nature portfolio

Corresponding author(s): Sijia Wang
Last updated by author(s): Jun 15, 2021

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 (*n*) 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 *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	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. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted *Give P values as exact values whenever suitable.*
☒	☐	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 d, Pearson's r), 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	Three-dimensional images composed of surface and texture maps were taken using the 3dMD Face (3dMD, Atlanta, GA). Genotyping was performed using the Illumina Infinium Global Screening Array or Illumina HumanOmniZhonghua-8.
Data analysis	KU Leuven provides the MeshMonk spatially dense facial-mapping software, free to use for academic purposes (https://github.com/TheWebMonks/meshmonk). The statistical analyses in this work were based on MATLABTM 2018a, SHAPEIT2 (v2.17), IMPUTE2 (v2.3.2), SNPLIB (https://github.com/jiarui-li/SNPLIB), MeshMonk, FUMA (v1.3.6), GREAT, Metascape, REHH2 (v3.2.0), plink 1.9, R (> v.3.6), ggplot2 (v3.1.0), and Python (v3.5.0) as mentioned throughout the Methods.

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

The participants making up the NSPT, NHC and TZL datasets were not collected with broad data sharing consent. Given the highly identifiable nature of both facial and genomic information and unresolved issues regarding risk to participants, we opted for a more conservative approach to participant recruitment. Broad data sharing of the raw data from these collections would thus be in legal and ethical violation of the informed consent obtained from the participants. This restriction is

not because of any personal or commercial interests. Additional details can be requested from Li Jin and Sijia Wang.
Publicly available data used were:
The 1000GP Phase 3 data: (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/),
The Roadmap Epigenomics Project: (https://egg2.wustl.edu/roadmap/web_portal/imputed.html#chr_imp)
The transcriptome resource from separated ectoderm and mesenchyme of the developing mouse face (GSE62214).
GWAS summary statistics are available on the National Omics Data Encyclopedia (NODE) (https://www.biosino.org/node/project/detail/OEP002283). The project ID of our study is OEP002283. Data usage shall be in full compliance with the Regulations on Management of Human Genetic Resources in China.

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	The samples in this study were collected from three independent cohorts, the National Survey of Physical Traits (NSPT) cohort (n = 3,322), the Northern Han Chinese (NHC) cohort (n = 4,767), and the Taizhou Longitudinal Study (TZL) cohort (n = 2,881). For the NSPT sample, individuals were recruited at three Chinese cities: Nanning, Guangxi province (n = 1,326); Taizhou, Jiangsu Province (n = 986); Zhengzhou, Henan province (n = 1,010).
Data exclusions	Relatedness among the individuals was inferred using KING-robust separately in the discovery and replication datasets1. Any relative up to the second-degree was excluded. As a result, 6968 (n = 4,089 in NHC cohort, n = 2,879 in NSPT cohort) and 2706 unrelated individuals in the discovery and replication dataset are used for further analysis.
Replication	The TZL cohort, where individuals were recruited in Taizhou, Jiangsu province, were used as the replication dataset.
Randomization	No randomization took place, group membership of identification and verification was determined by the separately obtained datasets available. Because canonical correlation analysis does not accommodate adjustments for covariates, we removed the effect of relevant covariates (sex, age, age-squared, height, weight, facial size, the first four genomic ancestry axes, and the camera system), on both the independent (SNP) and the dependent (facial shape pre segmentation) variables using partial least squares regression (plsregress from Matlab 2018a), and thus performed the canonical correlation analysis under a reduced model with residualized variables.
Blinding	Blinding was not relevant to this study.

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
☐	☒	Human research participants
☒	☐	Clinical data
☒	☐	Dual use research of concern

Methods

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	All participants are Han Chinese. For the NSPT cohort, the percentage of female participants is 63.81%, the average and standard deviation of age is 49.89 and 12.51, the average and standard deviation of BMI is 24.49 and 3.44. For the NHC cohort, the percentage of female participants is 51.16%, the average and standard deviation of age is 46.71 and 13.88, the average and standard deviation of BMI is 24.41 and 3.34. For the TZL cohort, the percentage of female participants is 64.45%, the average and standard deviation of age is 56.30 and 9.5, the average and standard deviation of BMI is 24.41 and 3.02.
Recruitment	The NSPT sample, individuals were recruited at three Chinese cities: Nanning, Guangxi province (n = 1,326); Taizhou, Jiangsu

Recruitment	Province (n = 986); Zhengzhou, Henan province (n = 1,010). The Northern Han Chinese sample, participants were recruited in Tangshan, Hebei province. The TZL cohort, where individuals were recruited in Taizhou, Jiangsu province, were used as the replication dataset. All individuals in the three cohorts were imaged using the 2-pod 3dMDface camera system. Our datasets and the characteristics did not contain any specific selection or bias that might influence this work.
Ethics oversight	All participants provided written informed consent, and all study protocols were approved by the institutional review boards of the pertinent research institutions. The Taizhou Longitudinal Study (TZL) was approved by the Ethics Committee of Human Genetic Resources at the Shanghai Institute of Life Sciences, Chinese Academy of Sciences (ER-SIBS-261410). The Northern Han Chinese cohort (NHC) was approved by the Ethics Committee of Human Genetic Resources at the Shanghai Institute of Life Sciences, Chinese Academy of Sciences (ER-SIBS-261410-A1801). The National Survey of Physical Traits (NSPT) is the sub project of The National Science & Technology Basic Research Project which was approved by the Ethics Committee of Human Genetic Resources of School of Life Sciences, Fudan University, Shanghai (14117).

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