

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 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

-Compendium of cancer driver genes, downloaded from IntOGen (www.intogen.org)
 -Mutations observed in cancer genes, downloaded from IntOGen (www.intogen.org)
 -TP53 saturation mutagenesis (Kakudo et al., 2005; Kato et al., 2003; Kawaguchi et al., 2005), downloaded from <http://p53.iarc.fr/DownloadDataset.aspx>
 -Rare Oncogenic Variants (Kim et al., 2016), downloaded from http://cancerdiscovery.aacrjournals.org/hi_gswire/filestream/39011/field_highwire_adjunct_files/3/162070_1_supp_3443528_25fwsn.xlsx
 -High-throughput phenotyping of somatic mutations in lung cancer (Berger et al., 2016), downloaded from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5003022/bin/NIHMS801008-supplement-5.xls>
 -CLINVAR variants (Landrum et al., 2018), downloaded from https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/
 -OncoKB variants (Chakravarty et al., 2017), downloaded from <https://www.oncokb.org>
 -COSMIC SBS signatures v2 (Alexandrov et al., 2013), downloaded from https://cancer.sanger.ac.uk/cosmic/signatures_v2
 -PhosphoSitePlus data (Hornbeck et al., 2015), downloaded from <https://www.phosphosite.org/>
 -PhyloP (hg38) data (Pollard et al., 2010), downloaded from <http://hgdownload.cse.ucsc.edu/goldenpath/hg38/phyloP100way>
 -Ensembl Variant Effect Predictor (version 92; McLaren et al., 2010), downloaded from https://www.ensembl.org/info/docs/tools/vep/dbNSFP_v4 (Liu et al., 2020)

Data analysis

We designed boostDM, a machine learning approach to identify all possible driver mutations in a cancer gene, thereby carrying out an in silico saturation mutagenesis of all cancer genes across tumor types. The design, implementation and benchmark of boostDM is thoroughly described in a supplementary note to the main manuscript. BoostDM code is available at bitbucket.org/bbglab/boostdm.

In the implementation of boostDM and in analyses shown in the manuscript, we employed the following programs.

- bgvep, developed by the Barcelona Biomedical Genomics Groups, available at <https://pypi.org/project/bgvep>
- deconstructSigs (Rosenthal et al., 2016), available at <https://github.com/raerose01/deconstructSigs>
- dNdScv (Martincorena et al., 2017), available at <https://www.sanger.ac.uk/science/tools/dndscv>
- jupyter software, available at <https://jupyter.org/>
- Matplotlib library (Hunter, 2007), available at <https://matplotlib.org/>
- numpy library (Oliphant, 2015), available at <https://www.numpy.org>
- Nextflow software (Di Tommaso et al., 2017), available at <https://www.nextflow.io/>
- pandas library (McKinney, 2017), available at <https://pandas.pydata.org/>
- Python programming language, available at <https://www.python.org>
- xgboost library (Chen and Guestrin, 2016), available at <https://xgboost.readthedocs.io/>
- scipy library (Virtanen et al., 2020), available at <https://www.scipy.org/scipylib>
- scikit-learn library (Pedregosa et al., 2011), available at <https://scikit-learn.org/stable/>
- shap library (Lundberg and Lee, 2017), available at <https://github.com/slundberg/shap>

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 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

Cancer genes and their observed mutations across ~28,000 tumors of 66 cancer types, the starting point to train boostDM models, are available through IntOGen (intogen.org). Mutational features computed across cancer genes, and positive and negative sets needed to train boostDM models are available in a Zenodo repository (DOI: 10.5281/zenodo.4813082). Data source files point towards this repository. Blueprints may be browsed through intogen.org/boostDM, and newly sequenced tumors may be analyzed employing boostDM models via the Cancer Genome Interpreter (cancergenomeinterpreter.org).

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	2080 cancer gene-tissue combinations are assayed for the representativeness of the set of observed mutations and the feasibility to build boostDM models; ultimately (see below) 185 good-quality models are built
Data exclusions	-270 cancer gene-tissue combinations were excluded because their estimated fraction of drivers among all observed mutations is below 0.85. -797 cancer gene-tissue combinations were excluded because their number of observed mutations does not allow to cull training sets of at least 20 mutations -98 cancer gene-tissue combinations were excluded because they did not attain the quality threshold (combining F-score50, number of observed mutations and discovery index).
Replication	The 185 selected models were systematically tested: a) A cross-validation of the base classifiers on positive (observed) and negative (synthetic) mutations was carried out b) Models of 133 cancer gene-tissue combinations were re-trained on 90% of the samples available in IntOGen and their performance on the mutations of the remaining 10% was tested c) Their performance was compared to that of experimental saturation mutagenesis assays and computational tools. In this comparisons, the base classifiers were used to avoid leakage of information in the evaluation of models. d) They were tested on a set of rare experimentally validated oncogenic and benign mutations (leaving them out of the training and test sets). e) Their performance of the discrimination between somatic pathogenic and germline benign variants in cancer genes was assessed
Randomization	Train (observed) and test (synthetic) sets of mutations employed by the base classifiers of the 283 models built (including the 185 finally selected) were randomly selected.
Blinding	In the tests described above, the models were blind to the sets of positive and negative mutations.

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