








Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells

Robin M Meyers^{1,7} , Jordan G Bryan^{1,7}, James M McFarland¹, Barbara A Weir¹, Ann E Sizemore¹, Han Xu¹, Neekesh V Dharia^{1–4} , Phillip G Montgomery¹ , Glenn S Cowley¹, Sasha Pantel¹, Amy Goodale¹, Yenarae Lee¹, Levi D Ali¹, Guozhi Jiang¹, Rakela Lubonja¹, William F Harrington¹, Matthew Strickland¹, Ting Wu¹, Derek C Hawes¹, Victor A Zhivich¹, Meghan R Wyatt¹, Zohra Kalani¹, Jaime J Chang¹, Michael Okamoto¹, Kimberly Stegmaier^{1–4}, Todd R Golub^{1–5} , Jesse S Boehm¹ , Francisca Vazquez^{1,2}, David E Root¹, William C Hahn^{1,2,4,6}  & Aviad Tsherniak¹ 

The CRISPR–Cas9 system has revolutionized gene editing both at single genes and in multiplexed loss-of-function screens, thus enabling precise genome-scale identification of genes essential for proliferation and survival of cancer cells^{1,2}. However, previous studies have reported that a gene-independent antiproliferative effect of Cas9-mediated DNA cleavage confounds such measurement of genetic dependency, thereby leading to false-positive results in copy number-amplified regions^{3,4}. We developed CERES, a computational method to estimate gene-dependency levels from CRISPR–Cas9 essentiality screens while accounting for the copy number-specific effect. In our efforts to define a cancer dependency map, we performed genome-scale CRISPR–Cas9 essentiality screens across 342 cancer cell lines and applied CERES to this data set. We found that CERES decreased false-positive results and estimated sgRNA activity for both this data set and previously published screens performed with different sgRNA libraries. We further demonstrate the utility of this collection of screens, after CERES correction, for identifying cancer-type-specific vulnerabilities.

Large-scale efforts using loss-of-function genetic screens to systematically identify genes essential for the proliferation and survival of cancer cells have been reported^{1–10}. Genes identified by these approaches may represent specific cancer cell genetic vulnerabilities that can be used to guide the development of treatment strategies and novel therapeutics. The CRISPR–Cas9 genome-editing system has proven to be a powerful tool for multiplexed screening, owing to its relative ease of application and higher specificity than that of RNA-interference technology¹¹.

However, we and others have recently observed that measurements of cell proliferation in genome-scale CRISPR–Cas9 loss-of-function screens are influenced by the genomic copy number of the region targeted by the single guide RNA (sgRNA)–Cas9 complex^{1,3,4}. Targeting Cas9 to DNA sequences within regions of high copy number gain creates multiple DNA double-strand breaks, thus inducing a gene-independent DNA-damage response and a phenotype of G2 cell-cycle arrest³. This systematic sequence-independent effect due to DNA cleavage (copy number effect) confounds the measurement of the consequences of gene deletion on cell viability (gene-knockout effect) and is detectable even among low-level copy number amplifications and deletions. In particular, this phenomenon hinders interpretation of experiments performed in cancer cell lines that contain many genomic amplifications, because genes in these regions represent a major source of false positives^{3,4}. Existing methods to handle the copy number effect adopt filtering schemes⁹, which preclude examination of data from within amplified regions and ignore the effect at low-level alterations. Here, we present CERES, a method to estimate gene dependency from essentiality screens while computationally correcting the copy number effect, thus enabling unbiased interpretation of gene dependency at all levels of copy number.

In our efforts to define a cancer dependency map, i.e., a catalog of cell-line-specific genetic and chemical vulnerabilities^{10,12}, we performed genome-scale CRISPR–Cas9 loss-of-function screens in 342 cancer cell lines representing 27 cell lineages (Supplementary Table 1; URLs), using the Avana sgRNA library¹³ (Supplementary Table 2), and assessed the cell-proliferation effects of introducing each sgRNA (Online Methods). After application of quality-control measures, receiver operating characteristic analysis of sgRNAs targeting ‘gold standard’ common core essential and nonessential

¹Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. ²Dana-Farber Cancer Institute, Boston, Massachusetts, USA. ³Boston Children's Hospital, Boston, Massachusetts, USA. ⁴Harvard Medical School, Boston, Massachusetts, USA. ⁵Howard Hughes Medical Institute, Chevy Chase, Maryland, USA. ⁶Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA. ⁷These authors contributed equally to this work. Correspondence should be addressed to A.T. (aviad@broadinstitute.org) or W.C.H. (william_hahn@dfci.harvard.edu).

Received 7 April; accepted 4 October; published online 30 October 2017; doi:10.1038/ng.3984

genes¹⁴ demonstrated high screen quality in all cell lines (Fig. 1a). This collection of screens surpasses the scale of existing comparable data sets by approximately tenfold. To confirm the generalizability of our results in independent screens performed with different sgRNA libraries, we reanalyzed two published data sets derived from screens across 33 cancer cell lines of diverse cell lineage (GeCKOv2)³ and 14 acute myeloid leukemia cell lines (denoted Wang 2017)⁹ (Supplementary Fig. 1a).

Using genomic copy number data from the Cancer Cell Line Encyclopedia (CCLE)¹⁵, we assessed the 342 cell lines screened in our Avana data set for sensitivity to the copy number effect, as in Aguirre *et al.*³. In consonance with previous observations, the relationship held in every cell line in our panel: sgRNAs targeting more genomic loci were on average more depleted, frequently to levels similar to or below the depleted levels of sgRNAs targeting cell-essential genes (Fig. 1b and Supplementary Fig. 1b,c). In each of the three data sets, some of the observed variability in sensitivity

was explained by the p53 mutational status of each line in CCLE (Supplementary Fig. 1d).

To quantify the extent to which this sgRNA-level effect translates into false-positive gene dependencies, we ranked the genes in each cell line by the average depletion of their targeting sgRNAs (average guide score). In an example breast cancer cell line, HCC1419, the list of high-ranking genes was enriched in both genes involved in fundamental cellular processes and genes with amplified copy number (Fig. 1c). The depletion ranks of the 100 genes with the largest copy number measurements were significantly higher than expected for most cell lines (298/342 with $P < 0.05$, one-sample one-tailed Kolmogorov–Smirnov test; Fig. 1d and Supplementary Fig. 2a), and the extent of enrichment was significantly correlated with the average copy number of these genes (Spearman $\rho = 0.61$, $P < 10^{-15}$), in agreement with the results of previous studies (Supplementary Fig. 2b).

To decouple the gene-knockout effect from the copy number effect, we developed CERES, which computationally models the measured

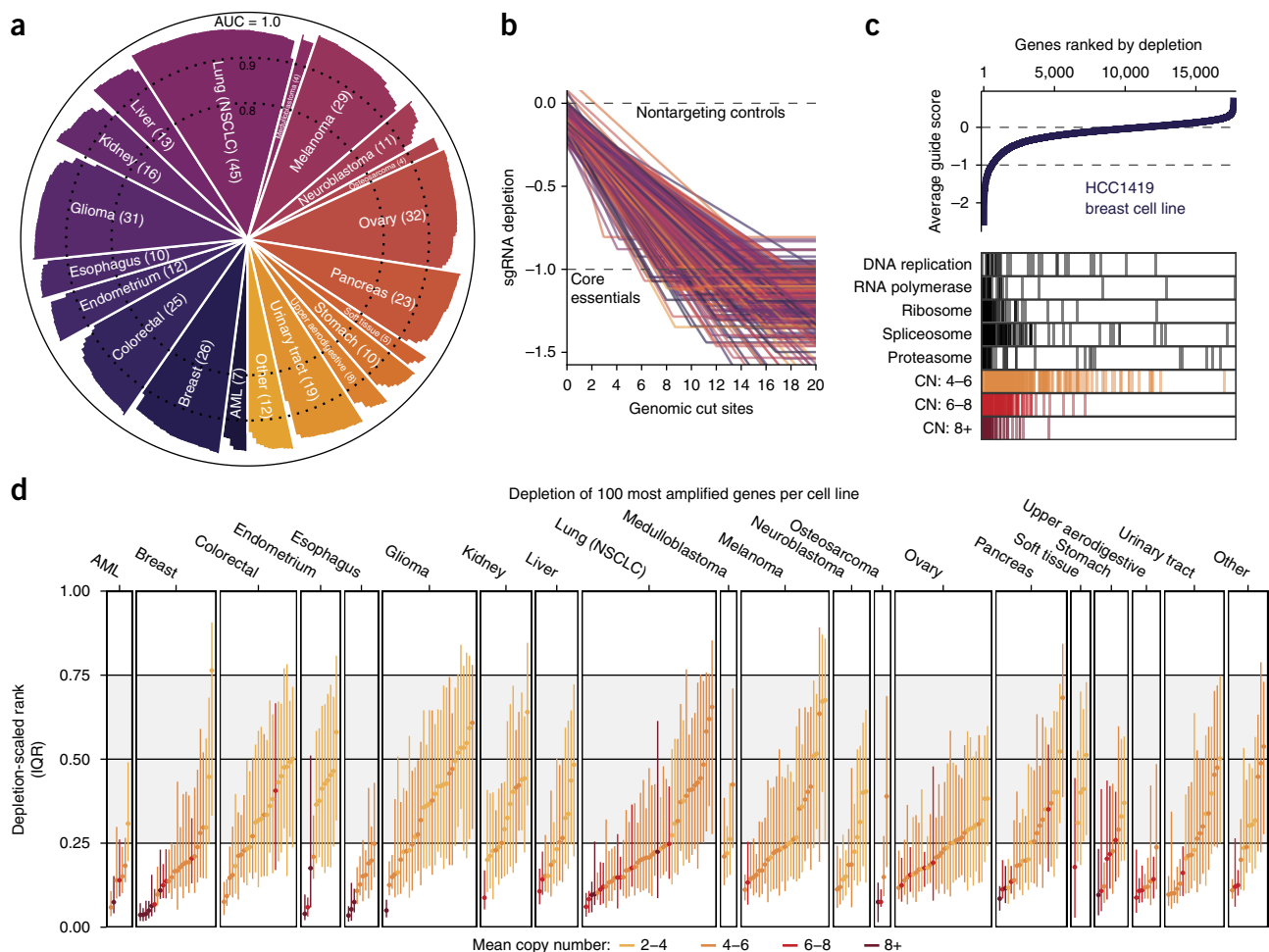


Figure 1 Genomic copy number confounds the interpretation of CRISPR–Cas9 loss-of-function proliferation screens of cancer cell lines. **(a)** Screen quality for each cell line in the panel ($n = 342$), as measured by the area under the receiver operating characteristic curve (AUC) in discriminating between predefined sets of common core essential and nonessential genes. AML, acute myeloid leukemia; NSCLC, non-small-cell lung cancer. **(b)** The depletion of sgRNAs is regressed against the number of perfect-match genomic cut sites by using a simple saturating linear fit, which is plotted for each cell line, colored by lineage, and scaled such that the median of sgRNAs targeting cell-essential genes is at -1 , marked by a dashed line. **(c)** Genes are ranked by the mean depletion of targeting sgRNAs (average guide score) and plotted for an example cell line. Values of 0 and -1 represent the median scores of nonessential and cell-essential genes, respectively, as indicated by dashed lines. Below, depletion ranks of genes involved in fundamental cell processes and genes at various ranges of copy number (CN) amplification are shown. **(d)** The median and interquartile range (IQR) of depletion ranks for the 100 most amplified genes per cell line are plotted. Color indicates the mean amplification level of these genes. The gray-shaded area indicates the IQR of all genes screened.

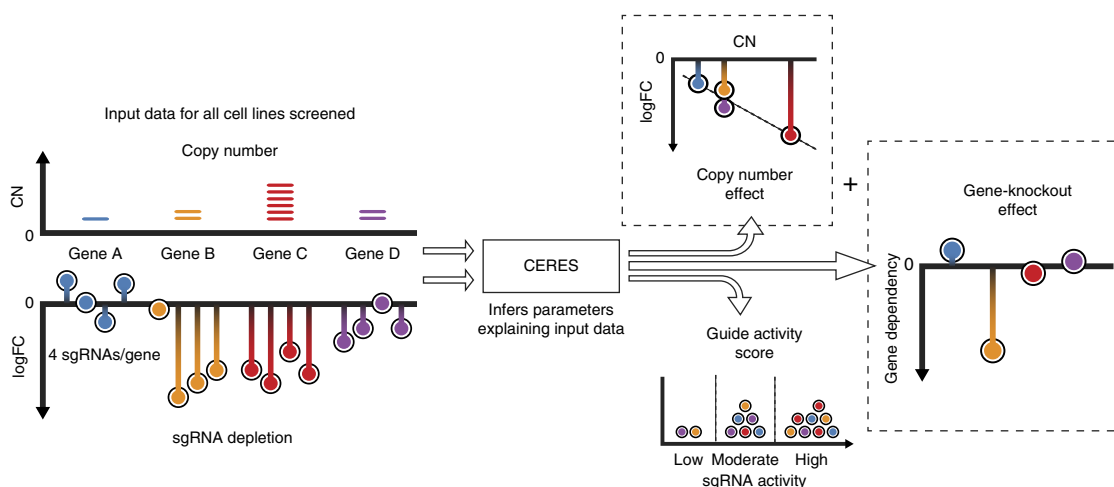


Figure 2 Schematic of the CERES computational model. As input, CERES takes sgRNA-depletion and copy number data for all cell lines screened. During the inference procedure, CERES models the depletion values as a sum of gene-knockout and copy number effects, multiplied by a guide activity score parameter. CERES then outputs the values of the parameters that produce the highest likelihood of the observed data under the model. FC, fold change.

sgRNA depletion (D) as a sum of these two effects (**Fig. 2** and Online Methods). Specifically, for each sgRNA i and cell line j , CERES assumes the following model

$$D_{ij} = q_i \left(\sum_{k \in G_i} (h_k + g_{kj}) + f_j \left(\sum_{l \in L_i} C_{lj} \right) \right) + o_i + \varepsilon \quad (1)$$

where ε is a zero-mean, independent Gaussian noise term. The gene-knockout effect is a sum of cell-line-specific (g_{kj}) and shared (h_k) effects, which are aggregated across any gene targeted by sgRNA i (G_i). The copy number effect is modeled by a piecewise linear spline, f_j , evaluated at the number of genomic sites targeted, determined by

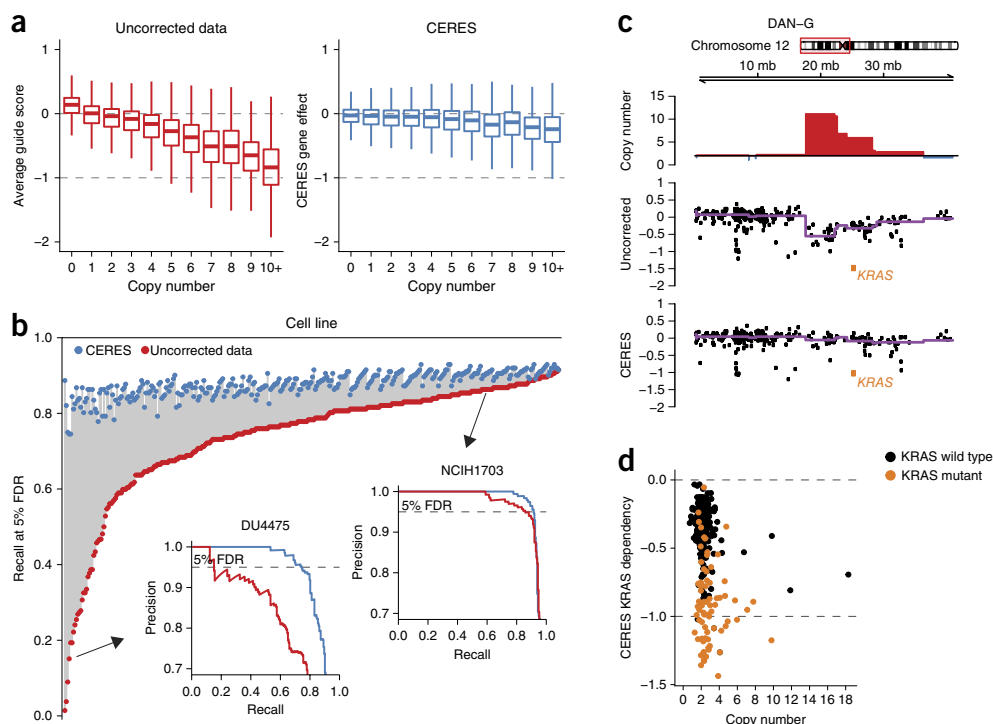


Figure 3 CERES corrects the copy number effect and improves the specificity of CRISPR-Cas9 essentiality screens while preserving true gene dependencies. **(a)** Box plots of gene-dependency scores are shown across copy number for uncorrected average guide scores and CERES gene-dependency scores. Data are scaled as in **Figure 1c**. Center line and box limits, median and upper and lower quartiles; whiskers, extremes of the data not more than 1.5× the interquartile range. **(b)** The recall of cell-essential genes at a 5% FDR of nonessential genes is plotted for each cell line before (red) and after (blue) CERES correction. Precision-recall curves are inset for example cell lines with poor recall (bottom left) and good recall (top right) before CERES correction. **(c)** An example amplified region on chr12p is shown for the DAN-G pancreatic cell line. The top track represents copy number, with amplifications shown in red. The middle and bottom tracks show the average guide score and CERES score, respectively, for each gene in this region. The purple line represents the median value in each copy number segment. *KRAS* is highlighted in orange. **(d)** *KRAS* gene dependency and copy number are shown for all cell lines after CERES correction, and mutant *KRAS* lines are shown in orange.

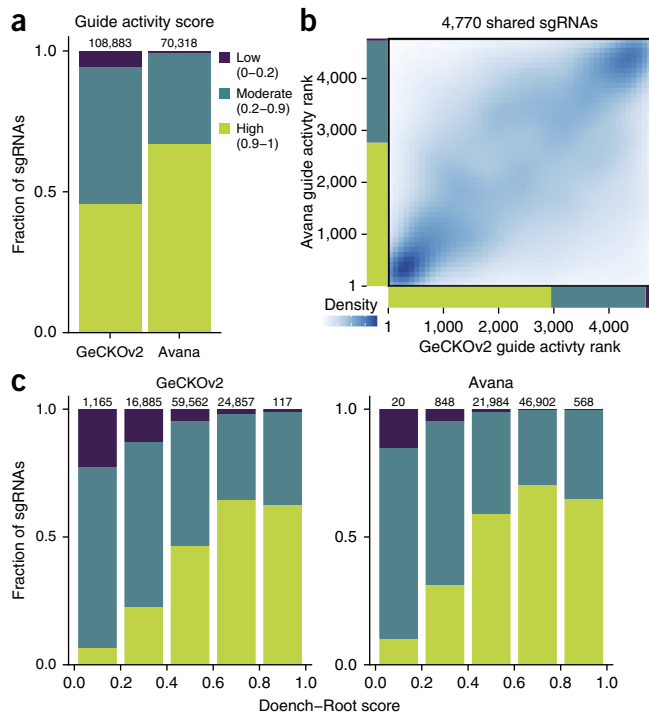


Figure 4 CERES estimates guide activity scores for each sgRNA. (a) sgRNAs are binned into groups with high (0.9–1.0), moderate (0.2–0.9), and low (0–0.2) guide activity scores. The compositions of guide activity scores are shown for the set of screens performed with the GeCKOv2 and the Avana sgRNA libraries. (b) For the set of 4,770 sgRNAs shared between the GeCKOv2 and Avana libraries, sgRNAs are ranked by guide activity scores in each data set and are plotted against each other, with darker blue representing a greater density of sgRNAs. (c) sgRNAs are binned by predicted on-target activity by using the Doench–Root score, and the composition of CERES-estimated guide activity scores is shown for each data set.

the target loci (L_i) and the copy number at each locus (C_{ij}) (Online Methods). The cumulative depletion effects are then scaled by a guide activity score (q_i), restricted to values between 0 and 1, to capture and mitigate the influence of low-quality reagents^{13,16,17}. The offset term σ_i accounts for noise in the measurement of sgRNA abundance in the reference pool (Online Methods). CERES infers the gene-knockout effects and all other parameters by fitting the model to the observed data via alternating least-squares regression (Online Methods). The inferred gene-knockout effects are then scaled per cell line such that scores of 0 and -1 represent the median effects of nonessential genes and common core essential genes, respectively.

We applied CERES to the Avana data set of 342 essentiality screens, as well as to the GeCKOv2 and Wang 2017 data sets, and analyzed the inferred gene-knockout effects (Supplementary Tables 3–5). As expected, CERES markedly decreased the correlation between copy number and gene dependency found in the uncorrected average guide scores (Fig. 3a and Supplementary Fig. 3a) and nearly entirely removed the relationship among unexpressed genes, as determined from CCLE expression data (Supplementary Fig. 3b). For each gene, we correlated its copy number measurements to its dependency scores before and after correction and found that CERES shifted the mean correlation to near zero (Supplementary Fig. 3c). CERES also improved the identification of essential genes in 339 of 342 screens, as measured by the recall of common core essential genes at a 5% false discovery rate (FDR) of nonessential genes², by

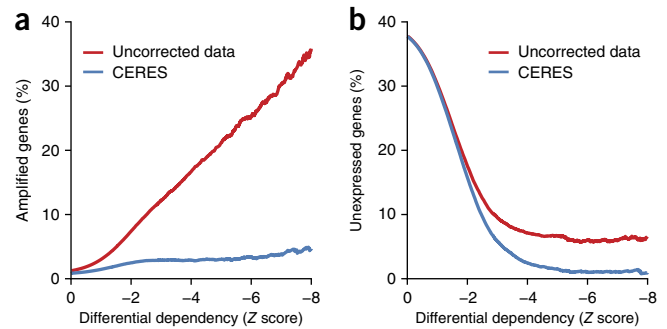


Figure 5 CERES decreases false-positive differential dependencies. (a) The percentage of genes in amplified regions (copy number >4) below a given differential dependency threshold is plotted for the uncorrected average guide score in red and the CERES gene-dependency score in blue. (b) The percentage of unexpressed genes ($\log_2(\text{reads per kb transcript per million mapped reads}) < -1$) below a given differential dependency score is plotted as in a.

an average of 13.8% (Fig. 3b and Supplementary Fig. 4a) (Online Methods). This improvement was substantially better than a simple linear model used to correct the relationship between average guide score and copy number (Supplementary Fig. 4b) (Online Methods). Furthermore, CERES preserved an average of 134 genes per cell line that would have been removed with a simple filtering scheme. On average, six of these filtered genes per cell line scored as essential below a threshold of -0.6 after CERES correction (Supplementary Fig. 4c). Reassuringly, CERES preserved expected cancer-specific dependencies, even in amplified regions, such as *KRAS* in an example amplification on chromosome (chr) 12p of the DAN-G pancreatic cancer cell line (Fig. 3c and Supplementary Fig. 5). Additionally, *KRAS*-mutant cell lines remained substantially enriched over wild type for *KRAS* gene dependency (Fig. 3d), a result that generalized to other known oncogenes (Supplementary Fig. 6).

CERES estimates a guide activity score for each sgRNA used in the screens (Supplementary Tables 6–8). Although experimentally validating the activity of all, or even most, sgRNAs in a genome-scale library is unfeasible, sequence determinants have proven useful in the prediction of on-target activity^{13,18,19}. The Avana sgRNA library was optimized through such predictions. Fittingly, CERES estimated higher guide activity scores on average for the Avana data set relative to GeCKOv2, with a nearly 20-fold increase in the ratio of high- to low-activity sgRNAs (161.3 to 1 and 8.3 to 1; Fig. 4a). The guide activity scores for the 4,770 sgRNAs common to both libraries showed substantial agreement (Spearman $\rho = 0.53$, $P < 10^{-15}$; Fig. 4b), thus demonstrating that CERES captured a measure of sgRNA activity that was reproducible across independent collections of screens (Supplementary Fig. 7a,b). For both the GeCKOv2 and Avana libraries, we compared CERES guide activity scores to sequence-based predictions of sgRNA activity (Doench–Root scores)¹³ and found significant correspondence (Avana, Pearson $\rho = 0.21$, $P < 10^{-15}$; GeCKOv2, Pearson $\rho = 0.37$, $P < 10^{-15}$; Fig. 4c). Together, these results demonstrate that the guide activity scores inferred by CERES are useful for estimating gene-knockout effects, and they further suggest that CERES could be used to aid in the selection of reagents for follow-up experiments.

To identify cancer-specific genetic vulnerabilities, we used a metric of differential dependency representing the strength of dependency in a cell line relative to all other lines screened (Online Methods). We assessed an upper bound on the number of false-positive differential dependencies due to copy number amplifications by calculating the

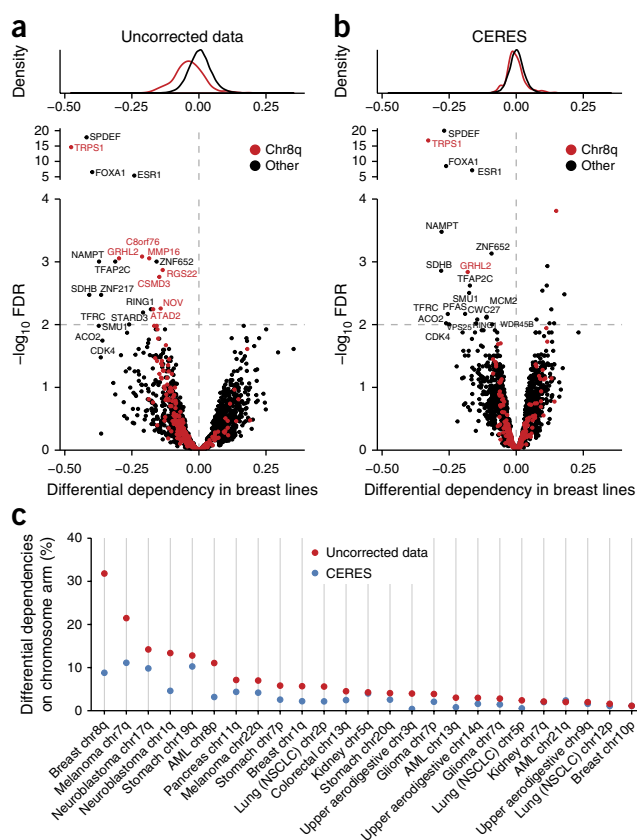


Figure 6 CERES decreases false positives among lineage-specific differential dependencies due to recurrently amplified chromosome arms. (a) The distributions of differential dependencies in breast lines are plotted in red for genes on chr8q (commonly gained in breast tumors) and in black for all other genes. Below, the differential dependency of each gene is plotted against the FDR-corrected P value, calculated from two-tailed Student's t test, with colors as above. The dashed line represents an FDR of 5%. (b) Data are shown for CERES-inferred gene effects as in a. (c) The percentages of lineage-specific differential dependencies (FDR < 0.05) that are on recurrently amplified chromosome arms are shown, before and after CERES correction.

percentage of amplified genes at every possible threshold of differential dependency. In the uncorrected data, the percentage of amplified genes increased at stronger dependency thresholds, climbing above 30% at the highest levels of differential dependency, an effect that CERES substantially decreased (Fig. 5a and Supplementary Fig. 8a). We next used a similar procedure to examine unexpressed genes, whose deletion or editing was not expected to induce phenotypic effects and which would represent an overt source of false positives if scored as differentially dependent. We found that, for genes below a differential dependency of -8 , CERES decreased the percentage of unexpressed genes from 6.6% to 0.9%, thus indicating a substantial improvement in specificity (Fig. 5b and Supplementary Fig. 8b).

A data set of this scale enables the discovery of genetic vulnerabilities specific to a subset of cancer cell lines defined by some cellular context, such as cell lineage. We hypothesized that, in this setting, copy number effects driven by recurrent copy number alterations, even with small effect sizes, could introduce false positives. For each gene, we compared average guide scores in 26 breast cancer cell lines to those of all other cell lines (Online Methods). Indeed, we found several differential dependencies resident on chr8q, which was recurrently amplified in breast tumors (Fig. 6a). However, when we used

CERES-corrected dependency scores, we found that only two of the original chr8q genes, *TRPS1* and *GRHL2*, remained (Fig. 6b). To confirm this finding by using a complementary assay, we analyzed this set of genes in a data set derived from genome-scale RNA-interference screens across 501 cancer cell lines¹⁰. We found that these were the only two genes on chr8q that scored as differentially dependent in the 34 breast lines, whereas most genes in other regions were validated (Supplementary Fig. 9a,b). Previous studies have implicated these transcription factors in breast cancer progression^{20,21}, and the high expression levels of these and other transcription factors in breast lines identified suggest that they are likely to be true differential dependencies (Supplementary Fig. 9c). We extended this analysis to all cell lineages with recurrently amplified chromosome arms and quantified the enrichment of differential dependencies before and after CERES correction in each context. We observed that CERES decreased the fraction of differential dependencies on the recurrently amplified chromosome arm in 24 out of 25 such cases (Fig. 6c and Online Methods).

Although CERES leverages data across many cell lines to infer guide activity scores, we confirmed that this approach can be applied to data sets of any size—even a screen of a single cell line—given predetermined guide activity scores. These scores may be precomputed from a larger set of screens, predicted with available tools, or assumed to be uniform. In random subsamplings of cell lines from the Avana data set, CERES performed nearly as it did when it was applied to the full set. Furthermore, we tested CERES on single cell lines, using fixed uniform guide activities, and found that the median improvement per cell line was more than 97% that of the run on all 342 cell lines (Supplementary Fig. 10 and Online Methods).

In summary, we introduce a large set of uniformly performed CRISPR–Cas9 essentiality screens of cancer cell lines, propose a methodology to estimate gene dependency while removing false positives due to copy number effects, and demonstrate the power of these two resources in identifying genetic vulnerabilities of cancer. To facilitate the use of the Avana data set and CERES, we have made the software available as an R package (URLs), along with all data and analysis scripts used in this study.

URLs. CERES, <https://depmap.org/ceres/>.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the [online version of the paper](#).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This work was supported by grants U01 CA176058, U01 CA199253, and P01 CA154303 (W.C.H.) and by the Slim Initiative for Genomic Medicine, a project funded by the Carlos Slim Foundation and the H.L. Snyder Foundation.

AUTHOR CONTRIBUTIONS

R.M.M., J.G.B., and A.T. conceived and designed the study. R.M.M., J.G.B., and J.M.M. performed computational analysis and interpretation of results. J.G.B. wrote and implemented the modeling software. R.M.M., B.A.W., and A.E.S. processed and managed data. H.X. and N.V.D. assisted with computational analysis. P.G.M. provided computational tools. G.S.C., S.P., and E.V. provided project management. A.G., Y.L., L.D.A., G.J., R.L., W.F.H., M.S., T.W., D.C.H., V.A.Z., M.R.W., Z.K., J.J.C., and M.O. assisted with data generation. R.M.M., J.G.B., J.M.M., W.C.H., and A.T. wrote and/or revised the manuscript with assistance from other authors. K.S., T.R.G., J.S.B., F.V., D.E.R., W.C.H., and A.T. supervised the study and performed an advisory role.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the [online version of the paper](#).

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

- Wang, T. *et al.* Identification and characterization of essential genes in the human genome. *Science* **350**, 1096–1101 (2015).
- Hart, T. *et al.* High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. *Cell* **163**, 1515–1526 (2015).
- Aguirre, A.J. *et al.* Genomic copy number dictates a gene-independent cell response to CRISPR/Cas9 targeting. *Cancer Discov.* **6**, 914–929 (2016).
- Munoz, D.M. *et al.* CRISPR screens provide a comprehensive assessment of cancer vulnerabilities but generate false-positive hits for highly amplified genomic regions. *Cancer Discov.* **6**, 900–913 (2016).
- Cheung, H.W. *et al.* Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Proc. Natl. Acad. Sci. USA* **108**, 12372–12377 (2011).
- Marcotte, R. *et al.* Essential gene profiles in breast, pancreatic, and ovarian cancer cells. *Cancer Discov.* **2**, 172–189 (2012).
- Cowley, G.S. *et al.* Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. *Sci. Data* **1**, 140035 (2014).
- Tzelepis, K. *et al.* A CRISPR dropout screen identifies genetic vulnerabilities and therapeutic targets in acute myeloid leukemia. *Cell Rep.* **17**, 1193–1205 (2016).
- Wang, T. *et al.* Gene essentiality profiling reveals gene networks and synthetic lethal interactions with oncogenic Ras. *Cell* **168**, 890–903.e15 (2017).
- Tsherniak, A. *et al.* Defining a cancer dependency map. *Cell* **170**, 564–576.e16 (2017).
- Fellmann, C., Gowen, B.G., Lin, P.-C., Doudna, J.A. & Corn, J.E. Cornerstones of CRISPR-Cas in drug discovery and therapy. *Nat. Rev. Drug Discov.* **16**, 89–100 (2017).
- Corsello, S.M. *et al.* The Drug Repurposing Hub: a next-generation drug library and information resource. *Nat. Med.* **23**, 405–408 (2017).
- Doench, J.G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).
- Hart, T., Brown, K.R., Sircoulomb, F., Rottapel, R. & Moffat, J. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Mol. Syst. Biol.* **10**, 733 (2014).
- Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
- Li, W. *et al.* MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol.* **15**, 554 (2014).
- Hart, T. & Moffat, J. BAGEL: a computational framework for identifying essential genes from pooled library screens. *BMC Bioinformatics* **17**, 164 (2016).
- Doench, J.G. *et al.* Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat. Biotechnol.* **32**, 1262–1267 (2014).
- Xu, H. *et al.* Sequence determinants of improved CRISPR sgRNA design. *Genome Res.* **25**, 1147–1157 (2015).
- Xiang, X. *et al.* Grhl2 determines the epithelial phenotype of breast cancers and promotes tumor progression. *PLoS One* **7**, e50781 (2012).
- Werner, S. *et al.* Dual roles of the transcription factor grainyhead-like 2 (GRHL2) in breast cancer. *J. Biol. Chem.* **288**, 22993–23008 (2013).

ONLINE METHODS

CRISPR–Cas9 essentiality screening assay. Cancer cell lines were transduced with a lentiviral vector expressing the Cas9 nuclease under blasticidin selection (pXPR-311Cas9). Each Cas9-expressing cell line was subjected to a Cas9 activity assay³ to characterize the efficacy of CRISPR–Cas9 in these cell lines (**Supplementary Table 1**). Cell lines with less than 45% measured Cas9 activity were considered ineligible for screening. Stable polyclonal Cas9⁺ cell lines were then infected in replicate ($n = 3$) at low multiplicity of infection (MOI < 1) with a library of 76,106 unique sgRNAs (Avana), which, after sex chromosomes were filtered out, comprised 70,086 sgRNAs targeting 17,670 genes (approximately four sgRNAs per gene) annotated in the consensus coding-sequence database, and 995 nontargeting control sgRNAs (**Supplementary Table 2**). Cells were selected in puromycin and blasticidin for 7 d and then passaged without selection while a representation of 500 cells was maintained per sgRNA until 21 d after infection. Genomic DNA was purified from endpoint cell pellets, the sgRNA barcodes were PCR amplified with sufficient gDNA to maintain representation, and the PCR products were sequenced with standard Illumina machines and protocols.

Preprocessing and quality control. After sequencing of the sgRNA barcodes, raw barcode counts were deconvoluted from sequence data with PoolQ software (http://portals.broadinstitute.org/gpp/public/dir/download?dirpath=protocols/screening&filename=Pooled_Screening_Deconvolution_using_PoolQ.pdf) and were summed across sequencing lanes. Samples were removed if they did not reach 15 million reads. We calculated normalized read counts for each sample according to the procedure in Cowley *et al.*⁷. We then calculated pairwise Pearson correlation coefficients between replicate samples from the same cell line to identify and remove poor-quality replicates, using a threshold of 0.7. All sample read counts were then divided by their representation in the starting plasmid DNA library (pDNA) to compute a fold change (FC). We computed robustly standardized mean difference (SSMD)²² statistics for the replicates, using FCs between nontargeting control sgRNAs and FCs from sgRNAs targeting the spliceosomal, ribosomal, or proteasomal genes in KEGG gene sets^{23–25}. We removed replicates with SSMDs that did not reach –0.5. We also followed standard fingerprinting procedures to remove mismatched cell lines⁷. The logFC data were then normalized within each cell-line replicate by subtracting the median logFC value and dividing by the median average deviation before input to CERES.

Copy number data. Copy number data for all cancer cell lines were obtained from the CCLE¹⁵ data portal (<https://portals.broadinstitute.org/ccle/>). Copy number data were derived from Affymetrix SNP6.0 arrays. Segmentation of normalized log₂ ratios was performed with the circular binary segmentation (CBS) algorithm. The data set is available at https://data.broadinstitute.org/ccle_legacy_data/dna_copy_number/CCLE_copynumber_2013-12-03.seg.txt.

Gene expression and mutation data. Gene expression and mutation data for all cell lines were obtained from the CCLE data portal. These data sets are available at https://data.broadinstitute.org/ccle/CCLE_RNAseq_081117.rpkm.gct and https://data.broadinstitute.org/ccle/ccle2maf_081117.txt.

sgRNA genome mapping. sgRNA sequences were mapped to the hg19 reference genome with the bowtie short-read aligner (version 1.1.2)²⁶. Bowtie was run with the options ‘-a -v 0’ to find all perfect matches in the genome. Only sgRNAs with fewer than 100 alignments were included, and alignments were filtered to include an NGG protospacer-adjacent motif. Alignments were then mapped to gene coding sequences by using the consensus coding-sequence database.

Model fitting. To fit CERES to input data, we solved the following optimization problem:

$$\begin{aligned} & \underset{\hat{D}}{\text{minimize}} \quad \sum_{i=1}^M \sum_{j=1}^N (D_{ij} - \hat{D}_{ij})^2 + \lambda_g \sum_{k=1}^K \sum_{j=1}^N g_{kj}^2 \\ & \text{subject to} \quad 0 \leq q_i \leq 1, i=1, \dots, M \\ & \quad f_j(C) \leq f_j(C'), \forall C \geq C' \in \mathbb{R}_{\geq 0}, j=1, \dots, N \end{aligned}$$

where \hat{D}_{ij} is computed according to equation (1). The constants M , N , and K in the objective function are, respectively, the total number of sgRNAs, cell lines, and genes in the data set. The right-hand term in the objective function acts as a regularizer on the cell-line-specific deviation from the shared gene-knockout effect, in which the hyperparameter λ_g modulates the strength of the regularization. The first constraint on the model parameters ensures that the guide activity scores are between 0 and 1. The second constraint ensures that the copy number–effect functions are monotonically decreasing in their arguments. Because the objective function is not jointly convex in the model parameters, we fit CERES by using alternating least squares, first solving for the gene-essentiality scores and copy number–effect parameters with the guide activity scores and offsets held constant, then solving for the guide activity scores and offsets as follows.

Algorithm 1.1. CERES alternating minimization.

Given $\varepsilon > 0$

Initialize:

1. Gene-knockout and copy number–effect coefficients $[g, f] := [0, 0]$
2. Guide activity scores and offsets $[q, o] := [1, 0]$

Repeat:

1. Solve for gene-knockout and copy number effects. Compute optimal parameters $[g^*, f^*]$
2. Update. $[g^*, f^*] := [g^*, f^*]$
3. Solve for guide activity scores and offsets. Compute optimal parameters $[q^*, o^*]$
4. Update. $[q^*, o^*] := [q^*, o^*]$
5. Evaluate mean squared error (mse). $mse_t := \|D - \hat{D}\|^2 / MN$
6. Evaluate decrease in error. $\Delta mse := mse_t - mse_{t-1}$
7. Stopping criterion. Quit if $\Delta mse < \varepsilon$

Because of the presence of constraints, we used numerical optimization techniques to solve for the optimal parameters $[g^*, f^*]$ and $[q^*, o^*]$ in steps 1 and 3 (ref. 27). We use the bracket notation $[g, f]$ to indicate that the enclosed parameters are inferred simultaneously as variables in a system of constrained linear equations.

Spline functions. The piecewise linear spline functions f_j in the CERES model equations allow for flexible modeling of the characteristic saturation of the copy number effect at high numbers of cuts. They are implemented with B-spline regression methods and are each parameterized by 25 slope coefficients plus a single intercept parameter. These are inferred directly in the regression that determines the gene-knockout effects. Each spline has an initial knot point at copy number = 0. The additional knot points are determined by running average linkage clustering on the copy number data for each cell line.

Hyperparameter optimization and test-set evaluation. To improve the generalizability of our model and to minimize overfitting of the training data, we regularized the cell-line-specific gene effects. To find the best value of λ_g , we evaluated the mean squared error obtained on a randomly selected held-out validation set (one-tenth of all observations) for each of 25 values of λ_g , sampled log uniformly from the interval [0.01, 1]. After the 25 models were evaluated, the value of λ_g yielding the lowest mean squared error was used to fit the final model on the full set of observations (**Supplementary Fig. 11**). The optimized value of λ_g was 0.562, 0.681, and 0.681 for the Avana, GeCKOv2, and Wang 2017 data sets, respectively.

Model complexity. Given a collection of CRISPR screening data, let N be the number of sgRNAs, M be the number of cell lines, and K be the number of targeted genes in the data set. CERES fits KM cell-line-specific gene-effect parameters and an additional K parameters for the shared gene effects. The model also fits $M(S + 1)$ copy number–effect parameters, where S is the number of copy number segments in each piecewise linear spline, and $2N$ parameters for the guide activity scores and offsets. Ignoring the degrees of freedom lost by regularization and constraints, CERES takes in MN data points and fits $MN(1/N + S/N + 2/M + K/N + K/MN)$ parameters.

Software and implementation. Matrix operations for the optimization procedure were implemented with the open-source C++ linear-algebra library Eigen, version 3.3, available at <http://eigen.tuxfamily.org/>. These operations

were then wrapped into the R statistical software with the ‘RcppEigen’ package, downloaded from <http://cran.r-project.org/>. The optimization routine and final fit for each data set were run with Google Cloud Platform services.

Precision-recall analysis. Precision-recall curves were generated by using the sets of common core essential and nonessential genes defined in Hart *et al.*¹⁴. The best threshold for which >95% of hits are essential genes was calculated for an FDR of 5%. The percentage of all essential genes that scored as hits at this threshold was calculated as the recall at 5% FDR.

Comparison with linear regression. For each cell line, average guide scores were regressed against gene-level copy number data by using a linear model. The fit residuals were taken as the LM-corrected gene-dependency scores. Precision-recall analysis was performed as above.

Subsampling analysis. We simulated CERES performance generalization to other data-set sizes by downsampling from the Avana data set. Specifically, for each number p in the set {1, 2, 4, 8, 16, 32, 64}, we ran $342/p$ trials (rounded up to the nearest integer), such that each cell line appeared once in each run of size p . For each p and each cell line, we evaluated the harmonic mean of precision and recall (referred to as the F1 measure) at the point of equiprobability between the essential and nonessential gene classes. We then compared this number to the F1 measure obtained by running CERES on the full Avana data set. For $P < 5$, we fixed all guide activity scores to a value of 1.

Differential dependency. Differential dependency was calculated as the difference between a single cell line’s dependency score for a given gene and the mean score for that gene across all lines screened, then Z -score-normalized to that cell line’s entire set of differential dependencies to decrease the influence of noisy cell lines. For calculating the fraction of differential dependencies that were amplified or unexpressed, only genes with a negative dependency score in at least one cell line were considered.

Recurrent chromosome-arm amplifications. We called recurrent chromosome-arm amplifications for a lineage across the entire CCLE copy number data set. A chromosome arm was called as amplified if the weighted median of copy number segments on that arm was greater than 2.8. Recurrently amplified chromosome arms for a lineage were then defined with a one-tailed Fisher’s

exact test to test for enrichment of amplified arms in that lineage, at an FDR-corrected P value of 0.05.

Lineage-specific differential dependencies. For every lineage in our data set with at least five cell lines, we calculated the difference in means in gene dependency between cell lines of that lineage and the rest of the data set, and assessed significance with a one-tailed Student’s t test ($df = 340$), for each gene screened. Differential dependencies were called with a negative effect size at a significance of FDR-corrected P value < 0.05 . For each chromosome arm that was recurrently amplified for that lineage, we calculated the fraction of significant differential dependencies on that chromosome arm before and after CERES correction.

Code availability. CERES software, documentation, and code for regenerating analyses and figures can be found at <https://depmap.org/ceres/>.

Data availability. All CRISPR–Cas9 screening data presented here are available at <https://depmap.org/ceres/>. We also have posted these data and all other data sets used for analysis in a Figshare record, available at <https://doi.org/10.6084/m9.figshare.5319388>. A Life Sciences Reporting Summary is available.

22. Zhang, X.D. A pair of new statistical parameters for quality control in RNA interference high-throughput screening assays. *Genomics* **89**, 552–561 (2007).
23. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
24. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2016).
25. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D1, D353–D361 (2017).
26. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
27. Boyd, S. & Vandenberghe, L. *Convex Optimization* 1–730 (Cambridge Univ. Press, 2004).

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work we publish. This form is published with all life science papers and is intended to promote consistency and transparency in reporting. All life sciences submissions use this form; while some list items might not apply to an individual manuscript, all fields must be completed for clarity.

For further information on the points included in this form, see [Reporting Life Sciences Research](#). For further information on Nature Research policies, including our [data availability policy](#), see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

► Experimental design

1. Sample size

Describe how sample size was determined.

N/A

2. Data exclusions

Describe any data exclusions.

see
Online Methods: sections "Cancer cell lines" and "Preprocessing and quality control"

3. Replication

Describe whether the experimental findings were reliably reproduced.

N/A

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

N/A

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

N/A

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or the Methods section if additional space is needed).

n/a Confirmed

- | | | |
|-------------------------------------|-------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | The <u>exact</u> sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.) |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly. |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | A statement indicating how many times each experiment was replicated |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section) |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | A description of any assumptions or corrections, such as an adjustment for multiple comparisons |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | The test results (e.g. p values) given as exact values whenever possible and with confidence intervals noted |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | A summary of the descriptive statistics, including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range) |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | Clearly defined error bars |

See the web collection on [statistics for biologists](#) for further resources and guidance.

► Software

Policy information about [availability of computer code](#)

7. Software

Describe the software used to analyze the data in this study.

see Online Methods: "Software and implementation"

For all studies, we encourage code deposition in a community repository (e.g. GitHub). Authors must make computer code available to editors and reviewers upon request. The *Nature Methods* [guidance for providing algorithms and software for publication](#) may be useful for any submission.

► Materials and reagents

Policy information about [availability of materials](#)

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

No unique materials were used

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

N/A

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

see Online Methods: "Cancer cell lines" CCLE

b. Describe the method of cell line authentication used.

see Online Methods: "Cancer cell lines" SNP fingerprinting

c. Report whether the cell lines were tested for mycoplasma contamination.

see Online Methods: "Cancer cell lines"

d. If any of the cell lines used in the paper are listed in the database of commonly misidentified cell lines maintained by [ICLAC](#), provide a scientific rationale for their use.

No commonly misidentified cell lines were used

► Animals and human research participants

Policy information about [studies involving animals](#); when reporting animal research, follow the [ARRIVE guidelines](#)

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

N/A

Policy information about [studies involving human research participants](#)

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

N/A