nature portfolio

Corresponding author(s): Charles Mullighan

Last updated by author(s): Jun 22, 2022

Reporting Summary

Nature Portfolio 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 Portfolio 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	Standard Illumina instrumentation software (including software used by HiSeq2000, HiSeq2500, and NovaSeq instruments) and Complete Genomics, Inc. (CGI) software were used to collect the DNA and RNA sequencing data reported in this study.
Data analysis	BWA version 0.5.9, CONSERTING version 1.0, CREST version 1.0, SvABA version 1.1.3, Manta version 1.6.0, Delly version 0.8.2, Bambino version 1.6, VEP version 95.2, MutSigCV version 1.4, GISTIC version 2.0, STAR version 2.4.2a, CGI Cancer Sequencing service pipeline version 2, SigProfilerSingleSample version 1.3, GraphPad Prism version 8.4, Python version 2, SAS version 9.4, R version 3, the Tapestri pipeline version 2, MedalCeremony version 1.0, CICERO (in-house, prepublication version), FusionCatcher version 1.0, HTSeq version 0.6.0, GRIN (not versioned), ANNOVAR version 2014-11-12.

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 Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Genomic data is publicly available and data accessions for RNA-seq, WES, WGS and SNP are listed for each case in Supplementary Table 1. TARGET ALL data may be accessed through the TARGET website at https://ocg.cancer.gov/programs/target. The sequencing BAM and FASTQ files are accessible through the database of

genotypes and phenotypes (dbGaP; http://www.ncbi.nlm.nih.gov/gap) under accession number phs000218 (TARGET) and at NCI's Genomic Data Comments http://gdc.cancer.gov under project TARGET. The remaining data has been deposited in the European Genome Phenome Archive, accessions EGAS00001000447, EGAS00001000654, EGAS00001001923, EGAS00001001952, EGAS00001002217, EGAS00001003266, EGAS00001004810, EGAS00001004998, EGAS00001005084 and EGAS00001005250 and is also accessible through St. Jude Cloud at https://platform.stjude.cloud/data/cohorts?dataset_accession=SJC-DS-1009. Somatic mutation data can also be explored interactively using ProteinPaint and GenomePaint on St. Jude Cloud at https://viz.stjude.cloud/mullighan-lab/collection/the-genomic-landscape-of-pediatric-acute-lymphoblastic-leukemia~15.

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	No sample size or power calculations were performed beforehand. The cohort size was selected based on the availability of existing genomic sequencing data for acute lymphoblastic leukemia (ALL) samples, or the presence of ALL tissue on which genomic sequencing could be performed. The goal was to create the largest possible ALL genomic dataset, resulting in 2,754 samples analyzed in the study.
Data exclusions	Samples with oversegmented copy number profiles were excluded from the study. Samples found to be duplicate samples from the same patient, based on genomic similarity, were excluded. These exclusion criteria were not pre-determined.
Replication	We did not verify reproducibility as our goal was to aggregate previously published and newly sequenced ALL genomic data, in order to create a large ALL genomic dataset.
Randomization	Patients were not allocated into experimental groups as this is a retrospective, descriptive study of the genomic basis of ALL.
Blinding	Patients were not allocated into experimental groups, so no blinding was performed. This was a retrospective, descriptive study of the genomic landscape of ALL.

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

Antibodies

Antibodies used

For flow cytometry to increase tumor purity, the following antibodies were used:
BD Biosciences: Human CD45-FITC (clone 2DI, cat# 340664, 1:20); Human CD19-PE (clone 4G7, cat #349209, 1:20).
Life Technologies: Human CD7-APC (clone CD7-6B7, cat #MHCD0705, 1:40).

Links:
Human CD45-FITC: https://www.bdbiosciences.com/en-us/products/reagents/flow-cytometry-reagents/clinical-diagnostics/singlecolor-antibodies-asr-ivd-ce-ivd/cd45-fitc.340664
Human CD19-PE: https://www.bdbiosciences.com/en-us/products/reagents/flow-cytometry-reagents/clinical-diagnostics/singlecolor-antibodies-asr-ivd-ce-ivd/cd19-leu-12-pe.349209
Human CD7-APC: https://www.thermofisher.com/antibody/product/CD7-Antibody-clone-CD7-6B7-Monoclonal/MHCD0705

For Tapestri analysis: Ten TotalSeq™ oligo-conjugated antibodies from Biolegend (concentration 0.5 mg/ml) were used for cell-surface protein analysis: D0048 anti-human CD45 (clone 2D1), D0054 anti-human CD34 (clone 581), D0066 anti-human CD7 (clone

CD7-6B7), D0063 anti-human CD45RA (clone HI100), D0062 anti-human CD10 (clone HI10a), D0050 anti-human CD19 (clone HIB19), D0389 anti-human CD38 (clone HIT2), D0052 anti-human CD33 (clone P67.6), D0034 anti-human CD3 (clone UCHT1) and D0138 anti-human CD5 (clone UCHT2). For each sample we combined 1 ul of each antibody and brought to a volume of 50 ul by adding Cell Staining Buffer according to guidelines from the TotalSeq™-D with Mission Bio Tapestri® Single-Cell DNA + Protein Protocol (https://www.biolegend.com/en-us/protocols/totalseq-d-with-mission-bio-tapestri-single-cell-dna-protein-protocol).

Links:
TotalSeq™-D0048 anti-human CD45 Antibody, https://www.biolegend.com/en-us/products/totalseq-d0048-anti-human-cd45-antibody-21096
TotalSeq™-D0054 anti-human CD34 Antibody, https://www.biolegend.com/en-us/products/totalseq-d0054-anti-human-cd34-antibody-20805
TotalSeq™-D0066 anti-human CD7 Antibody, https://www.biolegend.com/en-us/products/totalseq-d0066-anti-human-cd7-antibody-20808
TotalSeq™-D0063 anti-human CD45RA Antibody, https://www.biolegend.com/en-us/products/totalseq-d0063-anti-human-cd45ra-antibody-20992
TotalSeq™-D0062 anti-human CD10 Antibody, https://www.biolegend.com/en-us/products/totalseq-d0062-anti-human-cd10-antibody-20862
TotalSeq™-D0050 anti-human CD19 Antibody, https://www.biolegend.com/en-us/products/totalseq-d0050-anti-human-cd19-antibody-20856
TotalSeq™-D0389 anti-human CD38 Antibody, https://www.biolegend.com/en-us/products/totalseq-d0389-anti-human-cd38-antibody-20851
TotalSeq™-D0052 anti-human CD33 Antibody, https://www.biolegend.com/en-us/products/totalseq-d0052-anti-human-cd33-antibody-20881
TotalSeq™-D0034 anti-human CD3 Antibody, https://www.biolegend.com/en-us/products/totalseq-d0034-anti-human-cd3-antibody-20845
TotalSeq™-D0138 anti-human CD5 Antibody, https://www.biolegend.com/en-us/products/totalseq-d0138-anti-human-cd5-antibody-20834

Validation

All antibodies for flow cytometry were validated for detecting human proteins by the manufacturer and confirmed for each specific application using cells of known origin and differentiation state and compared to isotype controls and cells that are known to express or lack the antigen. Flourescence-labeled antibodies used for flow cytometry analysis were validated by the SJCRH Flow Core facility.

Validation information can be found on the company websites with the links listed above. For example, the Human CD45-FITC website has the following statement: "Applications Tested: This 2D1 antibody has been pre-titrated and tested by flow cytometric analysis of human peripheral blood leukocytes. This can be used at 5 µL (1.0 µg) per test. A test is defined as the amount (µg) of antibody that will stain a cell sample in a final volume of 100 µL. Cell number should be determined empirically but can range from 10^5 to 10^8 cells/test." Please see https://www.thermofisher.com/antibody/product/CD45-Antibody-clone-2D1-Monoclonal/11-9459-42.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Patients had received a diagnosis of ALL and were all under the age of 31 years at diagnosis (median 6.4 years). Of 2,749 patients with known gender, 1,148 were female.
Recruitment	Participants were recruited (and consent obtained) to participate in research on their malignant tissue during clinical care at St. Jude Children's Research Hospital or Children's Oncology Group member institutions.
Ethics oversight	St. Jude Children's Research Hospital IRB and the Children's Oncology Group provide ethics oversight for the study.

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

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	For patient samples containing less than 70% blasts, cryopreserved bone marrow or peripheral blood samples were thawed in IMDM media containing 20% FCS and flow sorted to enrich for tumor cells before sequencing.
Instrument	Cells were collected using a FACSAria III instrument (BD Biosciences).

Software	FACSDiva 9.0 software (BD Biosciences) was used for data acquisition and analysis.
Cell population abundance	Tumor purity was assessed to be >90% post-sort.
Gating strategy	Tumor cells were gated on live cells based upon FSC-A and SSC-A, then gated on viable cells based upon FSC-A vs DAPI, then gated on singlets based upon SSC-W and FSC-A, then gated on CD45-dim based upon CD45-FITC and SSC-A, then gated on CD19 expression based upon CD45-FITC and CD19-PE (P1T). Non-tumor cells were gated on live cells based upon FSC-A and SSC-A, then gated on viable cells based upon FSC-A vs DAPI, then gated on singlets based upon SSC-W and FSC-A, then gated on CD45-high based upon CD45-FITC and SSC-A, then gated on CD7 expression based upon CD45-FITC and CD7-APC (termed P2N).

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.