nature research

Corresponding author(s): Benjamin L. Ebert

Last updated by author(s): Aug 23, 2021

# 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<br>*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<br>*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 | Datasets were downloaded directly from the UK Biobank and the Mass General Brigham Biobank and no software was used for data collection. |
| Data analysis | Somatic variants in the study were identified by using Mutect2 (v4.1.6.0) using criteria explained in the Methods. The workflow to identify somatic variants from alignment bam files are available in WDL format in github (https://github.com/gatk-workflows/gatk4-somatic-snvs-indels). The mosaic chromosomal alterations were obtained from the UK Biobank (Return 3094) and were identified using the MoCha algorithm (https://github.com/freeseek/mocha). Following tools were also used: Samtools (v1.7). Custom code were used to analyze the data in R version 3.4.4 which is also available in GitHub (https://github.com/abhisheknrl/myeloid_lymphoid_CH). |

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

The source data are available to the approved researchers through the UK Biobank and Mass General Brigham Biobank. The data generated in this study, including the somatic variants and chromosomal alterations, are available as Supplementary Materials which will be submitted to the respective biobanks to enable linking with individual-level data and sharing with other approved researchers. Usage of these data will be covered by the data use agreements with the respective

biobanks and no additional restrictions apply. Individual-level MGBB data are available from https://personalizedmedicine.partners.org/Biobank/Default.aspx, but restrictions apply to the availability of these data, which were used under institutional review board (IRB) approval for the current study, and so are not publicly available. Individual-level UK Biobank data are available for approved researchers from https://www.ukbiobank.ac.uk. The present article includes all other data generated or analyzed during this study. Additional databases used in this study are: Genome Aggregation Database (gnomAD, https://gnomad.broadinstitute.org), cBioPortal for cancer genomics (https://www.cbioportal.org), the atlas of genetics and cytogenetics in oncology and haematology (http://atlasgeneticsoncology.org).

# 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 | We analyzed all available UK Biobank and Mass General Brigham Biobank samples with genotypes. |
| Data exclusions | We excluded samples with a history of hematologic malignancies, samples who had withdrawn consent, samples who had missing covariates, and one sample from each related pair. The data exclusion criteria were established prior to the analysis. |
| Replication | The Mass General Brigham Biobank cohort was used as a replication cohort. No experimental replication. |
| Randomization | n/a; We analyzed all samples together. |
| Blinding | n/a; The data was previously collected by the respective biobanks and all samples with genotype data were analyzed together. The genotype analysis were independent of the phenotype data. |

# 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 |