

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 (<i>n</i>) 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 <i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>
☐	☒ 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. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted <i>Give P values as exact values whenever suitable.</i>
☒	☐ 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 <i>d</i> , Pearson's <i>r</i>), 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	Data was obtained from pre-existing datasets. GDC Data Transfer Tool gdc-client v2.3 (https://gdc.cancer.gov/access-data/gdc-data-transfer-tool) was used for downloading TCGA data files from the GDC Portal.
Data analysis	Annotation of somatic mutations was performed using Annovar (latest version as of Dec 5, 2024). The predicted pathogenicity of mutations was assessed using scores from the CADD database (CADD v1.7 for GRCh37/hg19). The Pangolin software (v1.0.1) was used to predict splicing changes caused by each somatic mutation, and FABIAN-Variant (web service) was used to predict variant effects on transcription factor binding affinity. Borzoi (https://github.com/calico/borzoi) was used for generating predicted RNA-seq tracks for mutated genes. Custom code was written in Python (v3.9.19), bash, and R (v4.5.0). Python libraries used include pyfaidx (v0.7.2.2), and pysam (v0.22.1) for manipulating data files, numpy (v), pandas (v1.5.3), and pandarallel (v1.6.5) for data analysis, scipy (v1.9.3) for performing statistical tests, statsmodels (v0.13.5) for multiple testing correction, seaborn (v0.12.2, and matplotlib (v3.7.5) for visualization, statannotations (v0.7.1), and adjustText (v1.3.0) for annotating plots, papermill (v2.6.0) for running parameterized Jupyter notebooks, and lifelines (v0.30.0) for survival analysis (Kaplan-Meier curves and Cox proportional Hazard models). The clinfun (v1.1.5) package in R was also used. All code used in this study is publicly available at https://github.com/Georgakopoulos-Soares-lab/cancer_passengers and archived at Zenodo [https://doi.org/10.5281/zenodo.19432145].

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](#)

All open-access data analyzed during this study are available at Zenodo [<https://doi.org/10.5281/zenodo.19432064>] (datasets.zip file). The ICGC portion of the PCAWG dataset is publicly available via ICGC-ARGO through the AWS bucket icgc25k-open (<https://object.genomeinformatics.org>) under the PCAWG directory; access instructions are provided at <https://docs.icgc-argo.org/docs/data-access/icgc-25k-data>. TCGA data are available under controlled access and were obtained through dbGaP under accession phs000178.v11.p8 [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000178.v11.p8]. Access requires application and approval through dbGaP. Survival data were obtained from the TCGA Pan-Cancer Clinical Data Resource (TCGA-CDR) (<https://gdc.cancer.gov/about-data/publications/PanCan-Clinical-2018>). Driver mutation data were obtained from the ICGC/TCGA pan-cancer driver mutation compendium (TableS3_panorama_driver_mutations_ICGC_samples.public.tsv.gz and TableS3_panorama_driver_mutations_TCGA_samples.controlled.tsv.gz). Driver genes were obtained from IntOGen (<https://www.intogen.org>). The POG570 dataset is publicly available at <https://www.bcgsc.ca/downloads/POG570>. Hotspot mutations were obtained from the Cancer mutation hotspots database v2 [https://www.cancerhotspots.org/files/hotspots_v2.xls]. Source data are provided with this paper and are available via Zenodo [<https://doi.org/10.5281/zenodo.19432064>] (Source Data.zip file).

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	Not applicable.
Reporting on race, ethnicity, or other socially relevant groupings	Not applicable.
Population characteristics	Not applicable.
Recruitment	Not applicable.
Ethics oversight	Not applicable.

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 statistical methods were used to predetermine sample size. All samples in the PCAWG dataset with whole-genome somatic mutation data were included in the analyses. The sample sizes are comparable to those used in previous studies in the field and were sufficient to support the statistical analyses performed.
Data exclusions	For cancer type-specific analyses, only samples from cancer types with a defined list of driver genes in the IntoGen database were included. Survival analyses were restricted to TCGA samples with available clinical endpoint data. No additional exclusions were made.
Replication	All scripts and Jupyter notebooks used for data preprocessing and analysis are publicly available on GitHub (https://github.com/Georgakopoulos-Soares-lab/cancer_passengers) for reproducibility.
Randomization	Randomization was not applicable as all available data was used for analysis. Covariates were accounted for in statistical models.
Blinding	Blinding was not applicable, as all available data were analyzed using computational methods.

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

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.