

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (<i>n</i>) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided <i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted <i>Give P values as exact values whenever suitable.</i>
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i>), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	NA
Data analysis	FastQC (v0.11.9), https://github.com/s-andrews/FastQC ; BWA (v0.7.17-r1188), https://bio-bwa.sourceforge.net/ ; Picard (v2.27.4), https://broadinstitute.github.io/picard/ ; SAMtools (v1.16.1), https://samtools.sourceforge.net/ ; GATK4 (v4.2.6.1), https://gatk.broadinstitute.org/ ; VCFtools (v0.1.16), https://github.com/vcftools/vcftools/ ; EIGENSOFT (v3.0), https://github.com/argriffing/eigensoft ; VEP (v110), https://asia.ensembl.org/info/docs/tools/vep/index.html ; SnpEff (v5.1-2), https://pcingola.github.io/SnpEff/ ; SnpSift (v4.5), https://github.com/pcingola/SnpSift ; LOFTEE (v1.0.4), https://github.com/konradjk/loftee ; KING (v2.2.8), https://www.kingrelatedness.com/ ; LiftOver (v1.1.17), https://genome.ucsc.edu/cgi-bin/hgLiftOver ; PLINK2 (v2.0), https://www.cog-genomics.org/plink/2.0/ ; ADMIXTURE (v1.3.0), https://dalexander.github.io/admixture/ ; ProcrustesAnalysisT (v1.0.0), https://github.com/YurunLu/ProcrustesAnalysisT ; badMIXTURE (v1.0.0), https://github.com/danjlawson/badMIXTURE ; Treemix (v1.12), https://github.com/carolindahms/TreeMix ; MEGA (v11.0), https://www.megasoftware.net/ ;

BCftools (v1.9), <https://github.com/samtools/bcftools>;
 POPLDdecay (v3.43), <https://github.com/BGI-shenzhen/PopLDdecay>;
 Haplogrep2 (v2.4.0), <https://github.com/seppinho/haplogrep-cmd>;
 Y-LineageTracker (v1.0.0), <https://github.com/Shuhua-Group/Y-LineageTracker>;
 MSMC2 (v5.1.1), <https://github.com/stschiff/msmc2>;
 SMC++ (v1.15.2), <https://github.com/popgenmethods/smcpp>;
 Selscan (v2.0.3), <https://github.com/szpiech/selscan>;
 PolyGraph (v1.0.0), <https://github.com/FerRacimo/PolyGraph>;
 Haploview (v1.0.0), <https://www.broadinstitute.org/haploview/haploview>;
 DnaSP (v6), <http://www.ub.edu/dnasp/>;
 Arlequin (v35), <https://cmpg.unibe.ch/software/arlequin35/>;
 PopART (v1.7.1 beta), <https://github.com/jessicawleight/popart>;
 PacBio SMRT-Analysis package, <https://www.pacb.com/products-and-services/analytical-software/smrt-analysis/>;
 Minimap2 (v2.2), <https://github.com/lh3/minimap2>;
 Sniffles (v2.2), <https://github.com/fritzsiedlazeck/Sniffles>;
 PAV (v2.3.4), <https://github.com/EichlerLab/pav>;
 SVision-pro (v2.4), <https://github.com/songbowang125/SVision-pro>;
 Jasmine (v1.1.5), <https://github.com/mkirsche/Jasmine/>;
 BEDtools (v2.31.0), <https://bedtools.readthedocs.io/en/latest/>;
 Paragraph (v2.4), <https://github.com/Illumina/paragraph>;
 Hifiasm (v0.19.5), <https://github.com/chhylp123/hifiasm>;
 QUAST (v5.2.0), <https://github.com/ablab/quast>;
 NextPolish2 (v0.2.0), <https://github.com/Nextomics/NextPolish2>;
 Inspector (v1.0.1), <https://github.com/Maggi-Chen/Inspector>;
 SPrime (v20May22.855), <https://github.com/browning-lab/sprime>;
 IBDmix (v1.0.1), <https://github.com/PrincetonUniversity/IBDmix>;
 SHAPEIT4 (v4.2.2), <https://odelaneau.github.io/shapeit4/>;
 EAGLE2 (v2.4), <https://alkesgroup.broadinstitute.org/Eagle/>;
 Minimac4 (v4.1.3), <https://github.com/statgen/Minimac4>;
 Ggtree (v1.0.0), <https://github.com/YuLab-SMU/ggtree>;
 Mclust (v1.0.0), <https://github.com/mclust-org>;
 hclust (v1.0.0), <https://github.com/knightjdr/hclust>;
 TOPMed Imputation Server (v2.0.0-beta3-tis), <https://topmedimpute.readthedocs.io/>;
 Michigan Imputation Server (v2), <https://imputationserver.sph.umich.edu/>;
 ccs (v6.3.0), <https://github.com/PacificBiosciences/ccs>

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Datasets generated in this study were deposited in public repositories. WGS data are archived at Genome Sequence Archive for human (GSA-Human: <https://ngdc.cncb.ac.cn/gsa-human/>) under the accession number: HRA007135; Genome assemblies are archived at Genome Warehouse (GWH: <https://ngdc.cncb.ac.cn/gwh/>) under the accession number: PRJCA028104; Variant data are archived at Genome Variation Map (GVM: <https://ngdc.cncb.ac.cn/gvm/>) under the accession number: GVM000730; To protect participants confidentiality, the raw sequencing data are available to the scientific community for general research through a controlled access process. Access can be requested by submitting an application that includes a detailed research proposal and an IRB approval from the applicant's home institute to the Data Access Committee of Kunming Institute of Zoology, Chinese Academy of Sciences (KIZ, CAS). The other data are open access.

Datasets obtained from publicly available sources include: human reference genome GRCh38, https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38_reference_genome/GRCh38_full_analysis_set_plus_decoy_hla.fa
 human reference genome T2T-CHM13, https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/analysis_set/chm13v2.0.fa.gz;
 1KGP phase 3 dataset, <https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/>
 high-coverage 1KGP dataset, http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20201028_3202_raw_GT_with_annot/
 HGDP dataset, https://rosenberglab.stanford.edu/data/conradEtAl2006/data1_1.tar.gz
 SGDP dataset: https://sharehost.hms.harvard.edu/genetics/reich_lab/sgdp/vcf_variants/vcfs.variants.public_samples.279samples.tar
 HGSVC3 (freeze4), http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC3/working/20240415_Freeze4/;
 HPRC (v1.1), <https://s3-us-west-2.amazonaws.com/human-pangenomics/pangenomes/freeze/freeze1/minigraph-cactus/hprc-v1.1-mc-grch38/hprc-v1.1-mc-grch38.vcf.gz>;
 Genome of Altai Neanderthal and Denisovan: <https://www.eva.mpg.de/genetics/genome-projects/>;
 RefSeq genes, <https://www.cog-genomics.org/static/bin/plink/glist-hg38>;
 GWAS summary data, <https://www.ebi.ac.uk/gwas/api/search/downloads/full>;
 cCREs from ENCODE, <https://downloads.wenglab.org/V3/GRCh38-cCREs.bed>;
 ClinVar, https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/
 OMIM database, <https://www.ncbi.nlm.nih.gov/omim/>;
 eQTLs from GTEx, https://storage.googleapis.com/adult-gtex/bulk-qt/1v8/single-tissue-cis-qt/GTEx_Analysis_v8_eQTL.tar

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

We applied term “sex” to identify the biological differences between females and males. Both male and female were considered in our study. The sex was determined based on self-report and recheck by using genotype data from the X chromosome. For the 3,023 individuals in this study, 1,262 are males and 1,761 are females. When analyzing the Y-chromosome haplogroups, only 907 unrelated male samples were included. The disaggregate sex data are provided in the source data, and consent has been obtained for sharing individual-level sex data.

Reporting on race, ethnicity, or other socially relevant groupings

3,023 individuals were grouped as 30 ethnolinguistic populations, covering five major language families spoken in Southeast Asia: Austro-Asiatic, Austronesian, Hmong-Mien, Sino-Tibetan and Tai-Kadai. We also classified these 30 populations into indigenous and non-indigenous populations based on designation or classification of indigenous peoples of Southeast Asia from reported literatures.

Population characteristics

No covariate-relevant population characteristics were involved in this study.

Recruitment

The 30 populations across mainland Southeast Asia were approached and included in the research project. Participants were recruited from local communities with the assistance of local scientists. Research team members from Royal University of Phnom Penh in Cambodia (led by Dr. Tuot Sovannary) collected samples from Cambodia (14 populations: CMBR, CMJA, CMKA, CMKE, CMKH, CMKL, CMKR, CMKU, CMLA, CMLU, CMME, CMPH, CMST and CMTO); research team members from Chiang Mai University in Thailand (led by Dr. Jatupol Kampuansai) collected samples from Thailand (4 populations: THH1, THPA, THW1 and THW2); research team members from Ha Noi Medical University in Vietnam (led by Dr. Huy Ho Quang) collected samples from Vietnam (2 populations: VICH and VIKI); research team members from National University of Laos (led by Dr. Sommay Singthong) and Muang Sing District Hospital (led by Somphad Sochampa) collected samples from Laos (3 populations: LALL, LALS and LALT); research team members from Field School of Archaeology in Myanmar (led by Dr. U Win Kyaing) collected samples from Myanmar (4 populations: MYBU, MYCH, MYNA and MYRA); research team members from the Kunming Institute of Zoology, Chinese Academy of Sciences (led by Dr. Bing Su) collected samples from southwestern China (3 populations: CHBL, CHDA and CHWA). Samples were collected following the ethical guidelines by the local institutions, and all participants were given informed consent.

For each population, adult participants were randomly selected from local communities. Selection bias may be present, as participants more likely reside in specific rural areas, which may not fully capture the genetic diversity of the broader population. This could lead to the underrepresentation of certain subgroups and an incomplete picture of genetic variation across mainland Southeast Asia. These biases may limit the generalizability of the findings, potentially overlooking important genetic variations within underrepresented populations. Future research should prioritize targeted recruitment strategies to ensure a more comprehensive representation.

Ethics oversight

The study has been reviewed and approved by the Life Sciences Ethics Committee of Kunming Institute of Zoology, Chinese Academy of Sciences (approval ID: SMKX-20160311-45).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

Among the 30 studied populations, 23 populations have more than 20 individuals for reliable estimations of population parameters. The other 7 populations have less than 20 individuals due to practical reasons.

Data exclusions

17 individuals were excluded by stringent quality controls, including high genotype failure rate and apparent ancestry outlier from the cluster of Mainland Southeast Asia.

Replication

No replication experiments were conducted in this study.

Randomization

All samples were randomly collected from natural populations.

Blinding

No blinding since no treatment or experimental groups were involved.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.