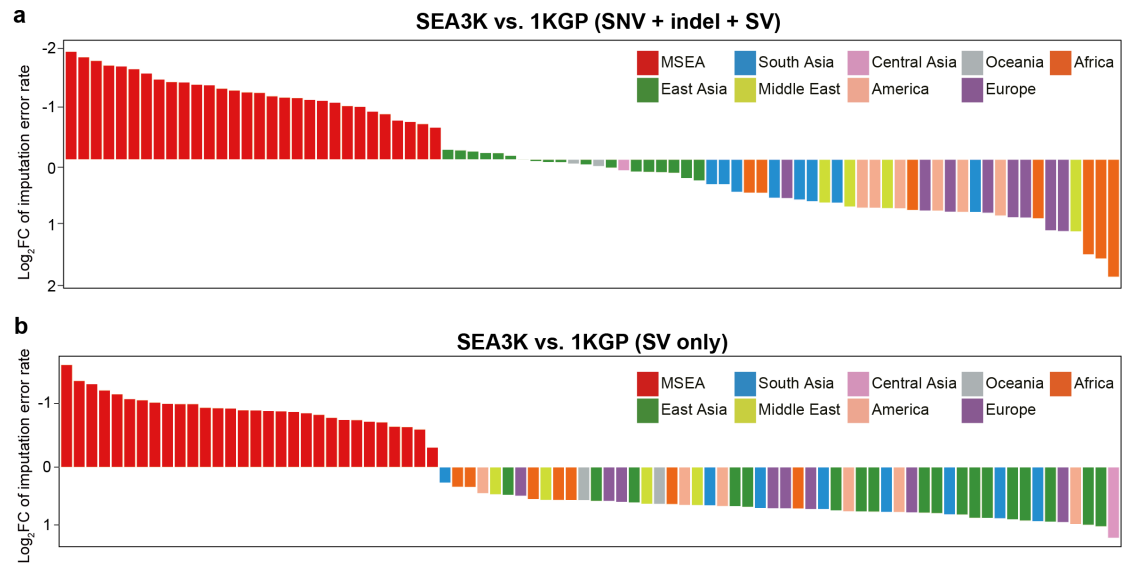
Supplementary Fig. 4 | The strategy of sample selection for long-read sequencing. The 40 samples were selected based on the genetic relationship (individual-based neighbor-joining tree by whole-genome SNVs). The countries and populations the samples belong to are colored and labeled in the legend. Considering the Cambodian samples are enriched in indigenous groups with high genetic diversity, we selected more Cambodian individuals into the long-read sequencing sample set. Seven populations of SEA3K were not included (labeled as * in the right legend: CHDA, CHWA, VICH, THW1, THW2, THPA and THH1) due to the poor DNA quality for long-read sequencing.



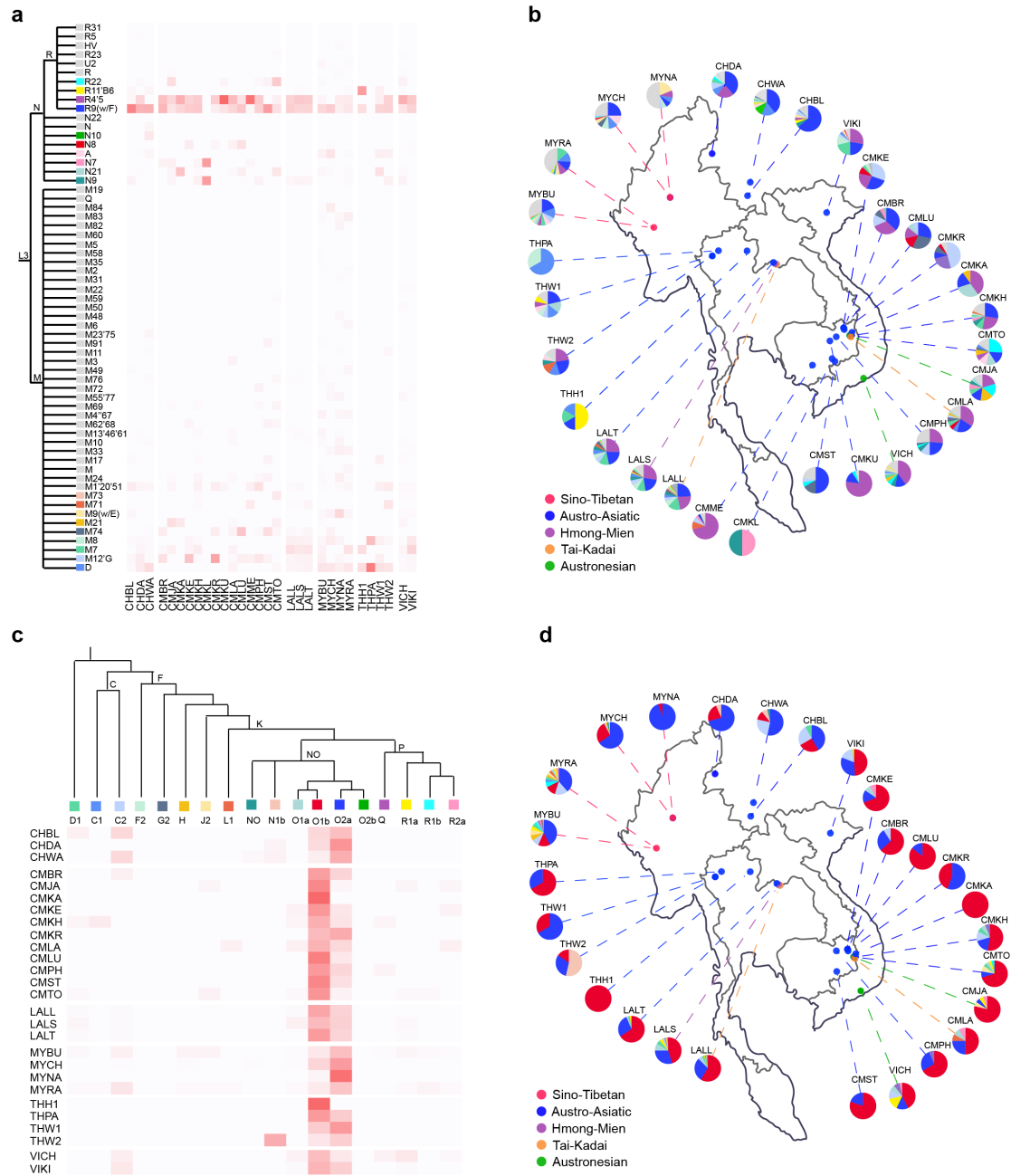
Supplementary Fig. 5 | The quality controls of three dropout samples (MN024, MN269 and MN422) in long-read sequencing. a, The plot showing the abnormal GC-content of three dropout samples in long-read sequencing. Compared to the theoretical distribution curve, the GC-content distributions of these three samples show additional peaks in mean value of 37% and 61%, suggesting a DNA contamination or low quality. **b**, The dot plots show the statistics of the assembled 40 MSEA genomes evaluated from Inspector tools (mapping to T2T-CHM13 and GRCh38 genome, respectively), including total length, GC-ratio and unaligned length per assembled genome. The three dropout samples are outliers.



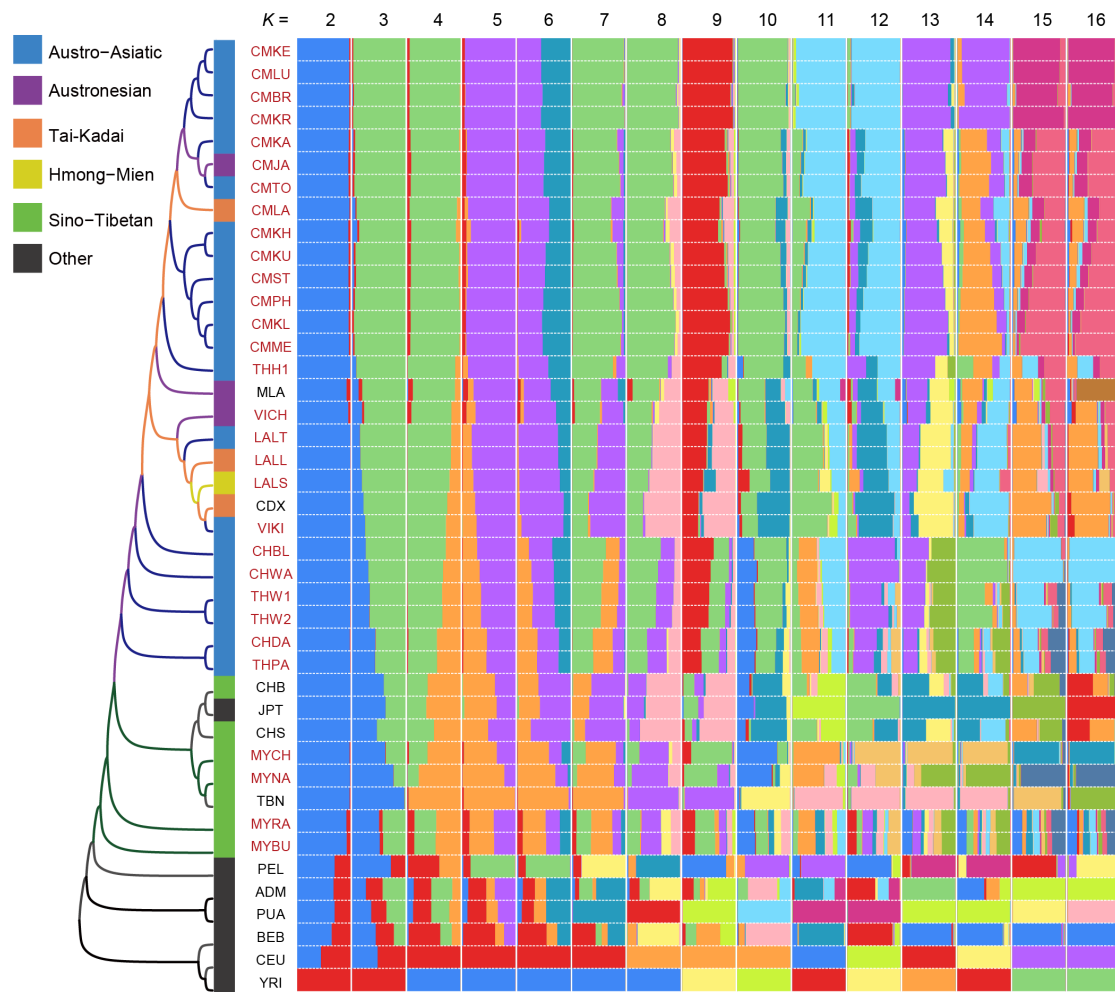
Supplementary Fig. 6 | Statistics of the 74 MSEA long-read assemblies. Each individual has two haploid assemblies (haplotype-1 and haplotype-2), plotted as two bars. A total of 74 haploid assemblies were produced for the 37 individuals. Total length, number of contigs, contig N50, and quality value (QV) per assembly is shown in the plots, respectively. The detailed statistics of the 74 assemblies are provided in Supplementary Table 6.



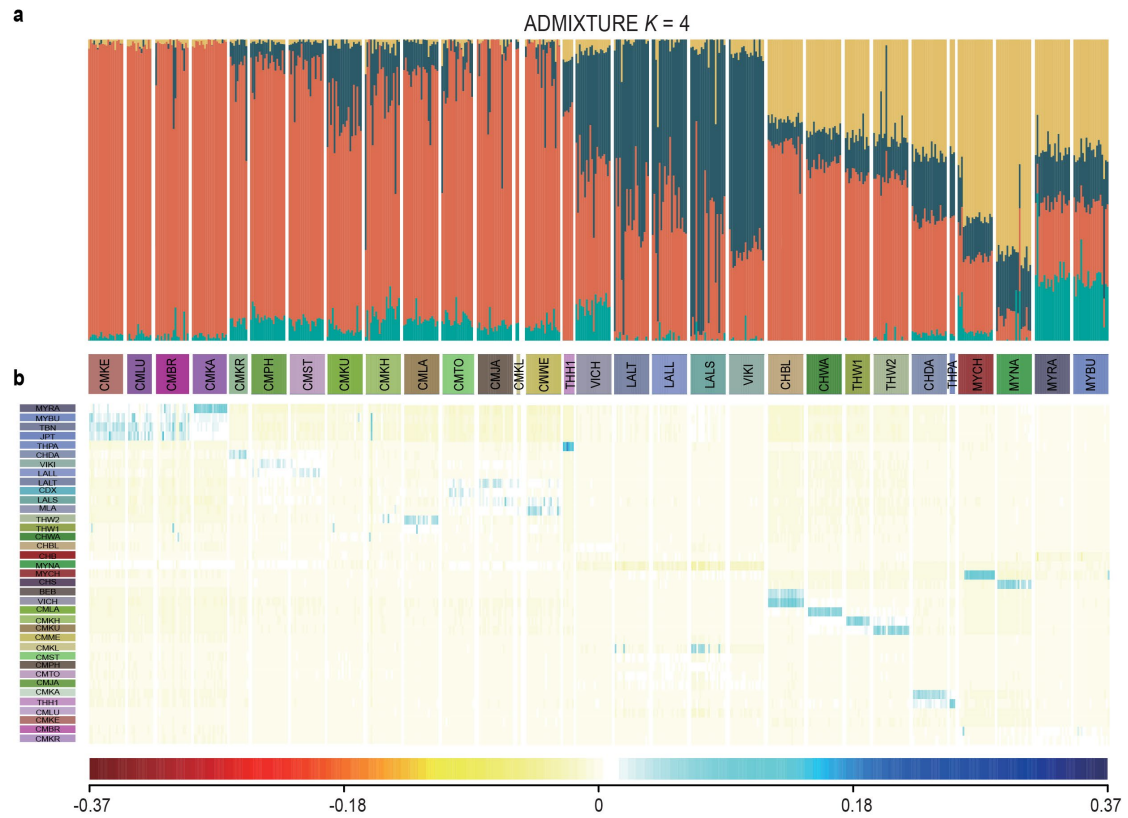
Supplementary Fig. 7 | Imputation performance of the SV-combined reference panel. Comparison of the fold changes ($\log_2\text{FC}$) of the imputation error rates for the worldwide populations using either the SEA3K panel or the 1KGP panel. **a**, Comparison of imputation performance for all variant combined (SNV + indel + SV). **b**, Comparison of imputation performance for SV only.



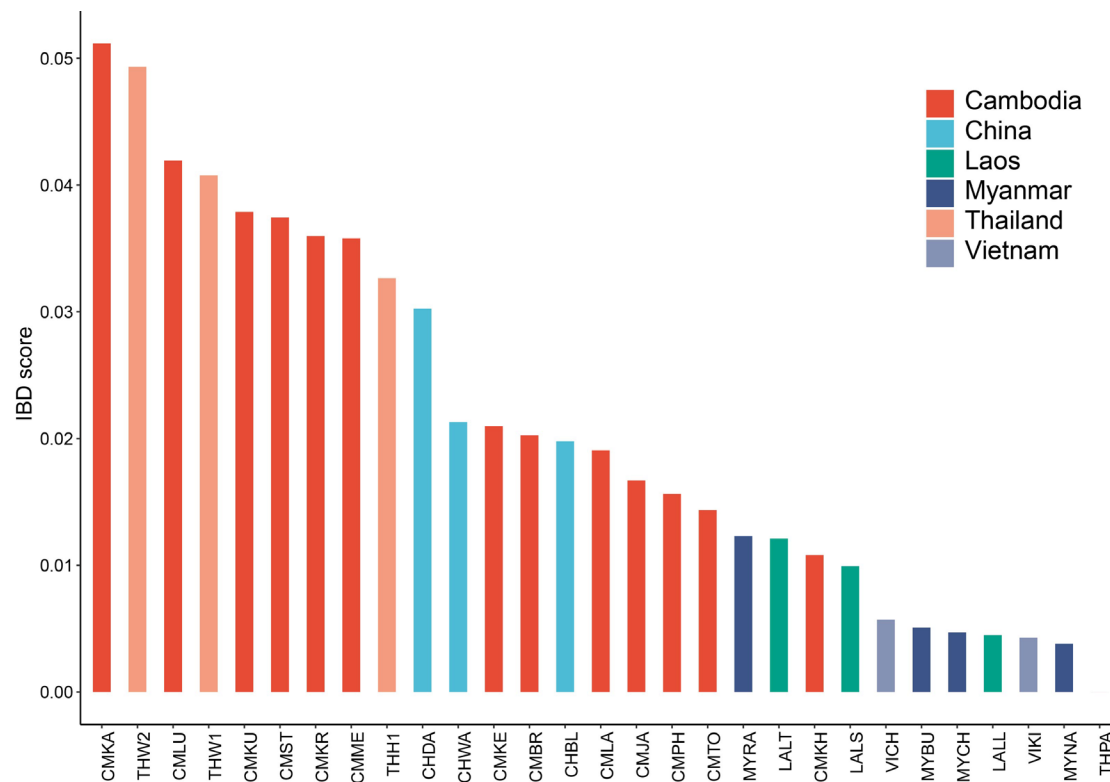
Supplementary Fig. 8 | Geographic distributions of mtDNA and Y chromosome haplogroups in MSEA populations. **a**, mtDNA haplogroup frequencies in 30 MSEA populations, and haplogroups with frequencies >10% in at least one population are marked in different colors (including D, M12'G, M7, M8, M74, M1'20'51, M21, M9, M71, N9, N21, N7, A, N8, N10, R9, B4'5, R11'B6 and R22), while the remaining are colored in gray. **b**, mtDNA haplogroup composition of each population in MSEA, the colors of haplogroups are consistent with those in panel-a. **c**, Y-chromosome haplogroup frequencies in 27 MSEA populations (three populations, i.e., CMKU (n = 0), CMKL (n = 1), CMME (n = 1), were excluded due to small sample sizes. Different haplogroups are indicated by different colors. **d**, Y-chromosome composition of each population in MSEA, the colors of haplogroups are consistent with those in panel-c.



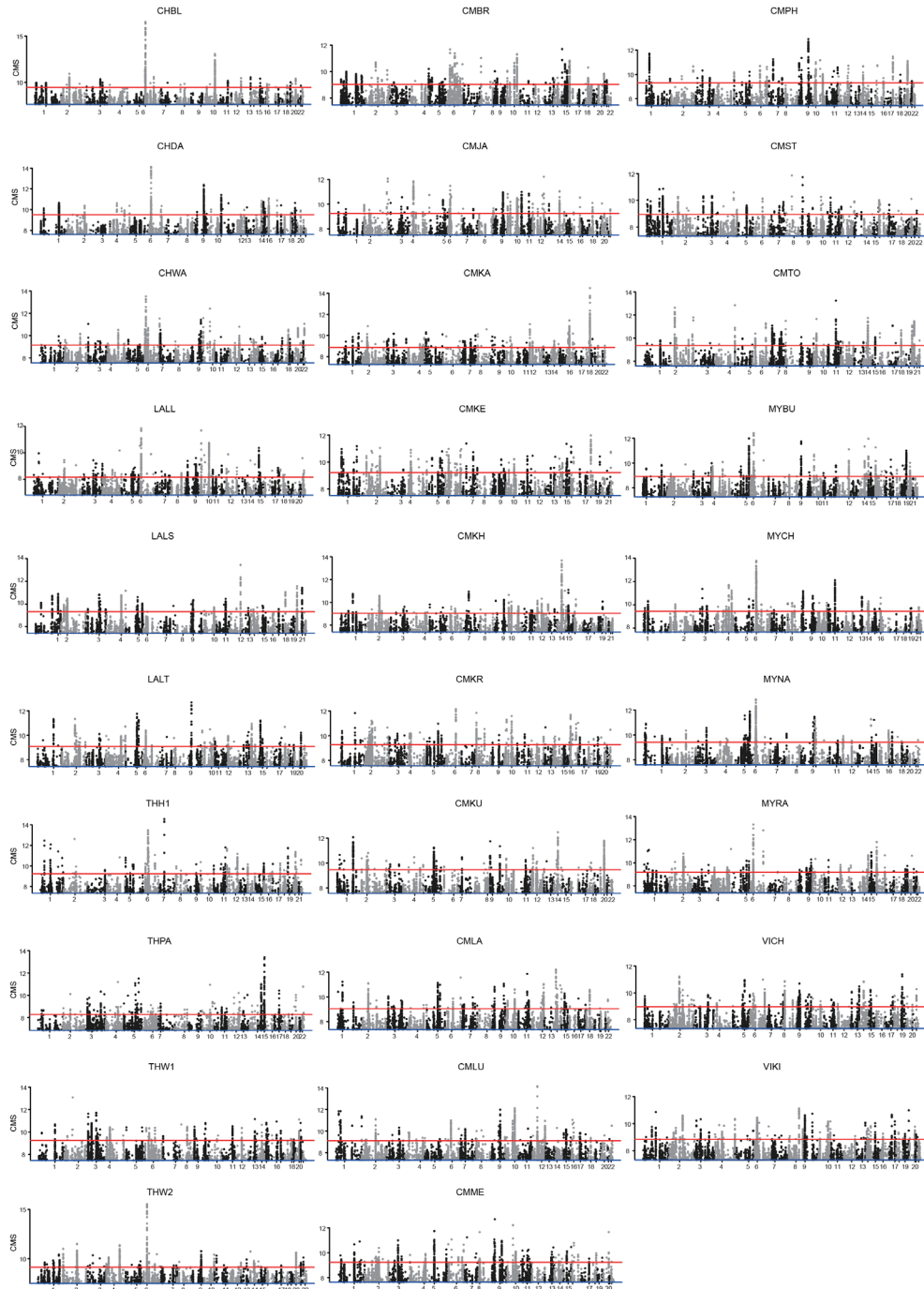
Supplementary Fig. 9 | ADMIXTURE clustering analysis of SEA3K and worldwide populations. The ML tree (left-panel) indicating the genetic relationship among the MSEA populations (highlighted in red) and other representative global populations from 1KGP and the published data (see Methods). The linguistic groups are indicated with colors as shown in the legend. The admixture results show K -value from 2 to 16 (right-panel). See Supplementary Table 2 for definitions of abbreviations.



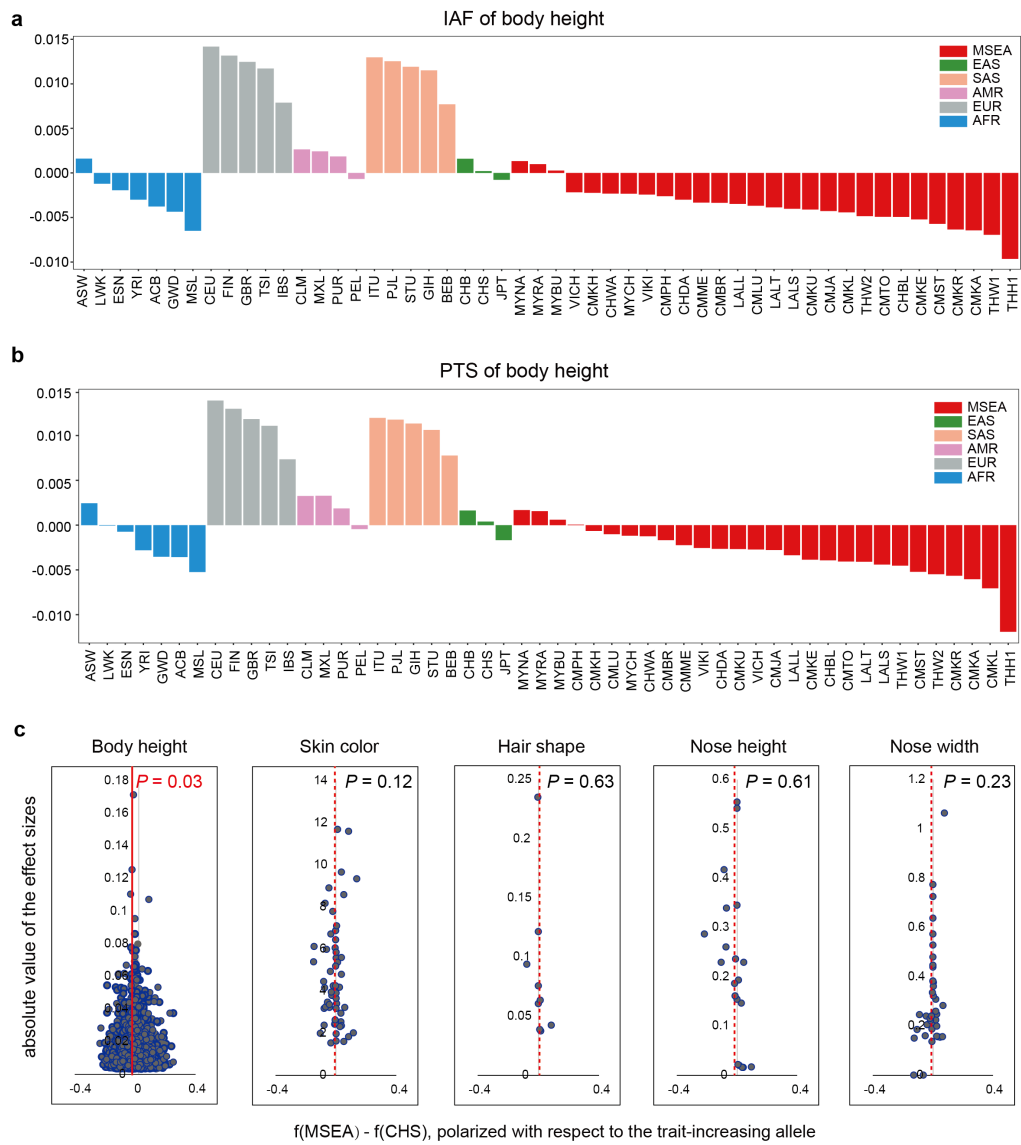
Supplementary Fig. 10 | Inspection of model fitting by badMIXTURE at $K = 4$. **a**, Ancestry proportions estimated through ADMIXTURE. **b**, The residual palette under the best-fit ancestral population admixture model. No systematic pattern of residuals was observed.



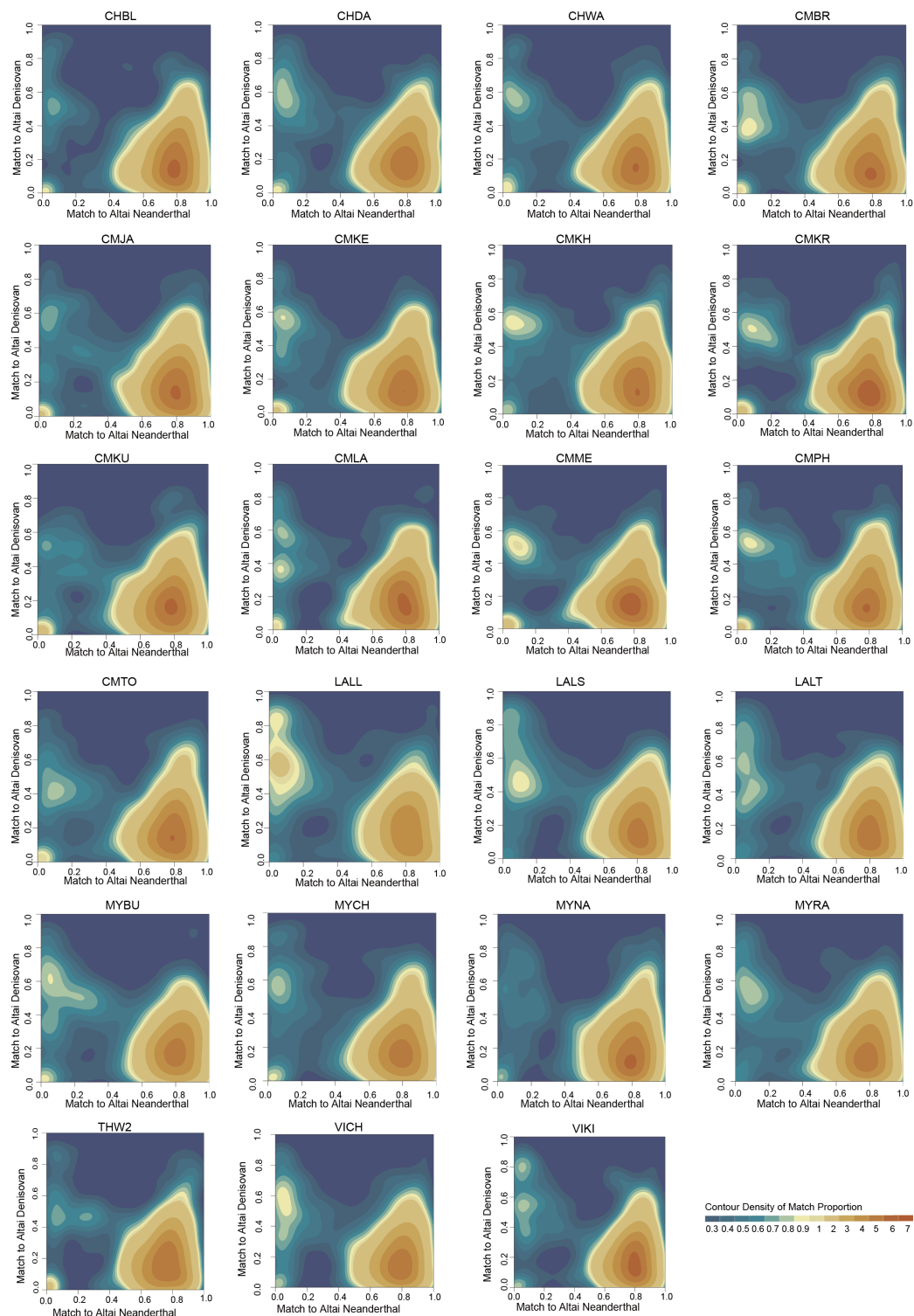
Supplementary Fig. 12 | The estimated IBD scores of MSEA populations. Populations with at least two samples were considered for IBD analysis, and only the unrelated samples were included in the analysis. The mean IBD score of each MSEA population was estimated by averaging the IBD-matrix within the population.



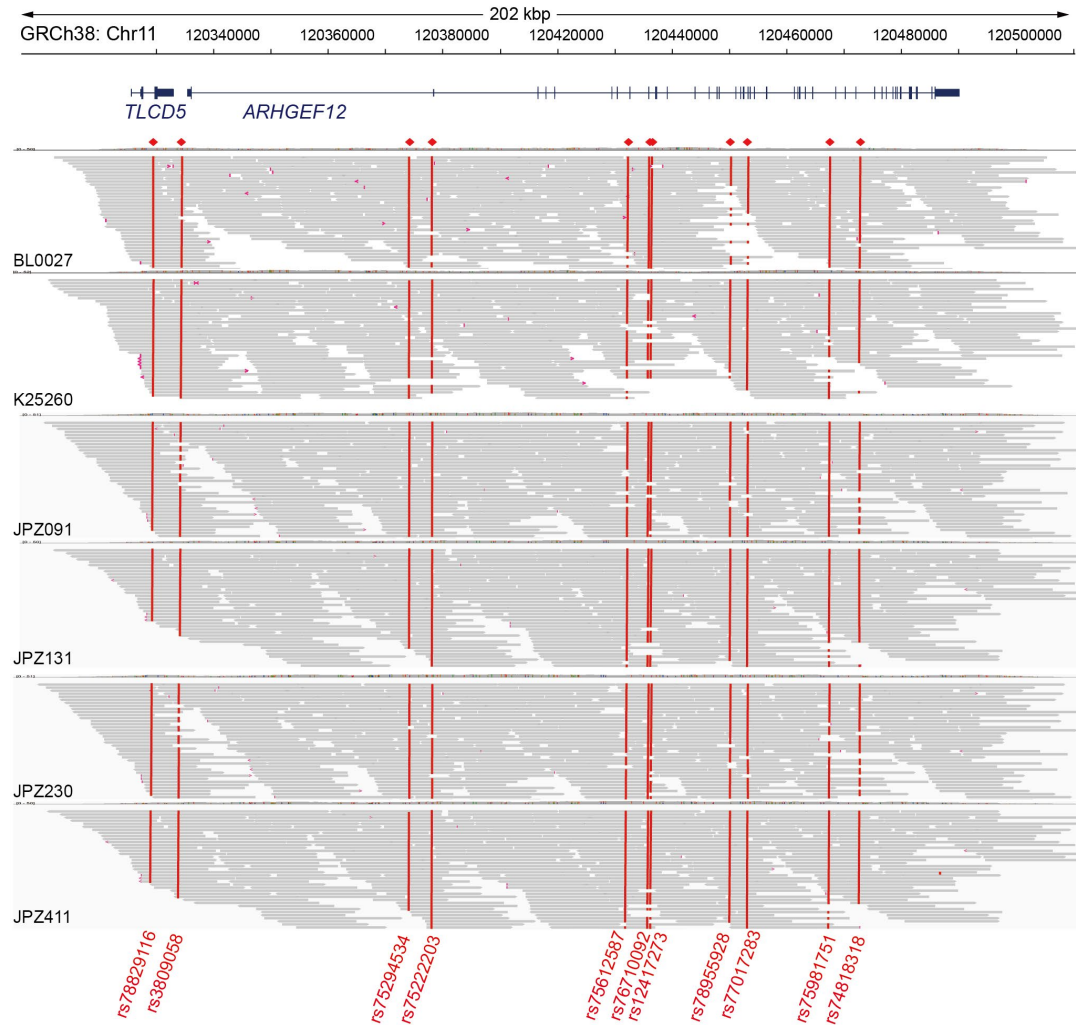
Supplementary Fig. 13 | Manhattan plot of the calculated CMS scores based on the genome-wide SNVs in each MSEA population. Only the signals falling in the top 0.1% of genome-wide CMS scores in each population are dotted. The red line denotes the top 0.01% cutoff. The plots show the population-shared, the population-specific and the indigenous-population-specific signals. To characterize the CMS signals in the entire MSEA populations, we averaged the CMS score of each SNV among the 29 investigated MSEA populations, by calculating an overall p-value by chi-squared distribution. Then we defined the 44 independent candidate selection signals by setting up the threshold of the top 0.01% maximal CMS values in the genome and LD-based clumping (Methods).



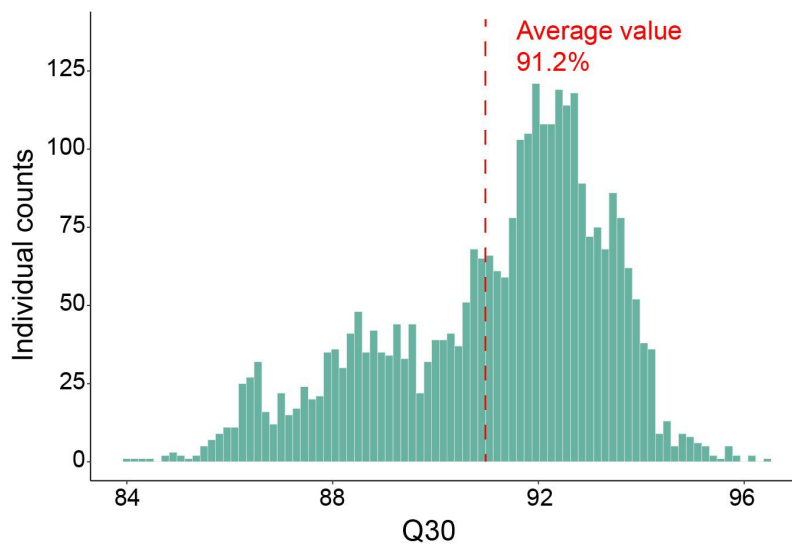
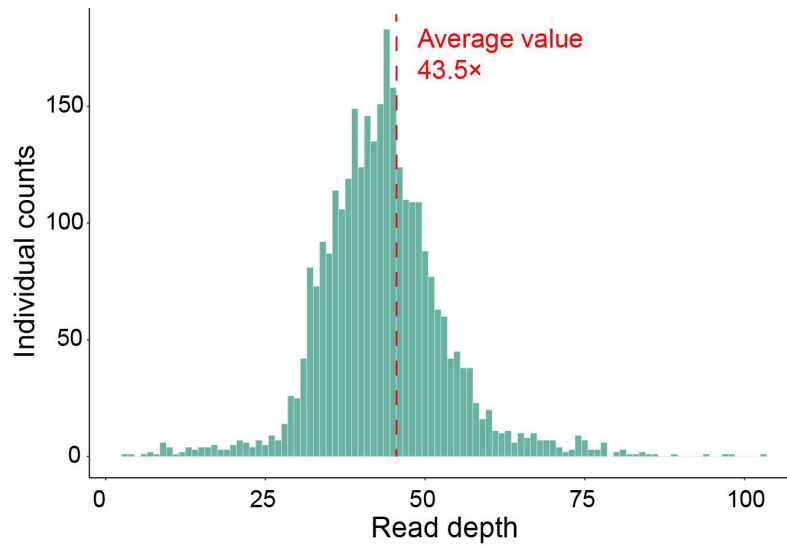
Supplementary Fig. 14 | Detection of polygenic effect on multiple traits in MSEA populations. **a**, Comparison of height-increasing allele frequency (IAF) among world populations. The great majority of MSEA populations show negative values. **b**, Polygenic trait scores (PTS) across populations show a similar pattern with IAF. The plots show the mean IAF and PTS of 9,972 independent height-associated SNVs among MSEA and other world populations in 1KGP, and the height-increasing allele was defined as the allele significantly associated with increased height in the GWAS data set. **c**, The absolute values of the effect sizes of the trait-associated SNVs for body height, skin color, hair shape, nose height and nose width, as a function of the difference in frequency between MSEA and CHS (southern Han Chinese): $f(\text{MSEA}) - f(\text{CHS})$. The panel frequencies are polarized with respect to the trait-increasing allele. The vertical lines refer to the mean value, which solid line and dashed lines denote to significant and non-significant P -value under 1000 permutation tests (one-sided). For body height, the mean value of $f(\text{MSEA}) - f(\text{CHS})$ is -0.00163, which reaches statistical significance under 1000 permutation tests (P -value = 0.032). This result suggests that the sum of height-increasing allele frequency is significantly depleted in MSEA populations.



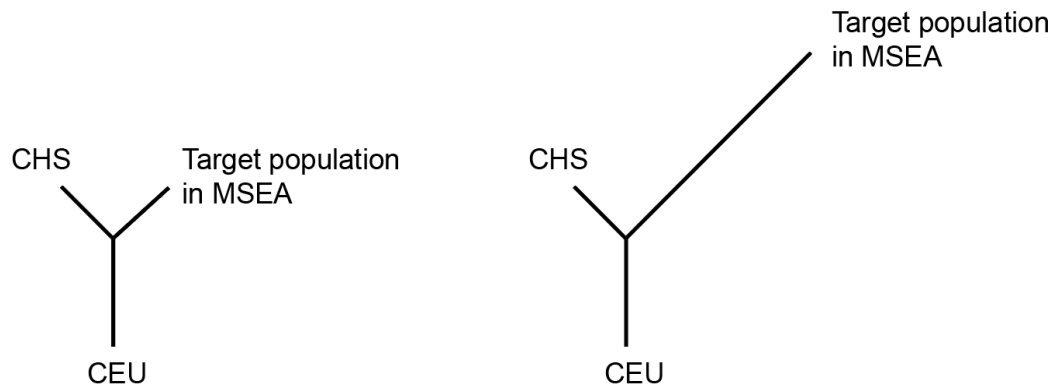
Supplementary Fig. 15 | Contour density plots of match proportion of the introgressed segments from the Altai Neanderthal and Altai Denisovan Genomes in each population of MSEA. All analyses were performed on autosomes, and YRI of 1KGP was taken as the outgroup in analyses. Only populations with sample size larger than 30 were included in this analysis, thus 1,976 individuals from 23 populations in SEA3K were included in the analysis (Methods). Numbers inside the plots indicate the height of the density corresponding to each contour line.



Supplementary Fig. 16 | Validation of the adaptive archaic introgression haplotype using the HiFi reads. The plot shows a Neanderthal-introgressed region on Chr11 (~150 kbp), containing the *ARHGEF12* and *TLCD5* genes. We identified 441 individuals in SEA3K exhibiting the archaic-introgression haplotype formed by 11 Neanderthal-derived PS-SNVs, and six individuals were included in the long-read sequencing panel of 37 individuals. The plots show the HiFi-read alignments of these six individuals to GRCh38 genome, and the 11 Neanderthal-derived PS-SNVs are denoted by rhombus in red and labeled at the bottom of the plot.



Supplementary Fig. 17 | Data coverage (read depth) and data quality (Q30) of the 3,023 MSEA individuals. We counted the read depth per sample for the 3,023 samples using mosdepth tool for bam file from initial mapping. Histogram shows the mean depth of all samples is 43.5x. Phred-based quality scores (Q-scores) are used to measure the accuracy of nucleotide identity data from a sequencing run. A higher quality score indicates a lower probability that an individual base is called incorrectly. Q30 indicates the probability of an incorrect base call of 1 in 1000. The green proportion of the graph represents bases with a high-quality score (defined as Q30) identified during the run. Histogram shows the mean Q30 of all MSEA samples is 91.2.



Supplementary Fig. 18 | The diagram showing the logic of detecting signal of positive selection in a target MSEA population using the PBS method. The left-panel shows the pattern of neutral evolution (no selection) with the expected F_{ST} -based branch lengths among populations, and the right-panel shows the pattern of Darwinian positive selection, with a substantial differentiation of the target MSEA population.

Members of Consortium of Anthropological Research in Southeast Asia and Southwest China (CASEAC)

Bing Su^{1,2,33}, Yaoxi He^{1,2}, Xiaoming Zhang^{1,2}, Xueping Ji³⁵, Jian-Hua Wang²¹, Hai-Mei Shen²¹, Min-Sheng Peng^{1,3,4}, Qingpeng Kong^{1,3,5,33}, Yuchun Li^{1,3,5}, Jianxin Guo¹, Ying-Hua Li³⁶, Yun Wu³⁷, Chuan-Chao Wang^{38,39,40,41}, Menghan Zhang^{42,43}, Yafei Mao^{17,18}, Lu Chen⁶, Xuebin Qi⁴⁴, Hong Shi⁴⁴, Hui Zhang⁴⁴, Tuot Sovannary²², Long Bunnath²², Heng Sophady²², Jatupol Kampuansai²³, Metawee Srikummool²⁴, Wibhu Kutanan²⁵, Sissades Tongsima⁴⁵, Rasmi Shoocongdej⁴⁶, Huy Ho Quang²⁶, Khoa Pham Dang²⁶, Lâm Thị Mỹ Dung⁴⁷, Sommay Singthong²⁷, Somphad Sochampa²⁸, Xai Xiong⁴⁸, Tany Sengvilai⁴⁹, Somchin Sitthideth⁵⁰, U Win Kyaing²⁹

¹State Key Laboratory of Genetic Evolution & Animal Models, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650201, China

²Yunnan Key Laboratory of Integrative Anthropology, Kunming 650201, China

³KIZ/CUHK Joint Laboratory of Bioresources and Molecular Research in Common Diseases, Kunming 650201, China

⁴University of Chinese Academy of Sciences, Beijing 100049, China

⁵Kunming Key Laboratory of Healthy Aging Study, Kunming 650201, China

⁶State Key Laboratory of Genetic Engineering, Center for Evolutionary Biology, School of Life Science, Fudan University, Shanghai 200438, China

⁷National Genomics Data Center, China National Center for Bioinformation, Beijing 100101, China

⁸School of Computer Science and Technology, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China

⁹MOE Key Lab for Intelligent Networks & Networks Security, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China

¹⁰Center for Mathematical Medical, the First Affiliated Hospital, Xi'an Jiaotong University, Xi'an 710061, China

- ¹¹School of Automation Science and Engineering, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China
- ¹²Genome Institute, the First Affiliated Hospital, Xi'an Jiaotong University, Xi'an 710061, China
- ¹³School of Life Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China.
- ¹⁴Faculty of Science, Leiden University, Leiden, 2311EZ, The Netherlands
- ¹⁵CEMS, NCMIS, HCMS, MADIS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
- ¹⁶School of Medicine, Yunnan University, Kunming 650091, China
- ¹⁷Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders, Ministry of Education, Shanghai Jiao Tong University, Shanghai 200230, China
- ¹⁸Center for Genomic Research, International Institutes of Medicine, The Fourth Affiliated Hospital, Zhejiang University, Yiwu 322000, China
- ¹⁹Laboratory Animal Center, Kunming Institute of Zoology, the Chinese Academy of Sciences, Kunming 650201, China
- ²⁰National Resource Center for Non-Human Primates, Kunming 650201, China
- ²¹Department of Anthropology, School of Sociology, Yunnan Minzu University, Kunming 650504, China
- ²²Department of Geography and Land Management, Royal University of Phnom Penh, Phnom Penh 12000, Cambodia
- ²³Department of Biology, Faculty of Science, Chiang Mai University, Chiang Mai 50200, Thailand
- ²⁴Department of Biochemistry, Faculty of Medical Science, Naresuan University, Phitsanulok 65000 Thailand
- ²⁵Department of Biology, Faculty of Science, Naresuan University, Phitsanulok, 65000 Thailand
- ²⁶Department of Immunology, Ha Noi Medical University, Ha Noi, Vietnam
- ²⁷The National University of Laos (NUOL), P.O.Box 7322, Dongdok, Vientiane,

Vientiane, Laos

²⁸Muang Sing District Hospital, Luang Namtha province, Laos

²⁹Field School of Archaeology, WG6X+6XV Twin Bye, Pyay-Paukkhaung Road, Paukkhaung, Myanmar

³⁰Queen Sirikit Botanic Garden (QSBG), The Botanical Garden Organization, Chiang Mai, 50200 Thailand

³¹Inter Mountain Peoples Education and Culture in Thailand Association (IMPECT) 252 Moo 2, T.Sansai Noi, Sansai, 50210, Thailand

³²Asia Indigenous Peoples Pact (AIPP) 112 Moo 1, Tambon Sanpranate, Amphur Sansai, Chiang Mai 50210, Thailand

³³Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China

³⁵Kunming Natural History Museum of Zoology, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China

³⁶Archaeological Institute for Yangtze Civilization, Wuhan University, Wuhan, 430072, China

³⁷Department of Paleoanthropology, Yunnan Institute of Cultural Relics and Archaeology, Kunming 650118, China

³⁸State Key Laboratory of Cellular Stress Biology, School of Life Sciences, Xiamen University, Xiamen 361102, China

³⁹Department of Anthropology and Ethnology, Institute of Anthropology, School of Sociology and Anthropology, Xiamen University, Xiamen 361005, China

⁴⁰State Key Laboratory of Marine Environmental Science, Xiamen University, Xiamen 361102, China

⁴¹Ministry of Education Key Laboratory of Contemporary Anthropology, Department of Anthropology and Human Genetics, School of Life Sciences, Fudan University, Shanghai 200433, China

⁴²State Key Laboratory of Genetic Engineering, and Collaborative Innovation Center for Genetics and Development, School of Life Sciences, Fudan University, Shanghai, China

⁴³Institute of Modern Languages and Linguistics, Fudan University, Shanghai, China

⁴⁴State Key Laboratory of Primate Biomedical Research, Institute of Primate Translational Medicine, Kunming University of Science and Technology, Kunming, Yunnan 650500, China

⁴⁵National Biobank of Thailand (NBT), National Science and Technology Development Agency (NSTDA), Bangkok, 12120 Thailand

⁴⁶Department of Archaeology, Faculty of Archaeology, Silpakorn University, Bangkok 10200 Thailand

⁴⁷Vietnamese Association of Archaeology, No. 1, Trang Tien - No. 216 Tran Quang Khai, Hoan Kiem, Hanoi, Vietnam

⁴⁸Ban nasala, Saythani District, Vientiane Capital, Laos

⁴⁹Ban Saphangthong, Sisattanak District, Vientiane Capital, Laos

⁵⁰Ban vangheua, Viangkham District, Luang prabang province, Laos