nature portfolio

Corresponding author(s): John C Chambers

Last updated by author(s): Sep 2, 2025

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	Whole genome sequencing was carried using the Novaseq platform, with data processing using DRAGEN v3.7.8. Bisulfite conversion of genomic DNA was performed using the EZ DNA methylation kit (Zymo Research, Orange, CA), with DNA methylation quantified using the Illumina Infinium MethylationEPIC BeadChip® array (EPIC) (Illumina, Inc, CA, USA) according to manufacturer protocols. Bead intensity was retrieved using the minfi package in R, with a detection P<0.01 used for marker calling. RNAseq libraries were prepared using samples of whole blood (n=1,234) collected in PaxGene RNA tubes at enrolment. RNAseq libraries were prepared using NEBNext® Ultra™ II Directional RNA Library Prep (New England Biolabs, Inc.), with GLOBINClear (Thermo Fisher Scientific) for depletion of globin gene RNA and Ribosomal RNA (rRNA). The libraries were sequenced on a NovaSeq6000, using a paired-end run of 2 x 150bp. The Metabolon Global Discovery Panel was used for untargeted mass-spectrometry-based metabolic profiling of 10,000 fasting EDTA plasma samples.
Data analysis	OneMap APIs (https://onemap.gov.sg) were called within the R environment to generate the latitude and longitude of each participant's postal code and planning area. Variants were annotated using the VEP tool, extracting the most severe consequence for each variant. PRS was estimated using the score function in PLINK2. We perform functional overlap analysis and annotation of the sentinel CpGs using eFORGEv2.0.We performed GSEA on the 2,972 genes that were differentially expressed in any one pairwise ancestry comparison. GSEA was performed using the gprofiler2 tool to determine enriched gene sets across 3 pathway databases. These include Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome (REAC) and Wiki Pathways (WP). cis-Expression quantitative trait loci (eQTLs) for the 2,972 diiferentially expressed genes were analysed using Matrix eQTL (R package MatrixEQTL). The analytic codes are available in the following GitHub repository (https://github.com/HELIOS-SG100K-LKC/HELIOS-Study-design-and-baseline-characteristics/tree/64bb5b581a28509fdb325a0c633714b0cca691ef)

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 Individual level data for study participants are available under restricted access due to data privacy and ethical considerations. Access request can be submitted to the HELIOS Data Access Committee (email: helios_science@ntu.edu.sg). The GWAS summary statistics for the metabolites generated in the present study have been deposited in the GWAS catalogue under accession codes: GCST90557814-GCST90557966 (https://www.ebi.ac.uk/gwas/). The cis-eQTL summary statistics for the significant genes generated in the HELIOS study have been deposited in our GitHub repository and are publicly available. Source data are provided with this paper.

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	The term sex was used throughout the manuscript. The current manuscript is focused on comparisons between the three main Asian ethnic subgroups. We adjusted for sex in our regression models that explore differences in traits and disease between the ethnic groups, but did not carry out any sex stratified analysis. We have clarified where we adjust for sex as a covariate in the revised manuscript.
Reporting on race, ethnicity, or other socially relevant groupings	Ethnicity information was recorded based on participants' Singapore National Registration Identity Card. In our cohort, three major ethnicities were recruited: Chinese or other East Asian background, Indian or other South Asian background, and Malay or other South-East Asian heritage.
Population characteristics	The cohort includes 10,004 participants, with 4027 (40.3%) males and a mean age of 52.9 years. A total of 6,784 people are self-identified as Chinese or other East Asian background, 1,807 people as Indian or other South Asian background, and 1,354 people as Malay or other South-East Asian heritage. There were 59 participants from other ethnicities.
Recruitment	Study participants were recruited from the Singapore general population through a range of community outreach programs to ensure participation from ethnic minority groups, as well as people across socio-economic groups. Community engagement included language-specific recruitment drives conducted in the worship places, religious associations, and community associations across Singapore; multilingual study advertisement and documents (English, Chinese, Malay, and Tamil); and collaboration with a range of employers and occupational groups. Individuals were excluded if they were pregnant or breastfeeding, or had acute illness, major surgery within the previous 3 months, current participation in a drug trial, or cancer treatment in the past year. We present a comparison of the characteristics of HELIOS study participants to national statistics for the general population of Singapore. In keeping with our recruitment strategy which aimed to oversample the Indian and Malay ethnic groups, the proportion of people reporting Chinese ethnicity was lower than for the general population. There was some evidence for responder bias, with a higher proportion of women, lower cigarette smoking rates, and greater total years in education amongst study participants, compared to national averages. Nevertheless, the prevalence rates for diabetes, hypertension, hypercholesterolemia and obesity among HELIOS participants were similar to those reported in national statistics, indicating a minor healthy cohort effect overall.
Ethics oversight	HELIOS asks permission from participants to use the data and samples that they contribute, for clinical and molecular epidemiological research focussed on improving human health. This includes the application of 'untargeted' molecular profiling techniques that assess genomic, proteomic, transcriptional, metabolomic and other 'omic' variation in the biological samples collected. Participant consent also includes permission for linkage to disease registers, medical records, social care datasets and other health-related datasets held by Singapore's public bodies. Linkage is enabled by the Singapore NRIC, a unique national identifier allocated at birth, and with universal coverage. Consent provides permission for use of the data and samples from participants, by both academic and industry researchers, and for recontact of participants, including recontact based on phenotypic or genotypic characteristics. The HELIOS study operates under the governance framework of the Nanyang Technological University, and with Institutional Review Board approval (Ref: 2016-11-030).

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	The sample size is based on a number of factors including feasibility, cost, prevalence of risk factors, incidence of the major common non-communicable diseases, and range and depth of the extensive phenotypic data and measurements proposed.
Data exclusions	Individuals were excluded if they were pregnant or breastfeeding, or had acute illness, major surgery within the previous 3 months, current participation in a drug trial, or cancer treatment in the past year.
Replication	Assessment of reproducibility was carried out in 398 participants recalled between September 3, 2019 and January 28, 2022.
Randomization	This is not an experiment study so randomization is not required.
Blinding	This is not an experiment study so blinding is not required.

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, mosiacism, off-target gene editing) were examined.*