Supporting Online Material for

Profiling Essential Genes in Human Mammary Cells by Multiplex RNAi Screening

Jose M. Silva, Krista Marran, Joel Parker, Javier Silva, Michael Golding, Michael R. Schlabach, Stephen J. Elledge, Gregory J. Hannon,* Kenneth Chang

*To whom correspondence should be addressed. E-mail: hannon@cshl.edu

Published 1 February 2008, *Science* **319**, 617 (2008)
DOI: 10.1126/science.1149185

This PDF file includes:

Materials and Methods

Figs. S1 to S4

Tables S1, S2, and S4

References

Other Supporting Online Material for this manuscript includes the following:

(available at www.sciencemag.org/cgi/content/full/319/5863/617/DC1)

Table S3 and Tables S5 to S8

SUPPORTING ONLINE TEXT

MATERIALS AND METHODS

RNAi libraries

Plasmids expressing shRNAmirs were individually expanded in bacteria by 20 hours of growth in 96 deep well plates (1 ml) in GS-96 media (Qbiogene # 3101-132). The clones were pooled and plasmids were purified. The individual shRNAs shown in Figures are: FF was FF-1309 and was previously described (*8*); ANAPC2.1 (V2HS_64696); ANAPC2.2 (V2HS_64697); ANAPC2.3 (V2HSS_ 64693); ANAPC2.1 (V2HSS_64695); ANAPC4.1 (V2HS_64630); ANAPC4.2 (V2HS_64631); ANAPC4..3 (V2HS_64632); ANAPC4.4 (V2HS_254660); ANAPC4.5 (V2HS_64629); MAD2L1.1 (V2HS_134179); MAD2L1.2 (V2HS_235168); CDC-5L.1 (V2HS_112872); DKC-1.1 (V2HS_113450); DKC-1.2 (V2HS_113449); CENPE.1 (V2HS_150547); CENPE.2 (V2HS_150546); CCNT2.1 (V2HS_202729); CCNT2.2 (V2HS_202367); CCNT2.3 (V2HS_203220); CDK9.1 (V2HS_112920); PTPRF.1 (V2HS_238791); PTPRF.2 (V2HS_170983); PTPN9.1 (V2HS_170937); PTPN9.2 (V2HS_170936); APOBEC3G.1 (V2HS_80856); PPP2R3B.1 (V2HS_71705); POLR2B.1 (V2HS_131396); POLR2B.2 (V2HS_131397). The sequences for all shRNAs can be obtained from http://codex.cshl.edu or http://www.openbiosystems.com.

Infection of the cell lines

Three parallel infections were done for each cell line to generate a total of 6.5 million infected cells per pool. Time zero (T=0; Fig. 1a) was defined following puromycin selection of infected cells and represented approximately day 4 following the initial infection. Cells were passaged without allowing confluence and time point. T=1 and T=2 were taken 1 week and 2 weeks following T=0.

Viral infection

Three parallel infections were done for each cell line to generate a total of 6.5 million infected cells per pool for the 6K pool and 10 and 20 million cells respectively for the 10K and 20K pools. Time zero (T=0; Fig. 1a) was defined following puromycin selection of infected cells and represented approximately day 4 following the initial infection. Cells were passaged without allowing confluence and time points T=1 and T=2 were taken 1 week and 2 weeks following T=0.

DNA preparation

Genomic DNA was extracted from cells sampled at each time-point by resuspending cells in lysis buffer (1% SDS; 50mM TRIS-HCl, pH 8.0; 100nM EDTA; 100mM NaCl). Lysed cells were incubated with Proteinase K at 55°C until lysates clarified and then treated with RNase A for 2-3 hrs. at 37°C. Samples were phenol/chloroform extracted 3 times, and DNAs were precipitated and resuspended in 10mM TRIS-HCL, pH 7.5.

Microarray hybridization and detection

For some experiments (initial 6K screens and probe validation), we used high-density oligonucleotide multiplex arrays (12-plex) (NimbleGen Systems Inc.) with around 13,000 probes synthesized per subarray. Probe were all 50mers. We employed a 2-color hybridization strategy where the sample was labeled with Cy3 and the reference was labeled with Cy5. A hybridization mix containing 180ng of each fluorescently labeled sample, biotin-CPK6 control oligo at 4ug/ml (Nimblegen Systems Inc.), Herring Sperm DNA at 80ug/ml, BSA at 0.4mg/ml, and 1X MES hybridization buffer (1X MES/ 0.89M NaCl/20mM EDTA/0.01% Tween20) was set up for each hybridization. Hybridizations were carried out using 12-plex slide chambers (Nimblegen Systems Inc.) at 44°C for 18-19 hrs. Slides were washed in NSWB (6X SSPE/0.01% Tween20) at room temperature and then in SWB (1X MES/25mM NaCl/0.01% Tween20) at 44°C and then washed again in SWB at room temperature. Slides were finally washed in NSWB before CPK6 oligo detection. Slides were incubated in 1X stain solution (1X MES/0.89M NaCl/0.1% Tween20)/BSA at 2mg/ml/ Cy3- and Cy5-streptavidin at 10ug/ml for 30-45 mins at room temperature. After which slides were quickly dunked in 2X SSPE. Slides were spun dried and scanned using Axon 4000B scanner. Feature extraction was performed using Nimblescan.

For other studies, custom agilent arrays with an 8-plex format were used at a probe density of 15,000 per subarray. Sample amplification and labeling were carried out in the same manner as with Nimblegen arrays. However 1.5ug of each labeled sample (per color) was used for hybridization. We used standard Agilent-recommended hybridization protocols, reagents, and equipment for microarray processing (Agilent Technologies Inc.). Slides were scanned using an Axon 4000B scanner and feature extraction performed using the Agilent FE software.

Microarray analysis

Data generated (Nimblescan software) from image scans (Axon 4000B Scanner) were imported into R version 2.4 for processing and analysis. Each array consists of 12,636 features, divided into 5742 barcode and 6887 half hairpin probes. Intensity distributions for barcode probes demonstrated increased intensities overall relative to the half hairpin probes. This probe type effect could bias normalization, thus barcode and half hairpin probes were separated for further processing. Array background measurements were estimated from 529 probes that were not specific to any hairpins introduced into the screen. These negative control spots were showed lower intensities than the experimental probes on each array (data not shown). The median of these values was taken to be the background estimate. Signal was utilized from 4248 barcode and 5499 half hairpin probes corresponding to hairpins introduced as described previously. These signal probes were compared to the background median on each array for T=0 samples. Probes were removed from all further analyses if Cy3 values were not greater than 1.5 times the background median in two of three T=0 samples (3937 barcode and 4542 half hairpin probes remained). The two channels for each array were then normalized using loess normalization. The final pre-processing step was to collapse probes that correspond to identical sequences by taking the mean. This led to 3369 barcode

and 3881 half hairpin probes in the processed data. These two probe types were combined for further analyses.

Data analysis for pooled shRNA screens

We devised a rigorous set of criteria and procedures for analyzing data from pooled shRNA screens. These analyses were performed using the Bioconductor (*1*) and Extraction of Differential Gene Expression (*2*) libraries in R. The data from hairpin and barcode probes were analyzed separately because barcode probes showed higher median fluorescent intensity than hairpin probes. Surprisingly, barcode probes also had a higher median background fluorescent intensity. This difference in hybridization potential between hairpin and barcode probes may be attributable to their difference in design. For a fuller discussion of hairpin probes, see the accompanying paper by Schlabach and colleagues (accompanying paper).

Analysis began with the removal of probes that fail to hybridize. Failures were defined as those that do not produce signal greater than 1.5 times the raw median intensity of control probes (features that do not correspond to any sequences within the sample hybridized to the array) in at least 2 out of 3 T=0 replicates. Typically ~90% of barcode and ~80% of hairpin probes were retained after filtering, indicating that a significant proportion of shRNAs could be screened using these parameters (**Fig. S2**). All remaining probes were then loess normalized (*3, 4*) to adjust global intensity differences prior to calculation of log ratios (cy3/cy5) for each array. While some instability of clones from some early generations of shRNA libraries have been observed in previous studies, we find that this is not the case with our second generation reagents since more than 90% of all of the library shRNAs can be detected by hybridization. Extensive re-sequencing and analysis of clones within the collection over its more than one year of distribution have also supported the stability of the plasmids in this platform (http://www.openbiosystems.com/collateral/rnai/pi/pSM2retroviralshRNAmir.pdf).

Correlations between biological replicates were calculated (*5*) for each time each point and principal component analyses (PCA) (*6*) were performed for experimental variables - cell-type, time-points, and biological replicates - to investigate the sources of variation in the data. As expected, the major principal components of variation for these straight-lethal screens were time-point and cell-type, while sample treatments (e.g. different replicates or hybridization on different days) were not associated with primary axes of variation (**Fig. S3**). Although correlations between experimental replicates at each time point were high, these diminished with time, while correlations between the reference channel remained unchanged (**Table S1**). This likely reflected increasing random variations as populations were stressed and drifted upon propagation in culture. However, strong positive correlations between experimental samples were still evident at T=2.

Lastly, candidate biologically significant signals were identified using the time course algorithm described by Storey et al. (2005) (*7*) to select probes reflecting

differences in relative abundance of shRNAs over time. We used the test termed, "within class temporal differential expression", which assumes independent samples of each feature across time and condition. This methodology performs a spline fit to the data for each probe and tests whether the fit is significantly better than that of a flat line (null hypothesis). q-values are calculated and used for identifying significant features.

As initial screens were of limited complexity, we could use both a half-hairpin and a barcode for detecting many of the shRNAs screened. We found that the correlation between these two assay modes was quite high (r = 0.87) (**Fig. S4**). Overall, a gene was scored as being a potential candidate if either its barcode or shRNA probe was found to fulfill our selection criteria, which was set at a threshold of greater than 2 fold change (FC) with false discovery rate (FDR) <10%. In order to uncover cell line-specific genetic sensitivities, initial comparisons focused on the 6K screens were done with MCF-10A and MDA-MB-435. Filtering for shRNAs that had a low false discovery rate (q<0.1) and at least 2-fold depletion in MCF-10A but no more than 1.2 fold depletion in MDA-MB-435. Although any cut-off is relatively arbitrary, we chose this as a conservative criterion to minimize false positives. The choice is validated empirically by examination of the validation studies that we subsequently performed.

In order to confirm that the results obtained by the microarray platform, five of the MCF10A time points, from which we had enough DNA material, were analyzed using a single molecule sequencing assay (*5, 8, 9*). In total, more than 30 million reads were sequenced. Solexa results showed that: 1- as previously seen in the arrays, the correlation between time points was highly consistent (r=0.94 at T=0 and r=0.91 at T=1); 2- comparisons between Solexa and microarray results for the same samples ranged from 0.44 to 0.70 indicating general agreement between the platforms. However, the reduced correlation between arrays and sequencing, despite maintaining correlations within replicates at different time points, does highlight the fact that each method may have intrinsic biases that will shape the subset of candidates that it reveals. Only very extensive validation studies will allow a determination of which platform shows better predictive power.

Cell culture

Tetracycline-inducible cell lines were generated by infecting MCF-10A and MDA-MB-435 with a retroviral vector expressing the transcriptional transactivator (tTA) (clontech) at high MOI (>70% infection). After selection with G418, resistant cells were used to generate individual tet-inducible knock-down cells. These cells were infected at high MOI with virus generated using the Tet-inducible vector (pTM, the backbone of which is identical to that of pSM2c, our second-generation library vector [7]) and antibiotic-resistant cells were selected with puromycin. Tet-inducible knocked down cells were maintained with doxycycline (50 ng/ul) to prevent shRNA expression. To test the effect of shRNAs on cell viability, dox-containing media was removed and replaced with fresh media (changing media on three consecutive days). Cells were grown in 24 well plates in triplicate and split every 3-4 days. During the

second and third passages (day 7 and 11) cell number was estimated by cell titer-glo (Promega). This experiment was repeated at least twice for each shRNA and the results of all experiments were averaged. Cell viability was calculated as the ratio of cell number in the candidate shRNA-expressing cells to the cell number of control shRNA (FF) cells.

To study the effects of the chemical inhibition of selected targets on the cell lines, each was cultured for 72 hours in the presence of increasing concentrations of the specific inhibitors (5,6-Dichlorobenzimidazole riboside (DRB) for CDK9; 5- aza for DNMTs ; Tarceva for EGF receptor and MG132 for the proteasome. Cell numbers were estimated with cell titer-glo, and the effect on cell viability was calculated as the ratio of treated to untreated cells. Experiments were done in 96 well plates with 8 replicas and repeated at least twice for each drug. The final results represent the average of all the repeats.

QRT-PCR

Total RNAs were extracted using RNAeasy (Qiagen) kit and reverse transcription using First Strand synthesis kit (Invitrogen) was employed to produce templates for SYBR-green QPCR analysis. Target genes were amplified using specific primers and expression levels were normalized to that of GAPDH.

Cluster Analysis

Cluster analysis was performed to visualize the relationship between shRNAs that showed changes in representation in each cell line. For each cell line, we required a hairpin to show a 2 fold depletion in T2 relative to T0 and a q-value < 0.1 from the time course analysis. This resulted in a total of 2186 hairpins (union of 352 from 231 cells, 1966 from 10A cells, 201 from T47 cells, 5 from ZR cells and 232 from 435 cells). Ratios were calculated for each cell line representing the average abundance at T2 relative to the average abundance at T0. This set was clustered using Pearson correlation and average linkage clustering for the hairpins and samples and visualized with TreeView (*12*).

Solexa

The half-hairpin/barcode fragment that was used for DNA microarray analysis was amplified using proofreading DNA polymerase for each time point. PCRs were purified by phenol/chloroform extraction and then ethanol precipitation followed by resuspension in nuclease-free water. The guide strand was amplified from these purified templates (500-750ng of each sample template) using solexa-adapted primers (p7+Loop : 5' - CAA GCA GAA GAC GGC ATA CGA TAG TGA AGC CAC AGA TGT A - 3' and p5+mir3 : 5' - AAT GAT ACG GCG ACC ACC GA C TAA AGT AGC CCC TTG AAT TC - 3') and Accuprime Pfx polymerase (Invitrogen Inc.). Amplified products were gel-purified (QIAEX II kit from Qiagen Inc.) and then resuspended to 10nM in nuclease-free water and then loaded onto an 8-lane flowcell for cluster generation. Please refer to Hodges *et al* for detailed information on the sequencing methodology (*9*). The following primer was used to sequence the

antisense strand for 35 cycles to generate 35nt reads: 5' - TAG CCC CTT GAA TTC CGA GGC AGT AGG CA - 3'.

SUPPLEMENTARY ONLINE FIGURES

Silva et al Suppl. figure 1

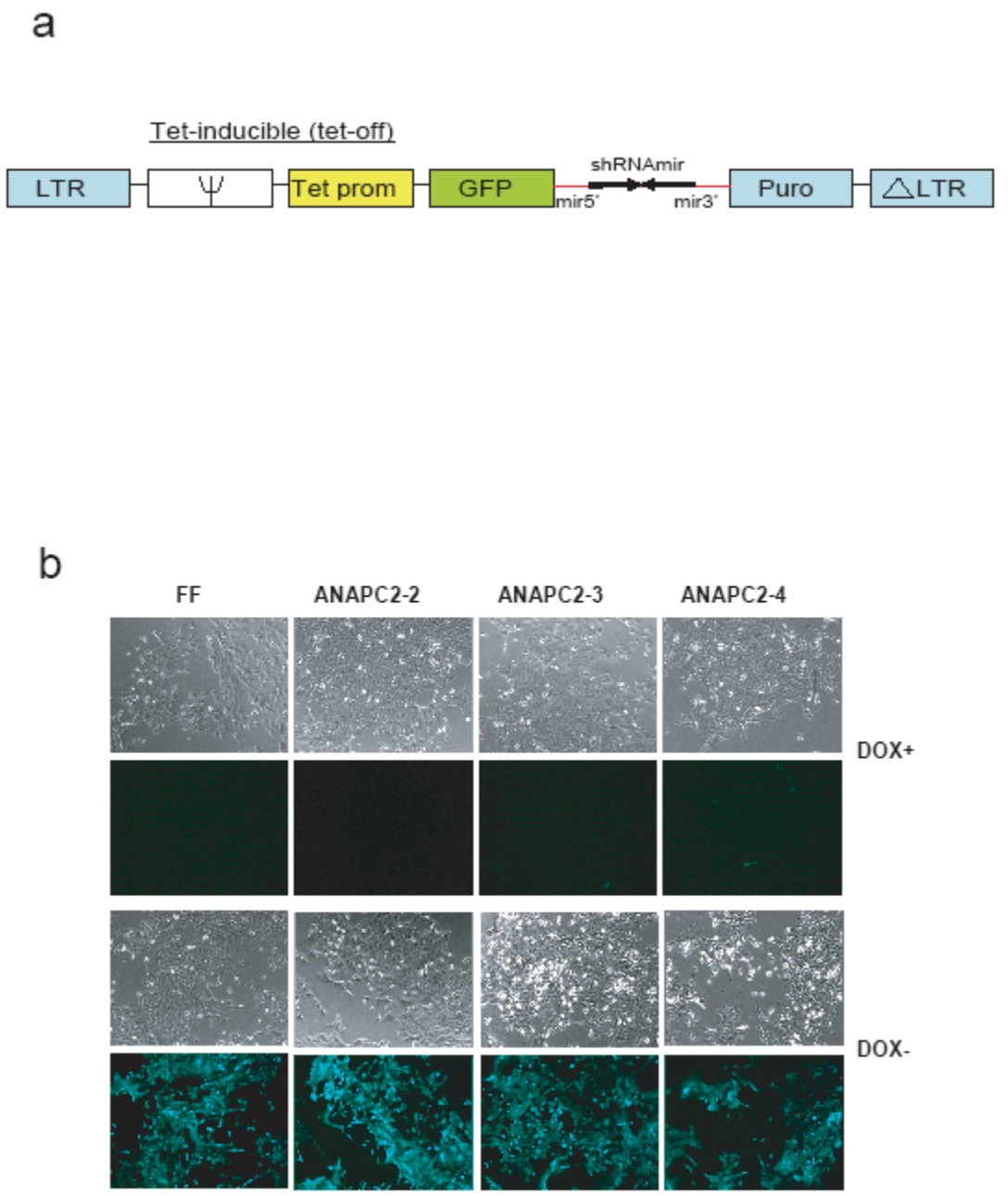

Figure S1. shRNAmir Tet-off inducible system. A, Tetracycline-inducible vector: LTR = MSCV long terminal repeat; ΔLTR = sinLTR; ψ = packaging signal; Tet-prom = tetracycline (tet) responsive promoter; GFP = green fluorescent protein; mir5'-shRNA-mir3' = miR30-based shRNA; PURO = PGK promoter-driven puromycin-resistance gene (PAC) expression. Regulated expression of shRNA can be monitored by level of GFP. **B.** MCF-10A cell line expressing the tTA transactivator was generated to create a tet-off inducible system. In the presence of doxycycline (dox), tTA cannot bind to the tet-responsive element (TRE) thereby preventing transcription of GFP and shRNA. Depletion of dox restores gene-silencing. Three different shRNAs targeting the anaphase-promoting complex (ANAPC2.2, ANAPC2.3, ANAPC2.4) and negative control shRNA targeting firefly luciferase (FF) are stably and conditionally expressed.

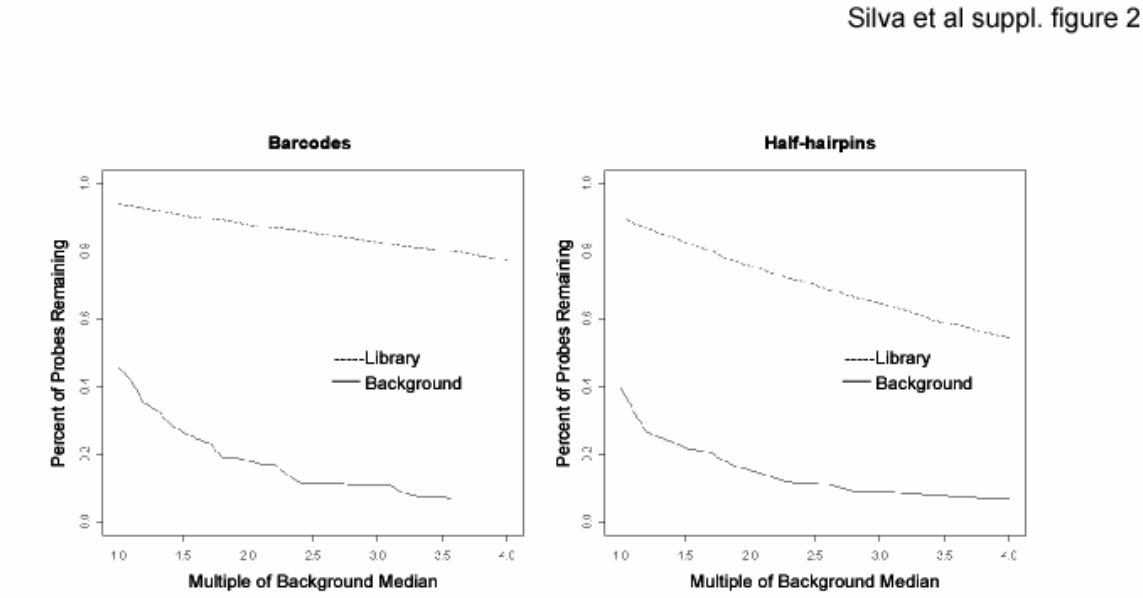

Figure S2. Probe performance. Percent of barcode (left) or half-hairpin (right) probes remaining at various multiples of median background intensity for the MCF-10A 6K experiment. Probes that do not correspond to any barcode or shRNA in the library (solid line) have significantly lower intensity at high thresholds than probes that should hybridize to library sample (dotted line). Typically, 80% of the barcode spots are still retained even when the cutoff is increased to 4 times the median of negative controls compared with 60% for the half-hairpin features.

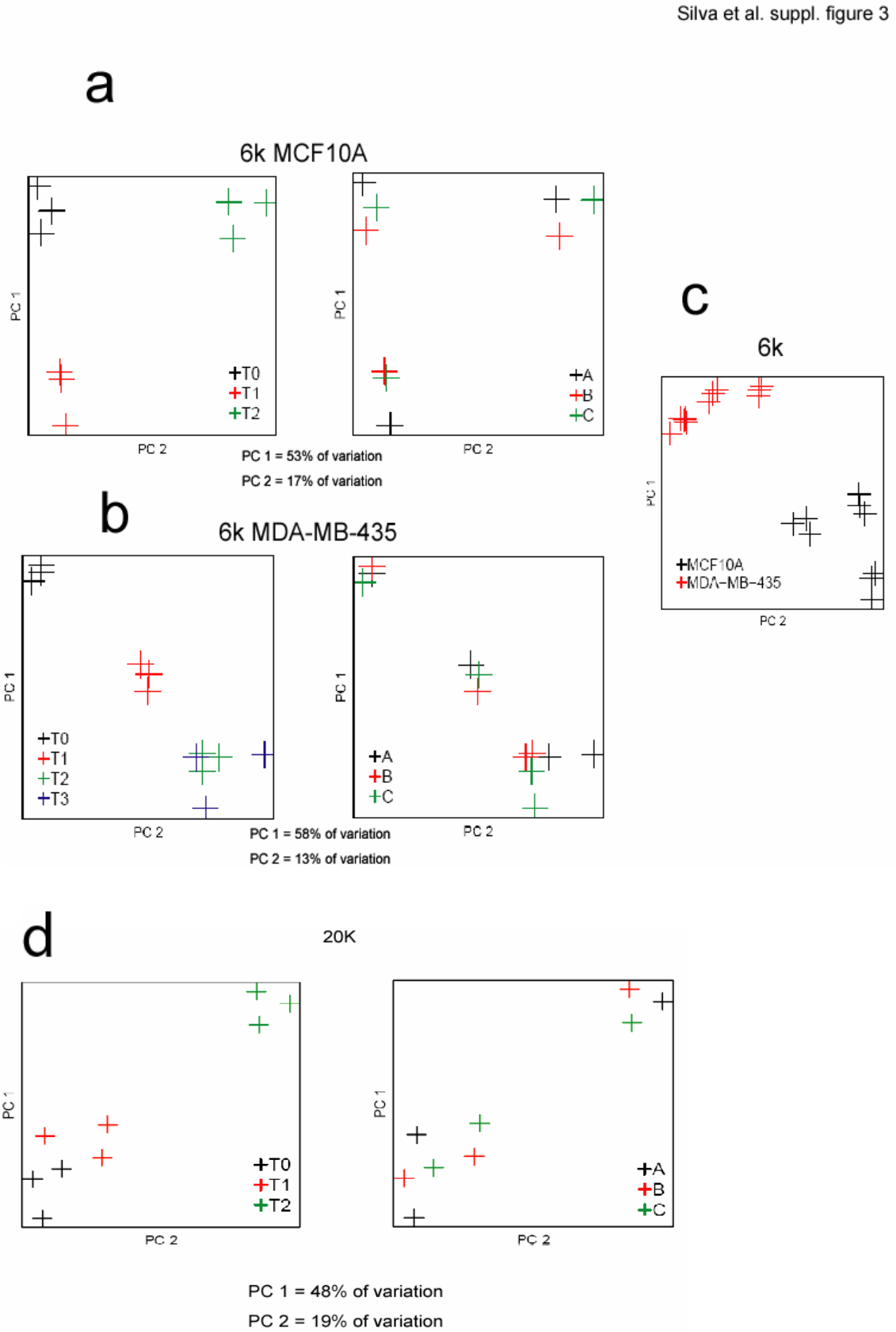
Silva et al. suppl. figure 3
a
6k MCF10A
PC 1
PC 2
+T0
+T1
+T2
+A
+B
+C
PC 1 = 53% of variation
PC 2 = 17% of variation
b
6k MDA-MB-435
+T3
PC 1 = 58% of variation
PC 2 = 13% of variation
c
6k
+MCF10A
+MDA-MB-435
d
20K
PC 1 = 48% of variation
PC 2 = 19% of variation

Figure S3. Representative principal component analysis (PCA). Variables are: time (T0, T1, T2, T3); cell population replicates (A, B, C) ; and cell-type. MDA-MB-435 6K screen was carried for nature of variable in 6K screen
an additional week (T3). (**A**) Shows the study for the MCF-10A 6K experiment. (**B**) Shows the study for the MDA-MB-435 6K experiment and panel (**C**) shows segregation of these cell lines as independent entities when analyzed by PCA. **(D)** This panel shows the PCA of the MCF-10A 20K screen for time (T0,T1,T2) and pool replicates (A,B,C).

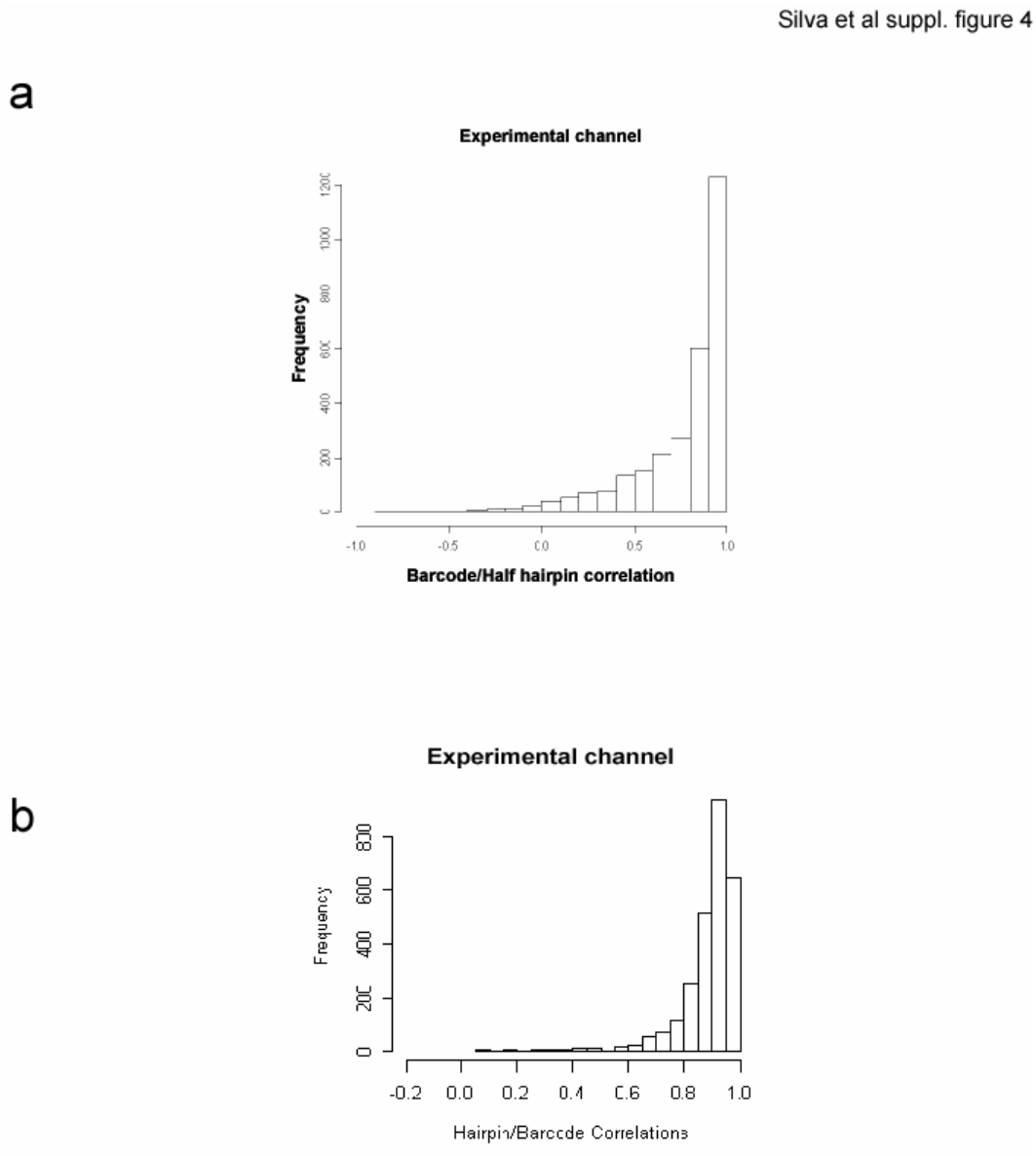

Figure S4. Representative correlation between the barcodes and half hairpins. The figure shows the correlation for 6K MCF10A screen (**A**) and for the 20K MCF10A screen (**B**).

SUPPLEMENTARY TABLES

Table S1. Correlation among replica experiments.		
	T=0 replica-A	T=0 replica-B
T=0 replica-B T=0 replica-C	.84 (.8) .83 (.8)	.87 (.82)
	T=1 replica-A	T=1 replica-B
T=1 replica-B T=1 replica-C	.66 (.61) .61 (.6)	.67 (.65)
	T=2 replica-A	T=2 replica-B
T=2 replica-B T=2 replica-C	.51 (.38) .53 (.42)	.5 (.42)

The numbers show the average of the correlation between the same probes of different replica experiments. First number represents correlation for barcodes and the number inside the brackets represents the correlations for half-hairpins.

Table S2. Resume of 6K screen results in MCF-10A and MDB-MB-435				
	Over 2 Fold change	# of shRNAs (%total)	Range	# of genes (%total)
MCF-10A	Depleted	461 (7.5)	2-8	357 (5.8)
MCF-10A	Enriched	556 (9.1)	2-8	450 (7.3)
MDA-MB-435	Depleted	333 (5.4)	2-10	285 (4.65)
MDA-MB-435	Enriched	169 (2.75)	2-5	158 (2.6)

The table shows the number and percentage of shRNAs and genes that showed greater than 2 fold change in representation. The range represents the magnitude of the change.

Table S3: Depleted genes for the 6K screens in MCF-10A and MDA-MB-435 cells.

*BC=Barcode; HH=Half Hairpin; FC= Fold change; CAG=Cyclosome associated gene; CAG(cyclosome)=Subunit of the Cyclosome; PAG=Proteasome associated gene; SAG; SWI/SNF associated gene**.**
* Fold changes are linear scale

See accompanying Excel file

Table S4

Representative 20k Screen: within-pool correlations. 20K_1000_T0, 20K_1000_T1, 20K_1000_T2 are T0, T1, T2 timepoints. Mean r and stdev for cy3 and cy5 are the mean correlation and standard deviation for the green (experimental) and red (normalization) fluorescence channels.

Table S7 (MCF-10A 20K)	cy3 mean r	stdev	cy5 mean r	stdev
20K_1000_T0	0.974	0.001	0.981	0.001
20K_1000_T1	0.962	0.003	0.920	0.012
20K_1000_T2	0.965	0.004	0.977	0.004

Table S5: Straight lethal 6K vs 20K screens in MCF-10A cells.

*BC=Barcode; HH=Half Hairpin; FC= Fold change.
* Fold changes are linear scale

See accompanying Excel file

Table S6: Overlapping depleted genes for the 6K and 20k screens in MCF-10A cells.

*BC=Barcode; HH=Half Hairpin; FC= Fold change.
* Fold changes are linear scale

See accompanying Excel file

Table S7: Genes specifically depleted in MCF-10A or MDA-MB-435 cells.

*BC=Barcode; HH=Half Hairpin; FC= Fold change; CAG=Cyclosome associated gene; CAG(cyclosome)=Subunit of the Cyclosome; PAG=Proteasome associated gene; SAG; SWI/SNF associated gene**.**
* Fold changes are linear scale

See accompanying Excel file

Table S8: Comparison of essential genes across cell lines (MCF-10A, MDA-MB-435, MDA-MB-231, T47D, ZR-75.1

*BC=Barcode; HH=Half Hairpin; FC= Fold change.
* Fold changes are log2 base

See accompanying Excel file

References and Notes

1. R. C. Gentleman *et al.*, *Genome Biol* **5**, R80 (2004).
2. J. T. Leek, E. Monsen, A. R. Dabney, J. D. Storey, *Bioinformatics* **22**, 507 (Feb 15, 2006).
3. Y. H. Yang *et al.*, *Nucleic Acids Res* **30**, e15 (Feb 15, 2002).
4. Y. H. Yang, Paquet, A., Dudoit,S., R package version 1.12.0. (2006).
5. Materials and Methods are available as supporting material on *Science* online.
6. F. Model, T. Konig, C. Piepenbrock, P. Adorjan, *Bioinformatics* **18 Suppl 1**, S155 (2002).
7. J. D. Storey, W. Xiao, J. T. Leek, R. G. Tompkins, R. W. Davis, *Proc Natl Acad Sci U S A* **102**, 12837 (Sep 6, 2005).
8. A. Barski *et al.*, *Cell* **129**, 823 (May 18, 2007).
9. E. Hodges *et al.*, *Nat Genet* (Nov 4, 2007).
10. J. M. Silva *et al.*, *Nat Genet* **37**, 1281 (Nov, 2005).
11. M. Schlabach.
12. M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, *Proc Natl Acad Sci U S A* **95**, 14863 (Dec 8, 1998).