nature research

Corresponding author(s): Liran I. Shlush

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

Sample size was determined by the availability of the samples from EPIC (European Prospective Investigation into Cancer and Nutrition) while assuring large enough cohort for the reported statistical analyses. (124 pre-AML samples and 676 controls). As detailed in Methods, we used weighting to minimise the biases introduced by the artificial case-control ratio and calculated hazard ratios relative to the (approximate) true cumulative AML incidence of about 1-3/1,000 in the given age range over a follow up of 10-20 years.

2. Data exclusions

Describe any data exclusions.

For the model described in figure 3 we have excluded samples taken less than 6 months prior AML diagnosis from model training and validation in order to avoid skewing the model towards significance (as per reviewer comments). This resulted in the removal of 4 individuals from the discovery cohort, leaving 91 individuals in the discovery cohort pre-AML group.

3. Replication

Describe whether the experimental findings were reliably reproduced.

Experimental replication was not attempted but rather validated in a second independent cohort

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

AML patients were identified based on the following ICD9 codes: 9861/3 9860/3 9801/3 9866/3 9891/3 9867/3 9874/3 9840/3 9872/3 9895/3 9873/3, which included only cases of de novo AML, and no secondary AML. Age and gender matched individuals that were not diagnosed for any hematological disorders during the average follow-up period of 12.6 were selected as the control group,

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

For assessment of statistical significant differences between the cases as controls blinding is impractical as the groups must be defined. Machine learning algorithms were blind in a sense as they been configure to be trained and tested in different set of samples.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

1) Sequencing reads were aligned using BWA-aln and BWA-mem
2) indel-realignment was done using GATK
2) Variant calling was done using deepSNV, Varscan2, Pindel v2.2 and CaVEMan v1.11.2. All of these algorithms are extensively described and validated in the published literature.
3) Code used to generate predictive models will be deposited on GitHub at the time of manuscript publication
For full description please refer to the appropriate method sections

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

No unique material were used in this study

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

No antibodies were used in this study

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

No cell lines were used in this study

b. Describe the method of cell line authentication used.

No cell lines were used in this study

c. Report whether the cell lines were tested for mycoplasma contamination.

No cell lines were used in this study

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

No cell lines were used in this study

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	No animals were used in this study

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	For the discovery cohort, samples were collected from individuals upon enrollment into the EPIC study between 1993 and 1998. 95 individuals (39 males and 56 females) who developed AML an average of 6.26 years (IQR=4.88 years) after the sample was collected were included in the pre-AML group. 414 age and gender (167 males and 247 females) matched individuals were selected for the control group, as they did not develop any hematological disorders during the average follow-up period of 11.6 years (IQR=2.13 years). The median age at recruitment was 56.75 years (range, 36.08 to 74.42) For the validation cohort, samples were ascertained from participants in the EPIC-Norfolk longitudinal cohort study enrolled between 1994 and 2010. 37 individuals (15 males and 22 females) who developed AML an average of 10.5 years from first sampling (IQR=8.3 years) were included in the pre-AML group. 262 age and gender (135 males and 127 females) matched individuals were selected for the control group, as they did not develop any hematological disorders during the average follow-up period of 17.5 years from first sampling (IQR=3.8 years). The median age at recruitment was 65.05 years (range, 43.9 to 88.1)