nature research

Corresponding author(s): Bert Vogelstein, Kenneth W. Kinzler, Nickolas Papadopoulos

Last updated by author(s): Mar 14, 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 Authors & Referees 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 data collection software was used in this study.
Data analysis	Unmapped BAM files were generated and adapter sequences were marked using Picard (v2.17.11) IlluminaBasecallsToSam. Reads were mapped to the human reference genome (hg19) using BWA-MEM (v0.7.17). Following alignment, reads were merged and sorted using SAMtools (v1.7). A custom Python (v2.7.15) pipeline was developed for the analysis of SaferSeqS data. The source code is available in a Zenodo repository (https://doi.org/10.5281/zenodo.4588264). Python packages pysam (v0.15.2), numpy (v1.14.0), tqdm (v4.19.4), umi_tools (v0.5.4) as well as R (v.4.0.3) packages ggplot2 (v3.3.2), ggsci (v2.9), ggpubr (v0.4.0), ggforce (v0.3.2), ggbeeswarm (v0.6.0), scales (v1.1.1), stats (base), binom (v.1.1-1), splines (base), data.table (v1.13.4), dplyr (v1.0.2), forcats (v0.5.0), tidyr (v1.1.2), and readxl (v1.3.1) were used for data analysis. Conda (v4.4.10) was used to setup and manage Python dependencies. RStudio (v1.1.463) was used to setup and manage R packages.

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/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 sequencing data generated in this study can be obtained from the European Genome-phenome Archive (accession number EGAS00001005048).

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 calculation was necessary for the design of this study. Several million bases were queried in aggregate across all samples analyzed, permitting reliable estimates of mutation frequencies.
Data exclusions	Reads containing bitwise flagwise values of 99 and 147 are derived from the Watson strand and those containing bitwise flags of 83 and 163 are derived from the Crick strand. Reads with any other bitwise flag values indicate that they either failed map or improperly mapped to the reference genome and were excluded from subsequent analysis. Two additional quality control criteria were imposed during UID family grouping to ensure accurate determination of the endogenous molecular barcode (i.e. fragment end coordinate): 1) reads with soft clipping at the 5' or 3' of the fragment ends were excluded, 2) reads were required to contain the expected constant tag sequence (GCCGTCGTTTTAT) immediately following the exogenous UID with no more than one mismatch. Furthermore, UID families resulting from "barcode collisions" (i.e. those sharing the same exogenous barcode but different endogenous barcodes) were excluded (see Supplementary Note). To exclude common polymorphisms, we excluded known germline mutations and all mutations in the Genome Aggregation Database (gnomeAD) present at an allele frequency greater than 0.01%. Reads comprising supercalimutants were subjected to a final manual inspection to exclude possible alignment artifacts.
Replication	Admixture experiments were performed with three different TP53 mutations and at three dilutions. Five plasma samples from cancer patients were assayed for eight different mutations that were originally identified in the corresponding primary tumors. All experimental conditions were assayed with a previously described molecular barcoding method ("SafeSeqS") and with SaferSeqS. Seventy four plasma samples from cancer patients were assayed for 86 primary tumor-derived mutations with multiplex and single amplicon SaferSeqS assays. Specificity of SaferSeqS was also evaluated in a collection of 24 plasma samples obtained from healthy donors.
Randomization	For experiments comparing SaferSeqS to previously described molecular barcoding methods, samples were randomly partitioned into two aliquots and assayed with both techniques
Blinding	Researchers were blinded to the identity and composition of the samples during library preparation.

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
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

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	Healthy donors were of average age 30 and had no known diagnosis or previous history of cancer. Cancer patients had biopsy-confirmed malignancies that were surgically resectable.
Recruitment	Plasma samples from healthy donors were acquired from BioIVT. Cancer patients were recruited as part of an ongoing clinical trial evaluating the use of liquid biopsies in the detection and management of residual disease.
Ethics oversight	The study was approved by the Institutional Review Boards for Human Research at participating institutions in compliance with the Health Insurance Portability and Accountability Act. Informed consent was obtained from all patients.

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