nature research

Corresponding author(s): Gad Getz
Catherine J. Wu

Last updated by author(s): Jun 1, 2022

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 for data collection.
Data analysis	A code availability statement is available in the main text, which includes reference to: Terra methods used in the study can be found at https://app.terra.bio/#workspaces/broad-firecloud-wupo1/CLLmap_Methods_Apr2021. Source code used in the study can be found at https://github.com/getzlab/CLLmap. The RFcaller pipeline is available at https://github.com/xa-lab/RFcaller. The new epiCMIT suitable for Illumina arrays and NGS approaches as well as the CLL epitype classifier can be found at https://github.com/Duran-FerrerM/CLLmap-epigenetics. Detailed descriptions of our analytical pipelines are provided in the Methods section and/or Supplementary Information. Software versions used: BWA-MEM v0.7.15-r1140, CrosscheckFingerprints from GATK v4.0.5.1, Oxog v17, MuTect v1.1.6, MuTect2 v4.0.5.1, Strelka2 v1.0.11, DeTiN v1.8.9, MutSig2CV v3.1, GISTIC2 v2.0.23, ABSOLUTE v1.5, PhylogicNDT v1 (2020-01-08), SignatureAnalyzer v1.2.3, CLUMPS (https://github.com/getzlab/getzlab-CLUMPS2; 2020-01-27), REBC_tools v1.1.3, g:profiler v2_0.1.9, igraph v1.2.5, IgCaller v1.1, IMGT/V-QUEST v3.5.18, MiXCR v.3.0.10, STAR v2.6.1d, RNA-SeQC v2.3.6, CIBERSORT v1.0.5, limma v3.42.2, GSVA v1.34.0, fgsea v1.15.1, sklearn v0.22.2 (for expression cluster classifier) and v0.21.3 (for KMeans), imblearn v0.6.2, pROC v1.17.0.1, PRROC v1.3.1, BSMAP v2.90, MOABS v1.3.9.6, ConsensusClusterPlus v1.52.0, CHMM v1.17, DESeq2 v1.34.0.

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

Sequencing, expression, and genotyping is available at European Genome-Phenome Archive (EGA, http://www.ebi.ac.uk/ega/), which is hosted at the European Bioinformatics Institute (EBI), under accession numbers EGAS00000000092 and in dbGaP under accession numbers: phs001473.v2.p1, phs000922.v2.p1, phs001431.v2.p1, phs001091.v1.p1, phs000435.v3.p1, phs002297.v2.p1, phs000879.v1.p1 and GEO accession number GSE143673. 450k array data is available at EGA under accession number EGAD00010001975. Patient data were collected from medical records by the contributing institutes. All other data that support the findings of this study are publicly available. Project data portal: https://cllmap.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	1.) Composition and description of the cohort: Table 1, Supplementary Tables 1,2; Extended Data Fig. 1a-b and Online Methods. 2.) For candidate cancer driver gene detection, our prior power estimates (Lawrence et al., Nature, 2014) suggested that with the ~1000 WES samples used in this study and somatic background mutation rate of ~1/Mb in CLL, we should be able to discover >90% of drivers mutated in 2% of patients.
Data exclusions	No data were excluded other than a limited number of samples excluded prior to analysis due to low quality or data redundancy.
Replication	The genetic analysis incorporates data from previous studies and shows replication with previous findings (e.g., Landau et al., Nature, 2015 and Puente et al., Nature, 2015). RNA expression clusters show comparable cluster membership frequencies in independent cohorts (Extended Data Fig. 8h). RNA expression clustering after downsampling the samples shows consistency with clustering using the full cohort (Supplementary figure 3) Epigenetic analysis shows replication of epigenetic feature inference across different platforms (Extended Data Fig. 7f) and consistency with previous epigenetic analyses (Kulis, Nature Genetics, 2012) IGHV characterization shows high agreement by different technologies (Sanger sequencing and computational inference from high-throughput sequencing data).
Randomization	Training and test sets for random forest classifiers were chosen randomly.
Blinding	Computational IGHV mutational status inference, expression cluster detection and epitype assignment were performed independently and blinded to clinical endpoints. Blinding was not relevant to other genomic analyses unrelated to patient outcome.

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	The characteristics of the 1148 CLL/MBL samples are described in Supplementary Table 1 and clinical characteristics of the 1009 CLL samples used in the clinical analysis are listed in Table 1. These included tumor and germline samples collected either during active surveillance (n=680), post-treatment (n=52), or at enrollment of a clinical trial prior to first cycle of therapy (n=416; treatment-naive n=371, relapsed/refractory n=45). Briefly, these trials included: (i) comparison of fludarabine and cyclophosphamide (FC) to FC-rituximab (FCR) in previously untreated patients (CLL8 trial, n=309); (ii) treatment-naive TP53 mutated patients within phase 2 CLL20 trial who all received alemtuzumab (n=31); (iii) ibrutinib or R-ibrutinib in relapsed/refractory (R/R) or untreated patients with del(17p), TP53 mutation, and/or del(11q) (n=76; treatment-naive n=31; R/R n=45).
Recruitment	CLL/MBL patients were recruited to studies at individual centers per local protocols regardless of race, gender, ethnicity, or other characteristics. Those enrolled on clinical trials were required to meet inclusion and exclusion criteria as defined in each respective study (note, patients were excluded if they declined to provide a research sample). Written informed consent was obtained from all patients.
Ethics oversight	Samples were collected via protocols approved by institutional review boards or ethics and policy committees from the International Cancer Genome Consortium, German CLL Study Group, Dana-Farber Cancer Institute, the CLL Research Consortium, National Heart, Blood and Lung Institute, and MD Anderson Cancer Center. All clinical trials were conducted in accordance with the Declaration of Helsinki and International Conference on Harmonization Guidelines for Good Clinical Practice.

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