Supporting Information

Cheung et al. 10.1073/pnas.1109363108

SI Materials and Methods

Cell Culture and Fingerprinting of Cell Lines. The culture conditions for all cancer cell lines are listed in Table S1. To verify the identity of each cell line, Sequenom genotyping assays for a panel of 48 SNP loci were performed on genomic DNA isolated from each replicate of cell lines at the conclusion of the screen at the Broad Institute Genetic Analysis Platform. A reference "fingerprint" containing 33 of these loci for each cell line was derived from Affymetrix 6.0 array data (http://www.broadinstitute.org/ccle) or prescreen Sequenom genotyping.

Construction of Pooled shRNA Library. The human 54K pool of 54,020 shRNA plasmids from the RNAi Consortium was assembled by combining 16 normalized subpools of ~3,400 shRNA plasmids each. Each subpool was used to transform ElectroMAX DH5α-E cells (Invitrogen) by electroporation and plated onto five 24 × 24-cm^2 bioassay dishes (Nunc). DNA was purified from the plated transformants by using a HiSpeed Plasmid Maxi Kit (Qiagen). These subpools were then combined to create the 54K shRNA pool. Then, 2 μg of this pool was used to transform DH5α cells and plated onto 50 24 × 24-cm^2 bioassay dishes. DNA was purified from the plated transformants and used for virus production. A complete list of shRNAs along with unique TRCN identifiers is publicly available (http://www.broadinstitute.org/rnai/public/).

Virus Pool Production, Infection, and Cell Propagation. Production of lentivirus from the 54K shRNA pool was performed as described (1). A single batch of ~5 L of virus was aliquoted and frozen at −80 °C for all infections.

Infections were performed as described (1) with the following modifications. To determine viral volume that would produce a multiplicity of infection (MOI) of 0.3–0.5 for each cell line, cells were infected with a titration of six different volumes (0–400 μL) of virus and cultured in the presence or absence of puromycin. Before large-scale infection, cells were filtered through a 40-μm cell strainer (BD Falcon). For each of the quadruplicate infections, 3.7 × 10^7 cells were resuspended in 24 mL of medium containing 4 μg/mL polybrene, and the appropriate volume of 54K library lentiviruses was added. This mixture was seeded into a 12-well plate at ~2 mL per well. A spin infection was performed by centrifugation at 930 × *g* for 2 h at 30 °C.

For suspension cells, supernatants were gently aspirated off after infection, and fresh medium was added to the 12-well plates. After 20 h, the 12 wells from each replicate infection were pooled, and the combined cells were transferred into a T175-flask containing 200 mL of medium containing puromycin. At 4 d after selection, for each of the four replicates, 2 × 10^7 cells were plated into a new T175-flask and cultured in 200 mL of medium containing puromycin. For all subsequent passages, 1.1 × 10^7 cells per replicate were carried over. The remaining cells for all passages were collected, resuspended in 1 mL of PBS, and stored at −20 °C for genomic DNA isolation.

For adherent cells, supernatants were gently aspirated off after spin infection, and fresh medium was added to the 12-well plates. After 20 h the 12 wells from each replicate were trypsinized, and cells combined and plated in two T225-flasks containing 60 mL of medium containing puromycin. Passaging for each cell line was continued for at least 16 population doublings or 28 d, whichever was longer. Puromycin selection was maintained for the entire experiment.

In-Line Infection Calculation. At 20 h after large-scale infection, a small fraction of cells (1.5–3 × 10^5) from each replicate were plated into each well of six-well plates in the presence or absence of puromycin. Control wells with 100% uninfected cells were included to verify complete puromycin killing of uninfected cells. Ninety-six hours later, viable cells were counted using trypan blue staining. The infection rate was determined by calculating the number of viable cells selected in puromycin divided by the number of viable cells without puromycin selection. Screening continued only when the infection rates were within the range of 30–65% to provide an MOI = 1 and to yield a sufficient number of cells to provide adequate shRNA representation.

Genomic DNA Isolation and Array Hybridization. Genomic DNA isolation, half-hairpin barcode production, and array hybridization were performed as described (1). For PCR amplification of shRNA sequences, minimum of 50 μg of genomic DNA was used as template for each replicate. Therefore, multiple PCR reactions were performed, each using 3 μg of genomic DNA per 50 μL reaction volume.

Quality Control of Hybridization. Scans of each array were visually inspected to detect spatial irregularities or hybridization profiles with signal out of the linear range. Such aberrant array hybridization data were discarded. Interreplicate agreement for experimental replicates of each cell line was assessed from their MvA plots using the GenePattern module MvAplots, which defines the interquartile range (IQR) value for each pairwise comparison of replicates of a cell line. Replicate pairs that have a calculated IQR value of <1.2 were retained for analysis. To confirm that experimental replicates derived from the same cell line exhibited very small discrepancies compared with intercell line differences, we performed unsupervised hierarchical clustering with a Pearson correlation. Replicates that failed to tightly cluster with each other were discarded. The arrays were also filtered based on the relative difference between the distribution of human and mouse probes in the raw data for each array. Arrays with <30% of human probes with signal above the mouse probe signal were removed. Any line with less than three replicates passing any QC measure was also removed.

Data Preprocessing for Custom TRC shRNA Arrays. Raw .CEL files from custom Affymetrix barcode arrays were processed with a modified version of dCHIP software (1). "Barcode" array type, "average" model method, and fifth percentile of region (PM-only) background selection were used as parameters. "Running median" and "All probes" were chosen as parameters for normalization, and data were logged before further analysis.

shRNA Scoring. After data preprocessing, the GenePattern modules "RNAigctconverter" and "MakeArrayInfo" were used to convert preprocessed data into a .gct file and make a file of array annotations, respectively. Then the module "shRNAscores" was used to collapse values derived from replicate measurements of the abundance of each shRNA in the initial DNA pool in comparison with its abundance at the completion of replicate experiments performed on each cell line using an adjusted log fold change score. The log fold change score is the difference in means between replicates of the cell line of interest and replicates of the initial DNA pool. This score was adjusted to deemphasize shRNAs that showed high variability among replicates of the DNA pool, which likely arises from technical artifacts including shRNA underrepresentation in the initial

DNA pool or suboptimal array probe performance. To penalize these variable scores, we divided the log fold change score by the SD of the DNA pool after it had been mean centered at 1 and floored at 1. The log fold change scores of the least variable shRNAs from reference measurements were unaltered and the scores of the most variable shRNAs were penalized proportional to the SD of their replicate measurements from the reference pool. This adjusted log fold change score was used for subsequent processing.

Scaling and Centering Data Ranges. To normalize the shRNA depletion values between cell lines, the distribution of adjusted log fold change scores of each line was scaled and centered with peak median absolute deviation (PMAD) normalization, a variation of Z score with median absolute deviation (1), using the GenePattern module "NormalizeCellLines." PMAD normalization first centers the shRNAscores per cell line at 0, by subtracting the value of each shRNA from the modeled peak value of the distribution of each cell line. The peak value was obtained by taking the maximum value of the Gaussian smoothed, kernel density estimate of the distribution. The shRNAscores for each cell line were then rescaled so that each line had similar data ranges by dividing the centered data for each shRNA by the median absolute deviation (MAD) of the shRNAs for each cell line.

Class Definitions. Comparisons of PMAD normalized shRNA relative abundance data were based on behavior of shRNAs within a class or differential behavior of shRNAs between classes of cell lines. Class definitions used included cell line lineage (e.g., ovarian cancer, NSCLC, etc.) or genetic alterations (*KRAS*, *BRAF*, or *PIK3CA* mutation) (2, 3). Class definition files (.cls) were made using the GenePattern module "SubsetGctandCls."

Scoring shRNAs by Class Comparisons. To compute the statistical evidence that a given shRNA contributes to the observed essentiality phenotype between two classes of interest, we used a weight of evidence (WoE) approach. This approach computes the likelihood that a given shRNA has the ability to discriminate between the two classes of interest in a statistically significant manner. Weights of evidence scores for a particular class comparison, as defined by a class definition file, were calculated using the GenePattern module "ScorebyClassComp." The probability that any given shRNA can provide this discrimination is inferred from its posterior log-odds ratio:

$$\mathrm{Ev}(r|x) = \log \frac{P(r = \mathrm{ClassA}|x = X_i)/P(r = \mathrm{ClassB}|x = X_i)}{P(r = \mathrm{ClassA})/P(r = \mathrm{ClassB})}, \quad [1]$$

where r is a binary variable and is either ClassA or ClassB, x is a single shRNA measurement, and X_i is the shRNA level score for that shRNA.

The total evidence that the shRNA level scores provide can be computed as the average absolute evidence (AvEv):

$$\mathrm{AvEv}(\mathrm{r}|x) = \sum_{i}^{k} P(x = X_i)|\mathrm{Ev}(r|x = X_i)|, \quad [2]$$

where the sum is over all of the k distinct shRNAscores X_i.

To compute the conditional probabilities, we used a logistic regression model because the set of X_i shRNA level measurements is a continuous distribution. The logistic regression model that approximates the conditional probability is:

$$P(r|x) = \frac{1}{1 + e^{-(A+Bx)}}. \quad [3]$$

A generalized linear model fit identified the values of the coefficients A and B so as to be able to compute the conditional probabilities in each cell line in each class. Because our primary focus was to identify shRNAs that were depleted in abundance, we ranked AvEv scores by their effects from most negative to most positive. Therefore, we used a signed AvEv from the value of coefficient B (sign(B)AvEv), preferentially ranking shRNAs from the most negative to the most positive WoE. In this manner, we identified shRNAs with the most discriminatory power among two classes as well as the shRNAs that were depleted in the particular class of interest (e.g., *KRAS* mutant). One advantage of this approach is that it does not assume that the shRNAscores are normally distributed within each class, an assumption that is central to other metrics of differential assessment including t tests and signal-to-noise ratios.

Data Files and GenePattern Modules. A portal with data files, accessory files, and GenePattern modules for reproducing the analysis to produce shRNAscores can be found on the Integrative Genomics portal (http://www.broadinstitute.org/IGP). Ranked shRNA and gene lists for all of the analyses presented in this paper can also be found there.

Subsampling Analysis. A group of 124 shRNAs, including control shRNAs and shRNAs targeting *KRAS*, *BRAF*, *PIK3CA*, and *PAX8,* as well as other genes, were used for an analysis of class size. WoE comparisons of *KRAS* mutant vs. WT, *BRAF* mutant vs. WT, *PIK3CA* mutant vs. WT, and ovarian vs. non-ovarian lines were performed for these shRNAs. Every comparison was performed on 100 random subsets of cell lines taken from each class, for a range of equal class sizes (1 vs. 1, 2 vs. 2, etc.) from 1 to number of cell lines screened in each target class (1–26 for *KRAS*, 1–10 for *BRAF*, 1–12 for *PIK3CA*, and 1–25 for Ovarian). Percentile of shRNA rank (shRNA rank divided by the total number of tested shRNAs, multiplied by 100) for shRNAs specific for *KRAS* (TRCN0000033262), *BRAF* (TRCN0000006291), *PIK3CA* (TRCN0000039607), and *PAX8* (see *Plasmids*) from the *KRAS*, *BRAF*, *PIK3CA*, and ovarian subsampled comparisons, respectively, were plotted as grouped boxplots by target class size.

Collapsing shRNAScores to Gene Rankings. The GENE-E program (http://www.broadinstitute.org/cancer/software/GENE-E) (1) was used to collapse shRNA differential essentiality scores to gene rankings by three complementary methods. These methods included (*i*) ranking genes by their highest shRNA depletion score, (*ii*) ranking genes based on the P value rank of their second best ranked shRNA, and (*iii*) ranking genes using a KS statistic in an approach similar to gene set enrichment analysis (RNAi gene enrichment ranking) for scoring genes based on the P value rank of the normalized enrichment scores (NES; ref. 1). The NES represents the bias of the set of shRNAs targeting each gene toward the phenotype of interest, for example, depletion in *KRAS* mutant lines.

The majority of the 11,194 genes were represented by 5 shRNAs (range 2–31 shRNAs per gene, excluding control shRNAs). Out of the initial 54,020 shRNAs in the pool, 979 shRNAs were excluded from the gene rankings because they contained overlapping sequence (offset of less than 3 base pairs) with another shRNA construct for the same gene. Nine additional shRNAs, representing 9 genes, were removed automatically by GENE-E, before gene ranking analysis. Control shRNAs target GFP, RFP, Luciferase, and LacZ, and each control shRNA is represented as 5 replicate measurements on the microarray.

To assess the significance of a gene score obtained by the second best or KS scoring methods described, P values were computed based on 10,000 random samplings of shRNAs to create artificial genes with the same number of shRNAs as the gene of interest (correcting for different set sizes of shRNA targeting different genes). The P value reflects the number of times such an artificially constructed gene received a score as

good as or better than the gene of interest. Therefore, the smaller the *P* value, the less likely such a gene score could have been obtained at random.

On average, 58% of the shRNA suppress the given target >70% using qPCR measurements of endogenous transcript levels (The RNAi Consortium); thus, a simple average of shRNAscores is not ideal because not all shRNAs are effective. Because the single shRNA and second best shRNA methods depend only on the 1–2 shRNAs of strongest effect, the influence of ineffective shRNAs on gene scores is minimized. The KS statistic however considers all shRNAs from each gene in producing a gene score. It is thus more sensitive to cases for example in which all five shRNAs score moderately for depletion. Because a higher false positive rate with the single shRNA ranking method is predicted due to off-target effects, only the top 150 genes identified by this method were selected for further analysis, whereas the top 300 genes from each of the other two methods were selected. A union was taken of the genes identified by these three methods.

Competition Assay. Of the 350 shRNA retested, 238 shRNAs were selected to represent a range of fold depletion in OVCAR-8 and OVCAR-4 cells, including shRNAs ranking from #1–19, #101–120, #501–525, #1,001–1,025, #5,001–5,025, #10,001–10,025, and #20,001–20,020). In addition, 112 shRNAs targeting 25 oncogenes or control genes were included. The 350 shRNAs are listed in Table S2. OVCAR-8 (5×10^4) cells were seeded into each well of a 96-well plate and spin-infected with 2 or 4 μL of lentiviruses (in duplicate) at 930 × *g* for 2 h at 30 °C in the presence of 4 μg/mL polybrene to transduce ∼50% of the cells. Cells were then trypsinized and replated into 24-well plates. The percent of GFP+ cells at 3 and 7 d postinfection was measured using BD LSR II flow cytometry system equipped with a high-throughput sampler (BD Biosciences). The fraction of GFP+ cells 7 d postinfection relative to 3 d postinfection was calculated. Data represent mean ± SD of duplicate infections.

Analysis of Primary Tumor Data. Regions of copy number amplification identified by Genomic Identification of Significant Targets in Cancer analyses were used from publications focused on various tumor lineages, including ovarian (4), NSCLC/lung adenocarcinoma (5), glioblastoma (6), colorectal, and esophageal squamous cancers (7). When necessary, coordinates were changed to hg18. Regions in the colon and esophageal squamous lineages were manually reviewed for segmentation artifacts; potential artifacts were removed. For all lineages, all RefSeq genes within the regions of amplification were identified and cross referenced with genes interrogated in the screening library. All primary high-grade serous ovarian cancer data were downloaded from the TCGA portal (http://tcga-data.nci.nih.gov/tcga). The frequency of amplification for *PAX8* genes was determined by using a threshold of $\log_2$ copy number ratio > 0.3 within a subset of tumors in TCGA project (345 tumors). Screenshots of the same tumor data were taken using the Integrative Genome Viewer (http://www.broadinstitute.org/igv).

Differential Expression Analysis. Expression analyses were performed on cell lines with gene expression data available (n = 83; http://www.broadinstitute.org/ccle). For every lineage with more than 6 lines with available expression data, Comparative Marker Selection was performed in GenePattern. The top 200 differentially overexpressed genes for each lineage compared with all other lineages were identified using a SNR. Significance testing of shRNAscores between high and low *PAX8* expressing lines was done with a t test (n = 83, mean *PAX8* expression dividing high and low classes).

Plasmids. To generate a plasmid coexpressing shRNA and GFP, a GFP cDNA fragment was cloned into the BamHI and KpnI sites of pLKO.1-puro-shRNA to replace the puromycin resistance gene. A pool of 85 control shRNAs targeting reporter genes (GFP, RFP, Luciferase, and LacZ) was used to generate control lentiviruses (Control shRNAs) (1). The sequences targeted by *PAX8*-specific shRNAs are as follows:

TRCN0000021274 (shPAX8#3: 5′-CCTTCGCCATAAAGCAGGAAA-3′),
TRCN0000021275 (shPAX8#5: 5′-GCAACCATTCAACCTCCCTAT-3′),
TRCN0000021276 (shPAX8#4: 5′-CTCTTTATCTAGCTCCGCCTT-3′),
TRCN0000021277 (shPAX8#2: 5′-CCCAGTGTCAGCTCCATTAAT-3′)
and TRCN0000021278 (shPAX8#1: 5′-CCGACTAAGCATTGACTCACA-3′).

Cell Proliferation Assay. Cells were seeded into each well of 96-well plates (Costar) 24 h before infection. Six replicate infections were performed for control shRNAs and each *PAX8*-specific shRNA in the presence of 4 μg/mL polybrene for 24 h. After the incubation, medium was replaced with fresh medium with triplicates containing 2 μg/mL puromycin, and cells were cultured for 5 d. The ATP content was measured using CellTiter-Glo luminescent cell viability assay (Promega). Data represent mean + SD of six replicate infections relative to infection with control shRNAs.

Immunoblotting. Cell lysates were prepared by scraping cells in lysis buffer [50 mM Tris HCl (pH 8), 150 mM NaCl, 1% Nonidet P-40, 0.5% sodium deoxycholate, and 0.1% SDS] containing 1× Complete protease inhibitors (Roche) and phosphatase inhibitors (10 mM sodium fluoride and 5 mM sodium orthovanadate). Protein concentration was measured using BCA Protein Assay kit (Pierce). An equal amount of protein (30 μg) was separated by NuPAGE Novex Bis-Tris 4–12% gradient gels (Invitrogen) and then transferred onto a poly(vinylidene difluoride) membrane (Amersham) using a Bio-Rad electrophoretic tank blotting apparatus. The membrane was then incubated with primary antibody for 1 h at room temperature. Antibody against PAX8 (sc-81353) was purchased from Santa Cruz Biotechnology. Antibody against poly(ADP-ribose) polymerase (#9532) was purchased from Cell Signaling Technology. After incubation with the appropriate horseradish peroxidase-linked secondary antibodies (Bio-Rad), signals were visualized by enhanced chemiluminescence plus Western blotting detection reagents (Amersham). β-actin was also assessed as an internal loading control by using a specific antibody (sc-8432-HRP, Santa Cruz).

1. Luo B, et al. (2008) Highly parallel identification of essential genes in cancer cells. *Proc Natl Acad Sci USA* 105:20380–20385.
2. Thomas RK, et al. (2007) High-throughput oncogene mutation profiling in human cancer. *Nat Genet* 39:347–351.
3. Forbes SA, et al. (2008) The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet* 57:10.11.1–10.11.26.
4. TCGA-Network (2011) Integrated genomic analyses of ovarian carcinoma. *Nature*, in press.
5. Weir BA, et al. (2007) Characterizing the cancer genome in lung adenocarcinoma. *Nature* 450:893–898.
6. Cancer Genome Atlas Research Network (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455:1061–1068.
7. Beroukhim R, et al. (2010) The landscape of somatic copy-number alteration across human cancers. *Nature* 463:899–905.

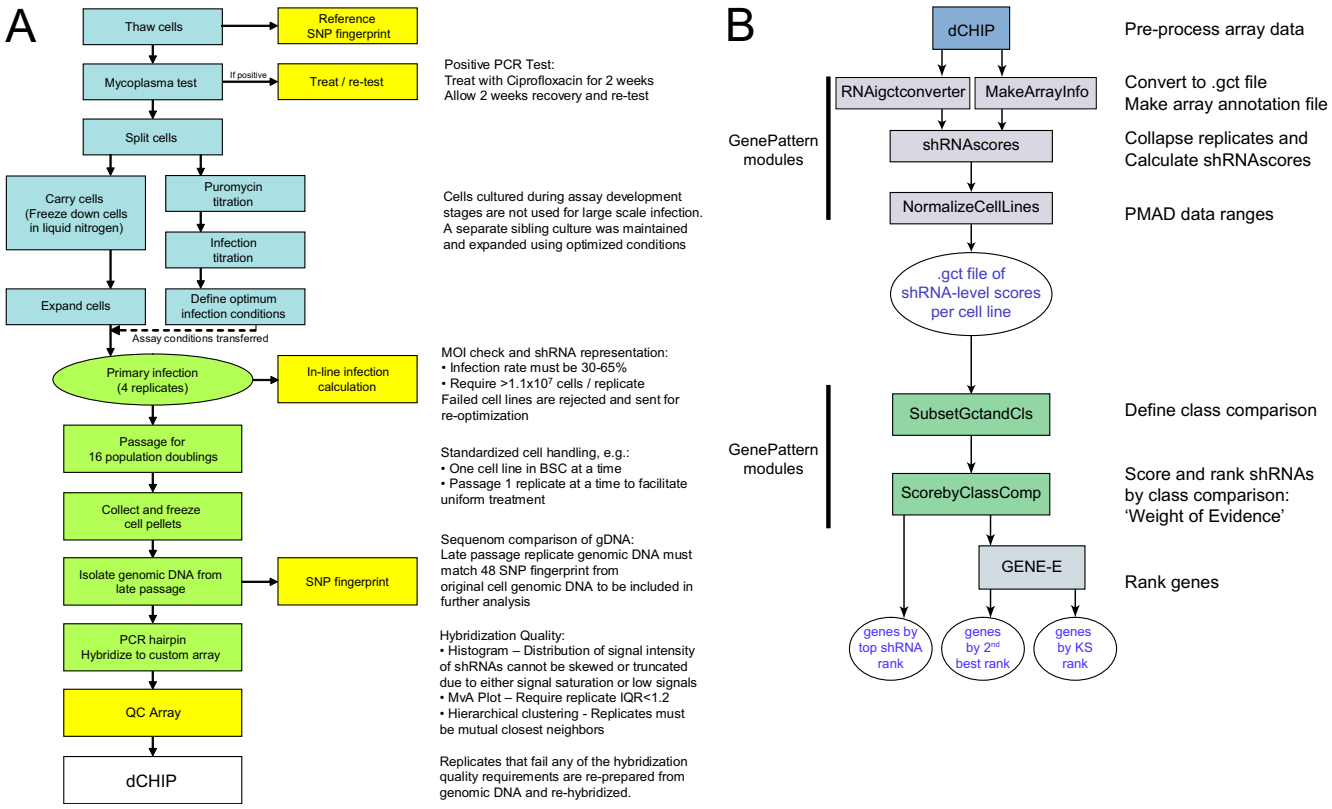

Fig. S1. Overviews of pooled screening and analytical pipelines. (*A*) We developed a streamlined, standardized process for cell line assay development (blue boxes), infection (green oval), and passaging (green boxes) to minimize variations in handling and culture conditions. Several quality control steps were implemented at different stages within the process (yellow boxes). All cell lines have a known reference 48 SNP Sequenom genotype before initiating the screen. Thawed cells were tested for the presence of mycoplasma by PCR before screening. Mycoplasma contaminated cell lines were cultured in the presence of 10 μg/mL ciprofloxacin for 2 weeks followed by 2 weeks of culture in standard growth medium. Cell lines that passed the mycoplasma PCR retest were allowed to reenter the screening pipeline. The cells used for shRNA lentiviral pool infections are parallel "sibling" cultures of those cells used for assay development. Puromycin sensitivity was determined by treating infected and uninfected cells with puromycin doses ranging from 0 to 10 μg/mL. Infection titration was performed over a range of 0–400 μL of virus per well of a 12-well plate using the same protocol as a large-scale infection (see *SI Methods* for details). After large-scale infection, an in-line measurement of infection rate was calculated by dividing the number of viable cells after puromycin selection over number of viable cells without puromycin selection. Infection rates between 30–65% were deemed acceptable for screening, and cell lines with infection rates outside this range were reoptimized. Cells were passaged for 16 population doublings or 28 d (whichever was longer) using a standardized passaging protocol. Genomic DNA from the final cell harvest was isolated, and cell line identity was confirmed by SNP genotyping and comparison with reference genotypes. Virally integrated shRNA sequences were PCR-amplified from genomic DNA, and products were hybridized to a custom microarray to determine the representation of shRNAs. The quality of the hybridization was assessed by examining probe distribution histograms. Replicate reproducibility was determined by examining both MvA plots and hierarchical clustering dendrograms. Outlier samples with respect to hybridization intensity distribution or replicate reproducibility were reevaluated starting from genomic DNA (see *SI Methods* for details). Three or four high-quality replicates were obtained for each of 102 cell lines screened. (*B*) Analysis pipeline. A schematic showing the analytic pipeline created to process pooled RNAi screening data. Raw array data (.CEL files) were first processed with a modified version of dCHIP (1). The rest of the pipeline used GenePattern modules designed to take in the dCHIP normalized array measurements ("RNAigctconverter") and produce shRNA-level data ("shRNAscores," "NormalizeCellLines"), then calculate a WoE score for each shRNA that measures a differential effect based upon a class comparison ("SubsetGctandCls," ScorebyClassComp"). The GENE-E program was used to take the ranked differential shRNAscores between two classes and collapse to gene-level data.

1. Luo B, et al. (2008) Highly parallel identification of essential genes in cancer cells. *Proc Natl Acad Sci USA* 105:20380–20385.

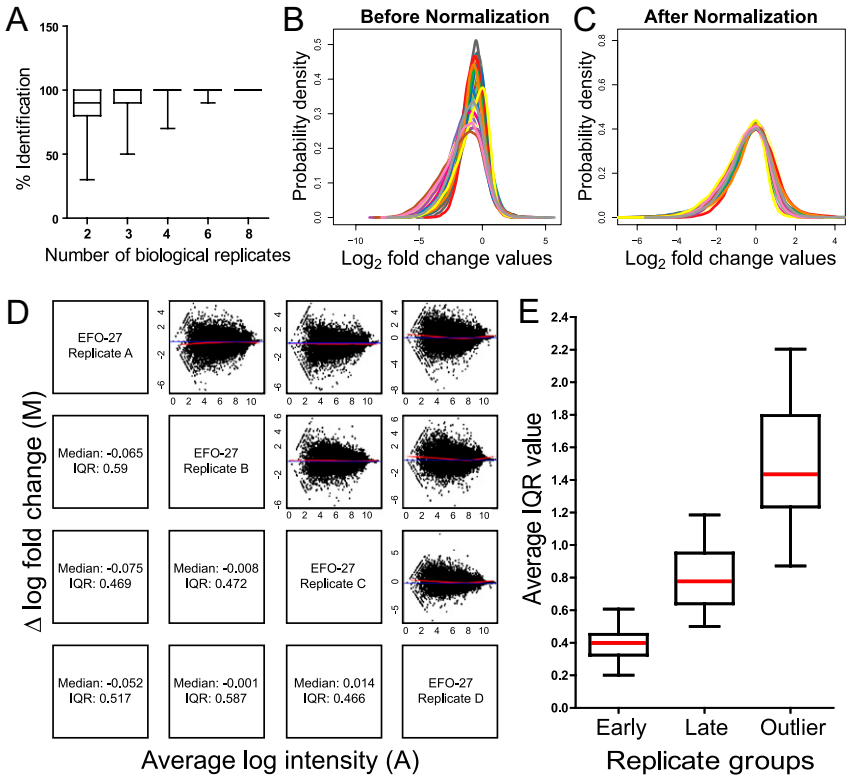

Fig. S2. Replicate number, normalization, and replicate reproducibility. (*A*) Published data (1) from 10 replicate infections of Jurkat cells were used to assess the minimum number of replicates required to generate an accurate list of shRNAscores and ranks. The log fold change of shRNA abundance in Jurkat late time point replicates relative to the initial reference plasmid DNA pool replicates were computed. The top 250 most depleted shRNAs in the 10-replicate set were identified. Randomly chosen subsets of replicates with replicate sizes of 2, 3, 4, 6, or 8 out of the 10 replicates were selected, and analysis was performed to determine the frequency at which the top 250 shRNAs from the 10-replicate set appeared within the top 1,000 ranked shRNAs in the smaller replicate set. The percent identification was averaged across the 10 subsampled datasets for each replicate size, where 100% identification indicates an ideal list identical to the list of shRNAs obtained in the 10-replicate set. The boxes represent the 25th to 75th percentile of the data, and whiskers extend to the extremes. The 4-replicate set was observed to accurately identify these top scoring constructs at high frequency. (*B* and *C*) Peak median absolute deviation (PMAD) normalization. The probability density (*y* axis) was plotted for the adjusted $\log_2$ fold change scores (*x* axis) of each cell line (colored by line) before (*B*) and after (*C*) PMAD normalization. PMAD normalization was performed by subtracting the value of each shRNA from the modeled peak value of the distribution of each cell line and dividing by the median absolute deviation of each line. (*D* and *E*) Replicate reproducibility. (*D*) MvA plots for four unnormalized replicates of EFO-27. For each pair of replicates, the difference between replicate values for $\log_2$ fold change of signal (*y* axis) is plotted against the average of $\log_2$ signal for those two replicates (*x* axis) (these plots shown in matrix positions above the diagonal). In addition, median and interquartile range (IQR) for the interreplicate differences in $\log_2$ fold change signal values are reported for each pairwise comparison (in corresponding matrix positions below the diagonal). Values for both IQR and median close to zero represent tightly clustered arrays. (*E*) The observed range of cell-line averaged IQR values across cell lines are displayed for early time point replicates (5 d postinfection), late time point replicates, and a generated set of artificial "outlier" replicates. The outlier IQR values were generated by combining three cell line replicates with a mismatched replicate from a different cell line. These artificial four-replicate sets thus model the expected distribution of IQR values in the case that one of the four chip replicates is a dramatic outlier. The red line in each box-plot is the median value for the group; boxes represent the 25th to 75th percentile of the data, and whiskers span the extremes.

1. Luo B, et al. (2008) Highly parallel identification of essential genes in cancer cells. *Proc Natl Acad Sci USA* 105:20380–20385.

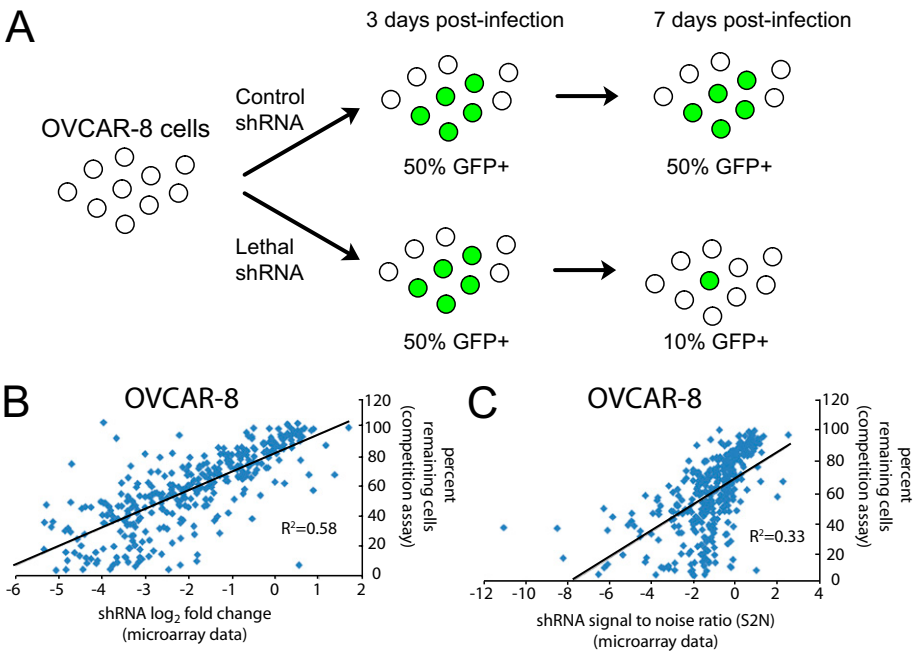

Fig. S3. Evaluation of different shRNA pooled screen scoring methods against individual shRNA proliferation tests. (*A*) Experimental schematic. OVCAR-8 cells were infected with each of the 350 shRNAs to transduce ~50% of the cells. Percentage of GFP+ cells 3 and 7 d postinfection was measured by FACS. (*B* and *C*) The relative abundance of OVCAR-8 cells infected with 350 individual shRNAs encoded in a GFP+ plasmid (*y* axis, relative to 3 d post infection) measured at 7 d post infection are plotted against the relative abundance of each shRNA in the pooled shRNA screen as quantified by different two different functions of the microarray hybridization data. Correlation plots are shown for $\log_2$ fold change (R^2 = 0.58) (*B*) and signal-to-noise ratio (R^2 = 0.33) (*C*). Based on these results, $\log_2$ fold change was selected as the basis for a shRNA scoring method for all subsequent analyses.

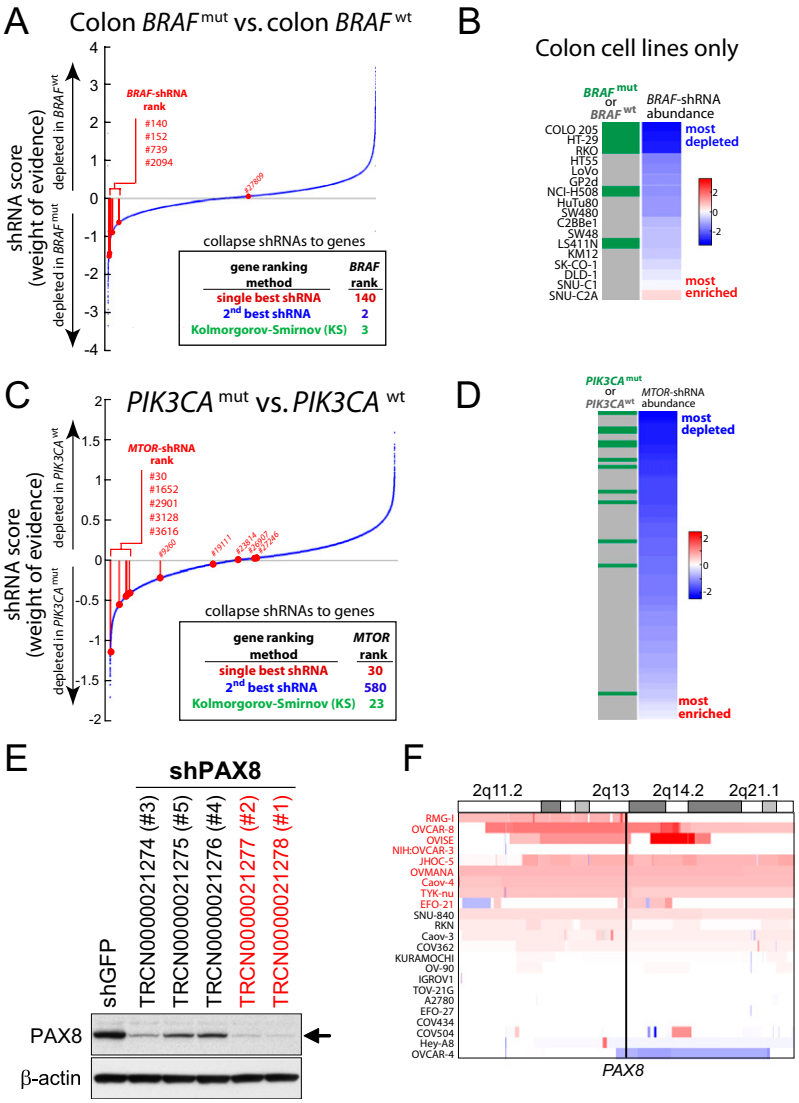

Fig. S4. Identification of essential genes in *BRAF* mutant, *PIK3CA* mutant, or 2q13-amplified cancer cell lines. (*A*) Identification of essential genes in *BRAF* mutant colon cancer cells. Distribution of shRNA ranks (*x* axis) by the WoE scores (*y* axis) for the class comparison of 5 *BRAF* mutant vs. 10 *BRAF* wild-type colon cancer cell lines only. shRNAs targeting *BRAF* are marked in red and their ranks are listed. *Inset* reports the gene ranks of *BRAF* for preferential proliferation-essentiality in the subset of cell lines with activating mutations in *BRAF*. (*B*) *BRAF*-shRNA depletion values correlate with *BRAF* mutation. Heatmap shows the fold depletion of a *BRAF*-shRNA (TRCN0000006291) in individual cell lines, sorted from most to least depleted. Mutation status is indicated in the top bar; mutant lines are in green, wild-type lines in gray. (*C* and *D*) Dependence on *MTOR* in *PIK3CA* mutant cancer cell lines. (*C*) Distribution of shRNA ranks (*x* axis) by the WoE scores (*y* axis) for the class comparisons of *PIK3CA* mutant vs. *PIK3CA* wild-type cell lines. shRNAs targeting *MTOR* are marked in red and their ranks are listed. *Inset* reports the gene rank of *MTOR* for preferential proliferation-essentiality in the subset of cell lines with activating mutations of *PIK3CA*. (*D*) *MTOR*-specific shRNA depletion values correlate with *PIK3CA* mutation status. Heatmap shows the fold depletion of the top-scoring *MTOR*-specific shRNA (TRCN0000038677) in individual cell lines, sorted from the most to least depleted. Mutation status of *PIK3CA* is indicated in the left bar; mutant lines are in green, wild-type lines in gray. (*E*) Validation of target gene suppression by *PAX8*-specific shRNAs. Immunoblot confirmed target gene suppression by top-scoring *PAX8*-specific shRNAs. OVCAR-4 cells were infected with a control shRNA targeting GFP or *PAX8*-targeting shRNAs, and cell lysates were collected 4 d after infection for immunoblotting. Two effective shRNAs, labeled in red, were further tested for their proliferation effects in a panel of ovarian cancer cell lines in Fig. 4. (*F*) Amplification of *PAX8* (2q13) in ovarian cancer cell lines. SNP array colorgram depicts genomic amplification of *PAX8*. Regions of genomic amplification and deletion are denoted in red and blue, respectively. Black vertical lines denote the boundaries of *PAX8* gene. Ovarian cancer cell lines are labeled in red if they harbor amplification of *PAX8* ($\log_2$ copy number ratio > 0.3).

Other Supporting Information Files

Table S1 (DOC)
Table S2 (DOC)
Table S3 (DOC)
Table S4 (DOC)
Table S5 (DOC)