nature portfolio

Corresponding author(s): Alexander Bagaev

Last updated by author(s): 03/21/2025

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	No specific software was used to collect the data.
Data analysis	This manuscript contains code for the reference samples processing and performance characteristics assessment. A newly developed software code used in the study was deposited at: https://github.com/BostonGene/Somatic_reference_standards/tree/main/performance_characteristics_test. BWA aligner v.0.7.17 was used to map WES data to the human genome (hg38). GATK v4.1.2 and mosdepth v0.2.1 were used for PCR duplicates read markup and sequencing metrics collection, including average coverage statistics. RNA-seq data were mapped to the human genome (hg38) using the default parameters of STAR aligner v2.4.2 with minor modifications. For gene expression quantification, reads were aligned to the human transcriptome (hg38) with Kallisto v.0.43.0 using default parameters. Standard quality control for WES was performed via fastQC v0.11.9 and FastqScreen v0.14.0. Picard v2.20.7 MarkDuplicates was used to remove duplicate reads. Off-target reads were calculated using samtools view v1.10 in intersection with target file with regions provided by the vendor (Agilent Technologies, Santa Clara, CA, USA). Unique reads were calculated using samtools view v1.10 with -F0x400 flag. Standard QC for RNA-seq was performed via RSeQC v3.0.1. OptiType v1.3.5 was used to obtain HLA types for quality control of sample mixing. Somatic INDEL (1-49 bp) calling was performed via Strelka v2.9.10 using small INDEL candidates from manta v1.5.0. Variant calling from RNA-seq data was performed via pisces v5.2.10.49. For additional verification of SNV/INDEL calls, presence samtools mpileup v1.10 and a downstream parsing custom script were used to precisely check discordant calls between two compared samples. MSISensor2 (https://github.com/niu-lab/msisensor2) was used to calculate the microsatellite instability (MSI) score. Macs3 v3.0.0a7, Samtools v1.10 (https://github.com/samtools/), bedtools v2.30.0 multiinter, and modified Sequenza v2.1.2 were used for CNV detection. For gene expression analysis, Kallisto v0.42.4, GENCODE v21, and STAR aligner v2.4.2 were used.

For fusion transcripts detection STAR-fusion v1.8.1 was used with subsequent annotation of FusionInspector (https://github.com/FusionInspector/FusionInspector/wiki) in validation mode.
Immune clonotypes were determined with MiXCR v3.0.12.

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

The data generated and analyzed during the current study is available at SRA repository PRJNA1134786 (https://dataview.ncbi.nlm.nih.gov/object/PRJNA1134786?reviewer=tf137c52lrit8k31q4j79om3m1) and will be available online at https://github.com/BostonGene at the time of publication.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	This information has not been collected.
Reporting on race, ethnicity, or other socially relevant groupings	This information has not been collected.
Population characteristics	This information has not been collected.
Recruitment	Not applicable
Ethics oversight	For the internal BostonGene samples, informed consent was obtained. The use of clinical samples was conducted in accordance with the Declaration of Helsinki and has been granted exemption from ethics approval by the Biomedical Research Alliance of New York (BRANY) Institutional Review Board (IRB) (BRANY study #22-12-938-853).

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

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 prior sample size calculation was performed.
Data exclusions	Highly polymorphic genes were excluded from the analysis to minimize false positive SNVs and INDELs (Supplementary Tables 1-3). Statistical analysis for test samples (i.e., cell line dilutions, clinical samples) were calculated for arm-level and gene-level CNVs. An arm-level CNV was defined as a segment that exceeds 50% of the actual arm length. Thus, the total copy number (tCN) of the arm was defined by the presence of arm-level segments. Notably, if the arm was fractured by a high amount of CNVs (e.g., chromothripsis), then it was excluded from the statistical analysis of that sample. A gene-level CNV was defined as a gene that intersects with its corresponding segment by at least 10% of its length. If a gene had an intersection with multiple segments, then the tCN of the longest of those segments was assigned for that gene. This approach allowed us to resolve discrepancies between annotation of segments, which allowed for two CNV profiles to be objectively compared. During quality control bioinformatics procedures for WES, MarkDuplicates was used to remove duplicate reads. During gene expression analysis, the noncoding RNA, mitochondria- and histone-related transcripts were removed from analysis and the protein-coding transcripts, IGH/K/L- and TCR-related transcripts, were retained, resulting in 20,062 protein coding genes. During fusion detection, detected fusion transcripts that met any of the following filtering criteria were removed: 1) Fusion fragments per million (FFPM) values greater than or equal to 0.1; 2) Number of junction reads supporting the fusion greater than or equal to 5; 3) A sum of all supporting reads (junction and spanning) greater than or equal to 7; and 4) genes overlapping each other or with a breakpoint distance greater than 10 kb. Functional annotation was also taken into account (mitochondrial genes and fusions from FusionCatcher black list were filtered out).

Replication	Biological replicates of the COLO829 cell line were used as a positive control for each clinical run of sequencing. The samples were used for longitudinal reproducibility assessment of both RNA-seq and WES defined events. To establish the sufficient number of reads for proper transcripts detection and expression level assessment, 10 technical replicates of the same extract from the COLO829 cell line were used. For assessment of SNV and INDEL variant calling performance, true positive mutation sets for this validation were defined by performing four to six NGS replicates on COLO829, HCC1143, HCC1937, HCC1395, and NCI-H1770 cell lines with 100% tumor purity. For assessment of CNV detection performance, we evaluated the consistency of our CNV calling algorithm by assessing the quality metrics for cell line dilutions. Due to the absence of gold standard references with widely acknowledged lists of CNVs, we created references for each cell line by using different and independent runs of samples with 100% tumor purity: 6 replicates for COLO829, HCC1143 and HCC1937; 4 replicates for HCC1395 and NCI-H1770. For cell line references, normalized regions were compared between 4 (NCI-0H1770, HCC1395) to 6 (COLO829, HCC1143, HCC1937) sequencing replicates. During gene expression correlation analysis, the reproducibility (CV for gene expression and gene signatures) assessment was performed on 10 FFPE clinical samples. All of these 10 samples were made into 3 technical replicates and genes' CV was assessed within each samples' replicates separately. To assess reproducibility of TME subtypes classification, seven clinical samples from patients with solid diagnosis were used. Each sample was prepared in three replicates within one day and on two more days additionally (five samples in total). Each set of samples was prepared with 15, 20, and 50 ng of input starting material, resulting in a total of 15 replicates for each clinical sample.
Randomization	We performed randomization of cell line specimens during establishment of thresholds for variant calling for SNVs, INDELs, CNVs, and fusions, prior to assessing the performance for variant calling.
Blinding	Blinding was not used during allocation of samples.

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
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	Cell lines were purchased from the American Type Culture Collection (ATCC, Manassas, VA, USA) and cultured according to the vendor's instructions.
Authentication	None of the cell lines used were authenticated in our laboratory.
Mycoplasma contamination	Cell lines tested negative for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used in this study.

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A