nature research

Corresponding author(s): Malachi Griffith, Elaine R Mardis

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.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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
☒	☐	An indication of 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 statistics 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
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	TCGA data was downloaded via the Genomics Data Commons (GDC). dbGAP study accession IDs have been included for all datasets under the 'Data availability section.
Data analysis	All software has been cited throughout the manuscript where used. Genome Modeling System(GMS) was used for alignment and variant calling of in-house datasets. Specific softwares used under the GMS were BWA, Picard, Samtools, Sniper, Strelka, and VarScan. Germline variant calling and phasing was done using Samtools for in-house samples and GATK for TCGA samples. HLA-typing was performed via OptiType. Variant annotation was done using Variant Effect Predictor using custom plugins. These plugins are available for download as part of the pVACtools suite. pVACtools was used to assess for the presence of proximal variants, and for predicting corrected neoantigen binding scores via NetMHCv 4.0. Github repository for proximal variant analysis code: https://github.com/griffithlab/pVACtools/tree/proximal_variants

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 upon request. 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

Several of the in-house sequencing datasets used in the study have been previously published and deposited in various databases. All sequence data for the HER2+ breast cancer samples can be accessed via the Database of Genotypes and Phenotypes (dbGAP; study accession: phs001291). Data for oral squamous cell carcinoma project and hepatocellular carcinoma samples are part of other manuscripts currently in preparation, and can be accessed under dbGAP study accession phs001623 and phs001106, respectively. Results for the glioblastoma case and small cell lung cancer have been published and can be accessed under dbGAP study accessions phs001663 and phs001049, respectively. TCGA data can be accessed under dbGaP study accession phs000178.

Field-specific reporting

Please select 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/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Samples were chosen such that they represented low, medium and high mutational burden tumors. For TCGA, we chose a sample size that could be stored and further analyzed on our local compute clusters. No statistical methods were used to predetermine sample size as power analysis was not relevant to this study. The study is primarily observational in nature, no statistical tests were performed and no p-values are reported.
Data exclusions	No data was excluded from the analysis.
Replication	For reproducibility of our results, details of relevant software and code have added in the code availability section. Conventional biological replicates are not relevant to this study.
Randomization	N/A. We did not group the samples into different cohorts, hence randomization was not applicable to this study.
Blinding	N/A. We did not compare cases and controls to study difference in effects, hence blinding was not applicable to this study.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

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