nature research

Corresponding author(s): Dr. Sekar Kathiresan
Dr. Pradeep Natarajan

Last updated by author(s): Sep 30, 2020

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 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	No software was used
Data analysis	CHIP Identification: putative somatic SNPs and short indels were called with GATK Mutect2 (https://software.broadinstitute.org/gatk). Single variant association analyses were performed with SAIGE version 0.29 (https://github.com/weizhouUMICH/SAIGE) Other statistical analysis was performed with R version 3.5 (https://www.r-project.org/).

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 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

Individual whole-genome sequence data for TOPMed whole genomes, individual-level harmonized phenotypes, harmonized germline variant call sets, the CHIP somatic variant call sets, RNA-Seq and peripheral blood methylation data used in this analysis are available through restricted access via the dbGaP. Accession numbers for these datasets are provided in Supplemental Table 1. Summary-level genotype data are available through the BRAVO browser (https://bravo.sph.umich.edu/). Full GWAS summary statistics are available at dbGaP accession phs001974: NHLBI TOPMed: Genomic Summary Results for the Trans-Omics for Precision Medicine Program.

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	As this was a genomic discovery effort, we sought to maximize sample size by aggregating a set of samples that was ~10 larger than all prior CHIP analysis efforts. In a post-hoc power calculation, we estimate that we had >80% power to detect variants at a minor allele frequency of >5% that confer at least a 1.15 fold genotype relative risk of CHIP. No statistical methods were used to predetermine sample size.
Data exclusions	Given that CHIP is unlikely to manifest in younger individuals, these individuals are effectively censored in our analysis set – that is, a young individual that does not presently have CHIP may still develop CHIP in the future. To avoid the power loss associated with misclassification of controls, we pruned these individuals from our analysis set. The single variant association analysis was run on a pruned set of samples that excluded those which had less than a 1% probability CHIP as estimated by the aforementioned model. This threshold was pre-established before performing the analysis. This excluded 21,712 samples leading to a final analysis set of 65,405 which was used for downstream association analyses.
Replication	We replicated the association with CHIP at the top loci (TERT) with prior analysis and replicated the TET2 locus using a second cohort of TOPMed samples distinct from our discovery analysis. We found support for all three single variant loci as well as the rare-variant CHEK2 loss of function burden signal in the cosubmitted paper on the closely related myeloproliferative neoplasm phenotype (Bao et al).
Randomization	Not applicable to genetic association studies.
Blinding	Not applicable to genetic association studies.

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

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	K562 cell lines were obtained from ATCC
Authentication	Identity validated using STR analysis
Mycoplasma contamination	Mycoplasma testing was routinely performed on all cells used in the study, and confirmed to test negative.
Commonly misidentified lines (See ICLAC register)	None.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Whole genome sequencing (WGS) was performed on 97,691 samples sequenced as part of 51 distinct studies contributing to

Population characteristics	the NHLBI TOPMed research program as previously described. (https://www.biorxiv.org/content/10.1101/563866v1; www.nhlbiwgs.org) Each of the constituent studies used in this analysis provided informed consent on the participating samples. Details on participating cohorts and samples is provided in Supplemental Table S1. Each of the studies contributing to TOPMed has a distinct study design and scientific focus. Study designs included community based prospective cohorts, case-control studies for heart lung, blood and sleep disease, including studies which focused on asthma, COPD, pulmonary fibrosis, hypertension, myocardial infarction, coronary artery disease, stroke, vascular disease, venous thromboembolism, congenital heart disease, atrial fibrillation, adiposity, blood traits, lipids, sleep traits. A subset of the studies contained extended family structures while most contained unrelated individuals. The sequenced individuals were highly diverse including ~40% of European ancestry individuals, ~30% of African ancestry individuals, ~ 15% Hispanic/Latino individuals and ~10% Asian ancestry individuals. Approximately equal proportions of male and female individuals were included. Sequenced individuals spanned the spectrum of ages from birth to >100 years old.
Recruitment	Recruitment of each of the 51 studies contributing to the data analyzed here has been previously described in detail (https://www.biorxiv.org/content/10.1101/563866v1; https://www.nhlbiwgs.org/parent-study-descriptions). Each of the studies contributing to TOPMed has a distinct study design. The most common study design were community based observational epidemiology studies. Recruitment for these most commonly included individuals from a given community who were recruited to participate at random (eg Framingham Heart Study) or through community schools/clinics/hospitals (eg Gene-Environment, Admixture and Latino Asthmatics study); (2) electronic health record/biobank based studies, where individuals volunteered for research studies and samples were later selected for sequencing (eg BioME); (3) disease cohort/registry based studies where individuals with a specific condition were selected (eg Boston Early-Onset COPD).
Ethics oversight	Written informed consent was obtained from all human participants by each of the studies that contributed to TOPMed with approval of study protocols by ethics committees at participating institutions. Secondary analysis of the TOPMed data as described in this manuscript was approved by the Partners Healthcare Institutional Review Board. All relevant ethics committees approved this study and this work is compliant with all relevant ethical regulations.

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