nature research

Corresponding author(s): Douglas Fowler, Jay Shendure

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	PacBio Base call files were converted from the bax format to the bam format using bax2bam (https://github.com/PacificBiosciences/pbccs). PacBio consensus sequences for each sequenced molecule in every library were determined using the Circular Consensus Sequencing 2 algorithm with default conditions (ccs, https://github.com/PacificBiosciences/pbccs). Each resulting consensus sequence was then aligned to either the TPMT or PTEN reference sequence using Burrows-Wheeler Aligner60 (http://bio-bwa.sourceforge.net/). Sequencing reads were converted to fastq format and de-multiplexed with bcl2fastq. Barcode paired sequencing reads for PTEN experiments 1 through 4 were joined using the fastq-join tool within the ea-utils package (http://expressionanalysis.github.io/ea-utils/) using the default parameters.
Data analysis	Enrich2 v1.1.0 was used to count the barcodes, associate each barcode with a nucleotide variant, and then translate and count both the unique-nucleotide and unique-amino acid variants. FACSDIVA v8 and FlowJo v10 were used for FACS / flow cytometry collection and analysis, respectively. RStudio v1.0.136 was used for the analysis, with the versions of the relevant packages used listed in Supplementary File 1.

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

Upon publication, all raw sequence data and function scores will be made freely available for all academic users, by nonexclusive license under reasonable terms to commercial entities that have committed to open sharing of PTEN and TPMT sequence variants, and under a free non-exclusive license to non-profits entities. The data presented in the manuscript are available as Supplementary Tables. The Illumina and PacBio raw sequencing files and barcode-variant maps can be accessed at the NCBI Gene Expression Omnibus (GEO) repository under accession number GSE108727.

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

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	Preliminary experiments were performed to assess noise and determine the number of samples that could be processed through the experimental work flow. A total of eight replicate experiments were then performed to obtain adequate estimates of variance for the majority of individual variants within the libraries. Sample sizes for downstream analyses were dependent upon the availability of the data present in the publicly accessible databases (e.g. ClinVar, TCGA, etc).
Data exclusions	As described in the methods section, experiments with poor technical replication across multiple bins were reamplified and resequenced in their entirety, resulting in the eight biological replicate experiments we report.
Replication	Confidence intervals were calculated for each abundance measurement. The majority of findings we discuss involve low abundance variants, which had reliable measurements across replicates. For a handful of variants, we also reproduced abundance scores and phenotypes with independent, orthogonal experiments (eg. testing variants for fluorescence individually, comparing variant scores with published western blotting phenotypes, and in-house western blot assays).
Randomization	There were no samples/organisms/participants allocated into experimental groups.
Blinding	There was no group allocation.

Materials & experimental systems

Policy information about availability of materials

n/a	Involved in the study
☐	☒ Unique materials
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Research animals
☒	☐ Human research participants

Unique materials

Obtaining unique materials	All unique materials are readily available from the authors.

Antibodies

Antibodies used	Western blotting was performed using a 1:2,000 dilution of anti-phospho-AKT (T308; 13038; Cell Signaling Technology) followed by detection with a 1:10,000 dilution of anti-rabbit-HRP (NA934V; GE Healthcare); a 1:2,000 dilution of anti-pan-AKT (2920; Cell Signaling Technology) followed by detection with a 1:10,000 dilution of anti-mouse-HRP (NA931V; GE Healthcare); a 1:4,000 dilution of anti-GFP antibody (11814460001;Roche), followed by detection with a 1:10,000 dilution of anti-mouse-HRP; 1:5,000 dilution of anti-HA-HRP (3F10; Roche); or a 1:5,000 dilution of anti-beta-actin-HRP (ab8224; Abcam).
Validation	These antibodies have been previously validated in our laboratories, and result in the predominant protein bands of the

expected molecular weight. Based on the manufacturer's website, the anti-beta-actin antibody has been used in 176+ publications as a loading control. Control cells lacking HA- or GFP- tagged proteins were first tested in western blotting experiments to validate the correct identification of the correct-sized band that appeared when HA- or GFP-tagged proteins were expressed. Furthermore, specificity of the antibodies were confirmed through observation of expected relative band intensities comparing variants of known effect. Based on the manufacturer's website, the pan- and phospho-AKT antibodies have been used in at least 19 and 104 publications, respectively. Control variants included in our experiments recapitulated observations made by other groups using similar antibodies to test for the same effects.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HEK 293T (ATCC® CRL-3216™) was the source cell line used, and it was obtained directly from ATCC. These cells were then modified through genome engineering steps performed in our laboratory to create the HEK 293T TetBxb1BFP Clone4 cells used in the experiments.
Authentication	The HEK 293T-based cell lines have not been authenticated.
Mycoplasma contamination	The HEK 293T TetBxb1BFP Clone4 and Clone37 cell lines used in this manuscript have tested negative for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.

Method-specific reporting

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ Magnetic resonance imaging

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Cells were prepared for sorting by lifting from 10 cm plates with Versene solution (0.48 mM EDTA in PBS), washing 1X in PBS, resuspending in sort buffer (1X PBS + 1% heat-inactivated FBS, 1 mM EDTA and 25 mM HEPES pH 7.0) and filtering through 35 µm nylon mesh.
Instrument	Cells were sorted on a BD Aria III FACS machine using an 85 or 100 µm nozzle. mTagBFP2, expressed from the unrecombined landing pad, was excited with a 405 nm laser, and emitted light was collected after passing through a 450/50 nm band pass filter. EGFP, expressed after successful recombination of the variant or library plasmid, was excited with a 488 nm laser, and emitted light was collected after passing through 505 nm long pass and 530/30 nm band pass filters. mCherry, also expressed after successful recombination of the variant or library plasmid was excited with a 561 nm laser, and emission was detected using 600 nm long pass and 610/20 band pass filters. Analytical flow cytometry was performed with a BD LSR II flow cytometer, equipped with filter sets identical to those described for the Aria III, with the exception of mCherry emission which was detected using 595nm long pass and 610/20 band pass filters.
Software	FACS Diva was used to collect the data. FlowJo_V10 was used for analysis.
Cell population abundance	The entire cell populations sorted were used for downstream analysis.
Gating strategy	Before analysis of fluorescence, live, single cells were gated using FSC-A and SSC-A (for live cells) or FSC-A and FSC-H (for single cells) signals. Recombinant mTagBFP2 negative, mCherry positive cells were isolated, with mCherry fluorescence values at least 10 times higher than the median fluorescence value of negative or control cells, and mTagBFP2 fluorescence at least 10 times lower than the median of the unrecombined mTagBFP2 positive cells (See Supplementary Fig. 1a for gating example)

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.