nature research

Corresponding author(s): Yukinori Okada

Reporting Summary

Nature Research 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 Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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
☐	☒	An indication of 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 statistics 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
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No software was used.
Data analysis	We used publicly available software for the data analysis (Omixon Target software version 1.9.3, BWA version 0.7.15, GATK version 3.6, OptiType version 1.3.1, POLYSOLVER version 4, HLA-HD version 1.2.0.1, Kourami version 0.9.6, eLD version 1, Rtsne R package version 0.13, DBSCAN R package version 1.1.1, Alluvial R package version 0.1-2, SNP2HLA version version 1.0.3, Eagle version 2.3, Minimac3 version 2.0.1, Rtsne R package version 3.2.3, GCTA version 1.91.1beta, Igraph R package version 1.1.2).

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/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

HLA data is deposited at the National Bioscience Database Center (NBDC) Human Database (Research ID: hum0114) as open data without any access restrictions. GWAS data of the BBJ subjects is available at the NBDC Human Database (Research ID: hum0014).

Field-specific reporting

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

☒ Life sciences ☐ Behavioural & social sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences

Study design

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

Sample size	Regarding the samples used for HLA sequencing, we used publicly available DNA obtained from cell lines of unrelated Japanese individuals established by the Japan Biological Informatics Consortium (JBIC, n = 1120). Regarding the samples used in the GWAS and PheWAS analysis, the BioBank Japan project has recruited roughly 200,000 participants. Clinical informations extracted from medical records and DNA / serum samples were also collected. Most of them were also genotyped by genome-wide SNP genotyping array, and relatively a small fraction of them were whole-genome-sequenced. For both, we selected samples as many as possible if they have available genotype or phenotype data.
Data exclusions	We excluded the samples and variants based on the standard quality control procedure in GWAS. Detailed information on quality controls were sufficiently described in our manuscript. The exclusion criteria were pre-established in the field of GWAS.
Replication	We used all the available data in the study, and did not conduct the two-staged discovery and replication studies.
Randomization	We did not apply randomization. All the samples with available accessibility to genotype and phenotype data were included in the analysis.
Blinding	We did not apply blinding of the samples because no intervention was conducted in our study.

Materials & experimental systems

Policy information about availability of materials

n/a	Involved in the study
☒	☐ Unique materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Research animals
☒	☐ Human research participants

Method-specific reporting

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ Magnetic resonance imaging