

# SCIENTIFIC REPORTS



OPEN

## Identification of novel cancer therapeutic targets using a designed and pooled shRNA library screen

David Oliver<sup>1,\*</sup>, Hao Ji<sup>1,\*</sup>, Piaomu Liu<sup>2,3</sup>, Alexander Gasparian<sup>1</sup>, Ellen Gardiner<sup>1</sup>, Samuel Lee<sup>1</sup>, Adrian Zenteno<sup>1</sup>, Lillian O. Perinskaya<sup>1</sup>, Mengqian Chen<sup>1</sup>, Phillip Buckhaults<sup>1</sup>, Eugenia Broude<sup>1</sup>, Michael D. Wyatt<sup>1</sup>, Homayoun Valafar<sup>4</sup>, Edsel Peña<sup>2</sup> & Michael Shtutman<sup>1</sup>

Received: 26 August 2016  
Accepted: 17 January 2017  
Published: 22 February 2017

Targeted cancer therapeutics aim to exploit tumor-specific, genetic vulnerabilities specifically affecting neoplastic cells without similarly affecting normal cells. Here we performed sequencing-based screening of an shRNA library on a panel of cancer cells of different origins as well as normal cells. The shRNA library was designed to target a subset of genes previously identified using a whole genome screening approach. This focused shRNA library was infected into cells followed by analysis of enrichment and depletion of the shRNAs over the course of cell proliferation. We developed a bootstrap likelihood ratio test for the interpretation of the effects of multiple shRNAs over multiple cell line passages. Our analysis identified 44 genes whose depletion preferentially inhibited the growth of cancer cells. Among these genes ribosomal protein RPL35A, putative RNA helicase DDX24, and coatomer complex I (COPI) subunit ARCN1 most significantly inhibited growth of multiple cancer cell lines without affecting normal cell growth and survival. Further investigation revealed that the growth inhibition caused by DDX24 depletion is independent of p53 status underlining its value as a drug target. Overall, our study establishes a new approach for the analysis of proliferation-based shRNA selection strategies and identifies new targets for the development of cancer therapeutics.

Genetic heterogeneity of human cancers drives the need to develop a broad panel of therapeutics specifically targeting tumor cells. Therapeutic development depends on the exploration of genes and pathways critical for the growth and survival of cancer cells. The most important genes are those whose inhibition ultimately kills tumor cells while having minimal impact on normal tissues. The most direct approach for target gene identification is functional profiling with genome wide libraries of trans-dominant genetic inhibitors (TGIs). Several types of TGI libraries have been developed. These include genetic suppressor element (GSE) libraries<sup>1</sup>, which are libraries of short cDNA fragments expressing either anti-sense RNAs or inhibitory peptides<sup>2</sup>, small hairpin RNA (shRNA) libraries<sup>3</sup> which are libraries expressing the stem and loop of shRNAs which are processed by cellular enzymes to double stranded inhibitory RNAs<sup>4</sup>, and more recently, single guide RNA (sgRNA) libraries<sup>5</sup>. The application of Next-Generation Sequencing (NGS) to the screening of TGI libraries allows quantitative analysis of screening results<sup>3,6</sup>. NGS-based screening procedures are designed to discover genes based on the analysis of enrichment or depletion of TGIs in cells subjected to selection, relative to non-selected cells. The major screening strategies for cancer therapeutic target identification are screens which detect the depletion of growth-suppressing TGIs over the course of cell propagation, or conversely, those which detect the enrichment of growth-suppressing TGIs<sup>7</sup>.

Most previously published functional screens are endpoint procedures<sup>7</sup>, where the abundance of TGIs is analyzed at the initial and final points only. However, NGS-based strategies allow the analysis of multiple data points to determine the kinetics of shifting TGI abundances. The latter approach, while it provides more information

<sup>1</sup>University of South Carolina, College of Pharmacy, Department of Drug Discovery and Biomedical Sciences, Columbia, SC 29208, USA. <sup>2</sup>University of South Carolina, College of Arts and Sciences, Department of Statistics, Columbia, SC 29208, USA. <sup>3</sup>Bentley University, Waltham, MA 02452, USA. <sup>4</sup>University of South Carolina, College of Engineering and Computing, Department of Computer Sciences and Engineering, Columbia, SC 29208, USA. \*These authors contributed equally to this work. Correspondence and requests for materials should be addressed to M.S. (email: shtutmanm@sccp.sc.edu)

about biological processes, is often not used due to the difficulty of analyzing such complex data sets. Previously we performed BrdU suicide selection of a genome wide GSE library in a panel of normal and cancer cell lines to select tumor-specific target genes<sup>8</sup>. The BrdU suicide selection procedure enriches for growth-suppressing GSEs. That is, growth-suppressing GSEs provide survival advantage during this selection resulting in their over-representation at the end of the experiment. Through the initial analysis of this dataset we identified the  $\zeta$ -subunit of the COPI complex (COPZ1)<sup>8</sup>. Inhibition of COPZ1 kills a majority of tumor cell lines, while not affecting the growth of normal cells. Here we re-sequenced and analyzed this dataset to reveal a large set of potential target genes. To reduce this list to the most effective targets, we designed an shRNA library focused on the identified gene set and performed a proliferation-based selection of the library in four tumor cell lines and normal fibroblasts. Enrichment or depletion of shRNAs were determined during the course of cell propagation. The screening results were then analyzed with a bootstrap likelihood ratio test (BLRT) statistical procedure we developed. This screen and analysis procedure revealed a subset of 44 genes with additional evidence for cancer-specific growth inhibition. The dependence of cancer cells on these genes was further tested with siRNAs against individual genes. Three genes successfully showed efficacy in all three selection procedures, ribosomal protein RPL35A, DEAD-box RNA helicase DDX24, and, in addition to the previously identified  $\zeta$ -subunit, the  $\delta$ -subunit of COPI, ARCN1. We show the association of the expression of the identified genes with development and mortality of human cancers and explore the mechanism of the effect of ARCN1 and DDX24 depletion in tumor cells.

## Results

**Genetic suppressor element library screen.** Previously we constructed a library of GSEs composed of cDNA fragments from the human transcriptome (average length 135 bp) prepared from a mixture of normalized (reduced-redundancy) cDNAs from 18 cell lines derived from different types of cancers<sup>8</sup>. The library was subjected to BrdU-suicide selection (Supplemental Fig. 1) for identification of growth-suppressing GSEs in three tumor cell lines: MDA-MB-231 (breast cancer), PC3 (prostate cancer), and HT1080 (fibrosarcoma), as well as immortalized normal human foreskin fibroblasts, BJ-hTERT. The cDNA fragments were recovered and amplified from genomic DNA of BrdU selected and non-selected cells<sup>8</sup>. We performed additional sequencing of the library by 454 sequencing (Roche), obtaining approximately 100,000 GSE reads per sample. Target genes were identified by mapping fragments back to the human genome. 221 genes were targeted by at least 1 GSE enriched more than 1.5 fold in two or three tumor cell lines relative to normal cells (Supplemental Dataset S1). Gene Ontology based functional analysis showed that the set was enriched for genes involved in RNA splicing (18 genes) and translation (21 genes). Also, eight genes involved in the PI3K-AKT signaling pathway were identified using KEGG pathways analysis. The results of functional analysis are presented in Supplemental Tables S1–S4.

### Design, construction, and screening of a focused shRNA library to select growth inhibitory shRNAs.

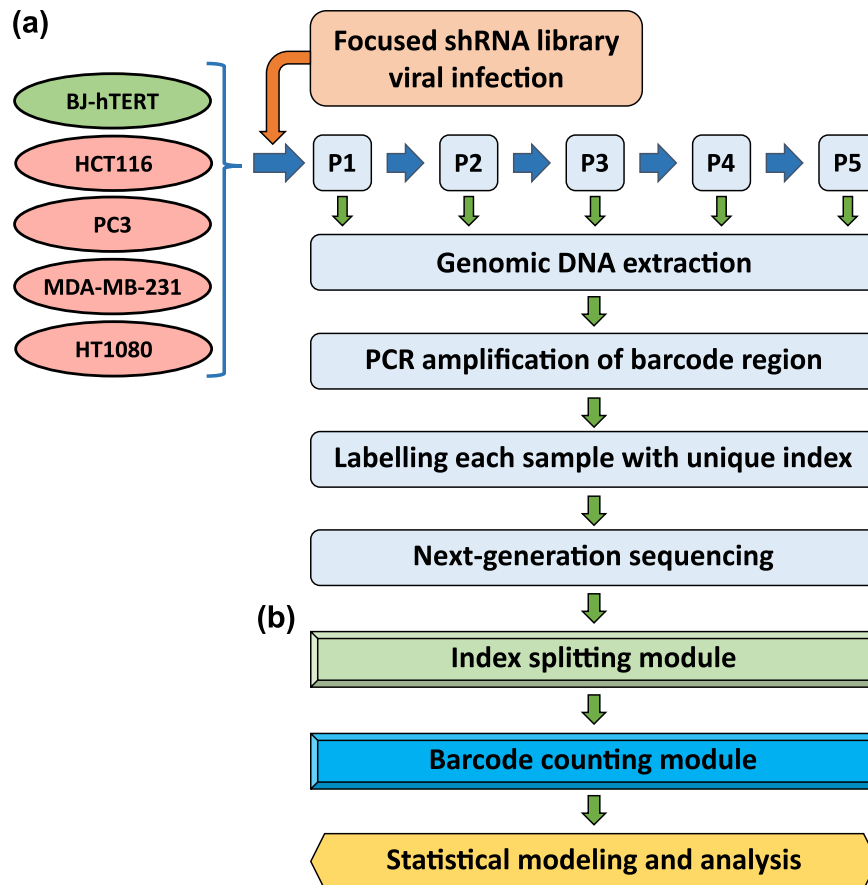
To analyze the cancer cell-specific efficacy of inhibiting the 221 genes identified in the GSE screen we designed a focused shRNA library. The main goal of the selection was to identify shRNAs depleted in proliferating tumor cell lines relative to normal fibroblasts. The shRNA library was designed to have six unique shRNAs against each of the 216 genes identified in the GSE data set. 6 unique shRNAs could not be designed for 5 of the identified genes (See Materials and Methods for details). Each shRNA was tagged with a unique barcode and an amplification adapter for rapid identification of unique shRNAs within the pool (Supplemental Fig. 2A; Supplemental Dataset S2). The shRNAs and barcodes were synthesized on-chip by parallel nucleotide synthesis, and subsequently inserted into the lentiviral vector.

Normal fibroblast (BJ-hTERT), breast cancer (MDA-MB-231), colon cancer (HCT116), prostate cancer (PC3), and fibrosarcoma (HT1080) cell lines were infected with the library-containing lentiviruses in duplicates. The library-infected cells were passaged five times to undergo approximately 20 divisions (Fig. 1a). At each passage (approximately every fourth division) genomic DNA was isolated from a portion of the cells and the barcode-containing sequences were PCR amplified from the genomic DNA. During the second round of amplification, indexes were added to each library for multiplexed sequencing (Supplemental Fig. 2B). The resulting 50 libraries were pooled and sequenced at an average depth of 3 million reads per sample. Sequenced samples were split by index and unique shRNA barcodes were counted using a bespoke algorithm. Barcode read data were then used to analyze enrichment and depletion of shRNAs over passages (Fig. 1b). The amplification and index-splitting procedure we developed to allow parallel sequencing of as many pooled libraries as needed (far more than 24 pooled libraries allowed by standard procedure of Illumina).

### Development of a bootstrapped likelihood ratio test procedure to identify genes of interest.

In order to analyze this dataset we developed a modified general linear model approach using normalized shRNA reads as the response variable and passages as a predictor (Fig. 2a(i),b). This method for analysis allows us to identify shRNAs with cancer specific efficacy by comparing them to the control BJ-hTERT cell line without the need for specific sets of positive and negative control shRNAs.

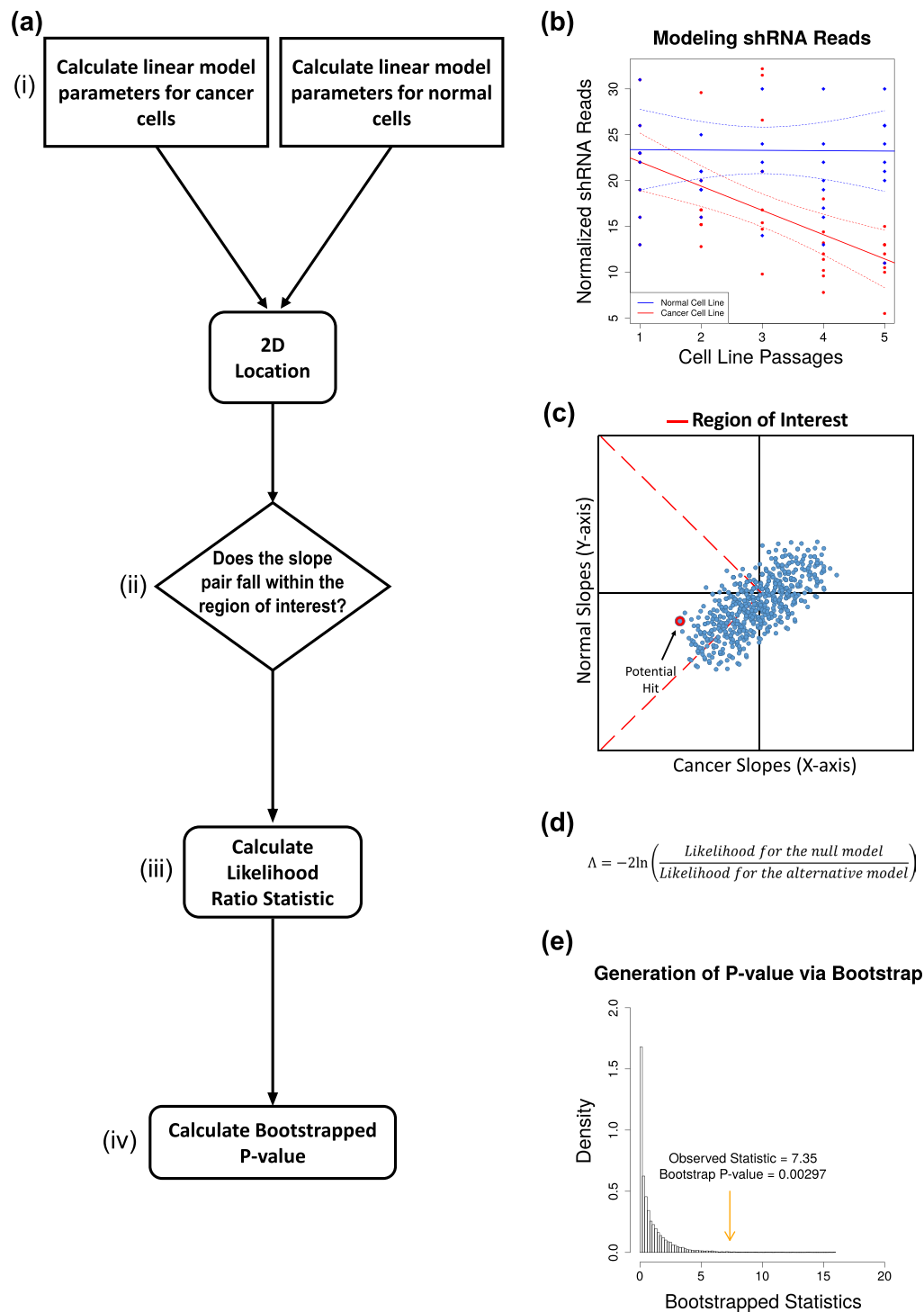
We defined a region of interest where the slope of the fitted line for the six shRNAs against a gene in the cancer cell line being queried is less than zero while the corresponding slope for the shRNAs against the same gene in the normal cell line is not more extreme than the cancer cell line (Fig. 2a(ii),c). This criterion allows for flexible normal cell line response to the shRNAs targeting the gene of interest while constraining the cancer cell line to depletion of the same shRNAs over passages. Specifically, shRNA barcode reads were normalized using upper-quantile normalization followed by Box-Cox transformation ( $\lambda \approx 0.5$ ). Since we used a general linear model, we checked that the normality assumptions were met before modeling (Supplemental Fig. 3). Once the model was fitted, we identified genes whose shRNAs affected cell growth such that the slope of the shRNA reads as a function of passage fell within the region of interest described above (Fig. 2a,b). If the estimated slopes were found to be within



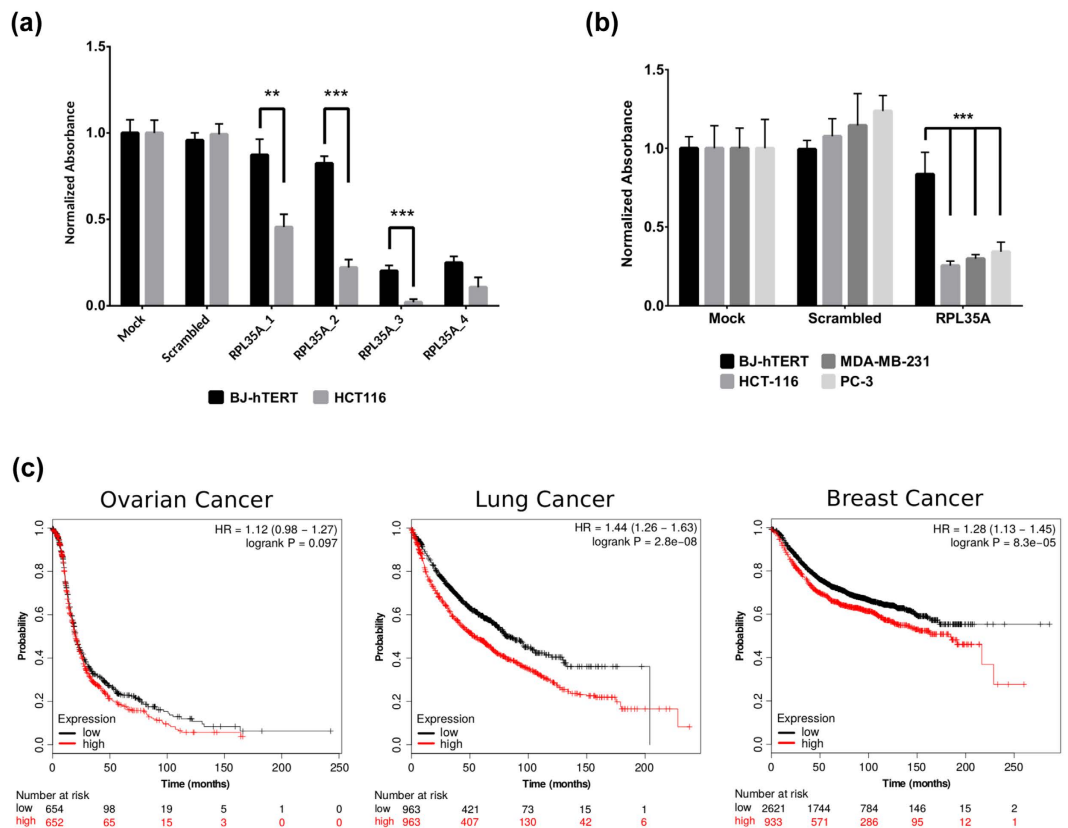
**Figure 1. Experimental design of focused shRNA library screening and sequencing.** (a) The focused shRNA library was infected into BJ-hTERT, PC3, MDA-MB-231, HCT116, and HT1080 cells. 48 hrs later cells were collected. One fourth of the cells were re-plated (P1) and the rest were used for genomic DNA extraction. The procedure was repeated for 4 passages (P2-P5). During the 3-step barcode amplification (See Materials and Methods for details) unique index-sequences were added to amplified fragments from each sample. Resulting indexed libraries were mixed together and sequenced on an Illumina HiSeq 2500. (b) Sequencing data were processed first by identifying the sample index and then by identifying the specific shRNA barcode resulting in cell line, passage number, and shRNA specific reads. These data were then submitted to statistical analysis.

the region of interest, then we calculated the likelihood ratio test (LRT) statistic for that gene (Fig. 2a(iii),d). To calculate a P-value for the gene-specific LRT statistic, we performed bootstrapping by re-sampling the residuals 10,000 times, calculating the LRT for each bootstrapped sample to determine the probability of observing a LRT statistic at least as extreme as the observed LRT statistic value (Fig. 2a(iv),e). See methods section for detailed derivation of the BLRT.

**Identification of shRNAs that inhibit the growth of tumor cells.** For the first three passages, the variability between the biological replicates was small and the correlations were strong. After passage 3, there was variability between the biological replicates that prevented any genes from passing our threshold for statistical significance ( $p < 0.05$ ) after multiple testing corrections (Supplemental Fig. 4). The observed variability is likely caused by the cumulative impact of off-target effects of shRNAs<sup>9</sup>. However, 72 genes had some evidence for selective targeting ( $-2\ln(\text{LRT}) > 0$ ) in the cancer cells compared to the normal control cell line through 5 passages. We stratified these genes by evidence in the four cancer cell lines. Genes in group 1 had a  $-2\ln(\text{LRT}) > 0$  in all four cell lines. Genes in group 2 had a  $-2\ln(\text{LRT}) > 0$  in 3 cell lines. Genes in group 3 had a  $-2\ln(\text{LRT}) > 0$  in 2 cell lines. For group 4, since the evidence is only in a single cell line, the requirements were made more stringent, by requiring the LRT statistic to be greater than 0.1. Of the original 72 genes, 44 genes fell into at least one of these four categories (Supplemental Table S5). 26 of the 44 genes have previously been reported to be involved in cancer. Of the 6 genes in group 1, 4 were previously identified as having a role in cancer. Of the 9 genes in group 2, 5 have previously been identified as cancer related genes. 19 genes were identified in group 3, of which 12 have previously been reported to have an involvement in cancer. Finally, 10 genes were identified in group 4, 5 of which have previously been reported to be involved in cancer (Supplemental Table S6). The fact that many of these genes had been previously identified as having a role in cancer supports the validity of our method to identify potential cancer therapeutic targets.



**Figure 2. Statistical analysis of shRNA depletion over multiple passages.** (a) The analysis of shRNA reads over multiple passages was performed by (i) calculating the linear regression model for the cancer cell line of interest and the normal cell line. (ii) shRNAs whose regression parameters fall within the region of interest were then identified. (iii) The likelihood ratio test statistic was calculated for shRNAs within the region of interest. (iv) A P-value for the likelihood ratio test statistic was then calculated via a bootstrap method. (b) A representation of the linear regression model calculated for all shRNAs targeting a single gene over five cell line passages for cancer (red) and normal (blue) cells. (c) The region of interest (demarcated in red) was defined by depletion of shRNA reads in cancer cell lines (x-axis) and less extreme depletion or enrichment in normal cells (y-axis). (d) The likelihood ratio test statistic for genes within the region of interest was the  $-2\ln$  of the likelihood of the regression parameters being located outside the region of interest (null model) over the likelihood that the regression parameters were within the region of interest (alternative model). (e) A P-value for the observed likelihood ratio statistic, for a given target gene, was calculated as the probability of observing a value at least as large as the observed value out of 10,000 residual re-sampling bootstraps.



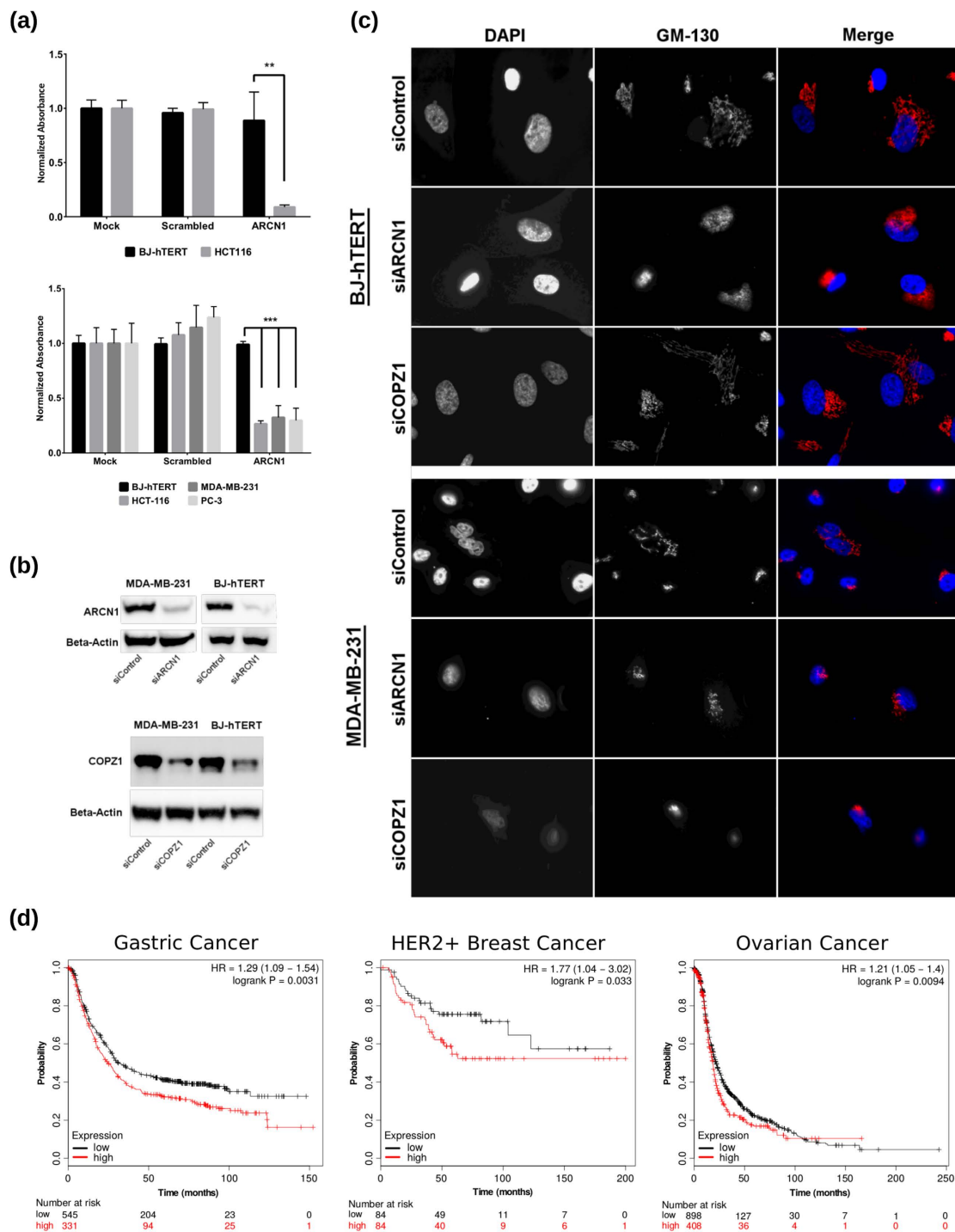
**Figure 3.** RPL35A depletion significantly reduces cancer cell growth compared to normal cells. (a) BJ-hTERT and HCT116 cells were transfected with four individual siRNAs, 5 nM each (Qiagen). The number of cells was quantified 5 days after transfection using SRB assay. Bars represent mean normalized absorbance at 540 nm  $\pm$  1 SD (P-values: 0.0035, 7.6356e-05, 0.0009, N.S.). P-values were calculated using single tailed Welch's t-test. (b) BJ-hTERT, HCT116, MDA-MB-231, and PC-3 cell lines were transfected with 2.5 nM of pooled siRNAs (GE Healthcare). Cell number determined and represented as in (a) (P-values: all <0.0001). Significance calculated as above but with additional Holm-Sidak correction for multiple testing. All siRNA transfections presented on panel A and B performed in 4 replicates. (c) Association of RPL35A expression in Affimetrix array datasets with survival of ovarian, breast, and lung cancer patients was analyzed using KM-plotter<sup>70</sup>. Plots represent survival of patients with high (red) and low (black) level of RPL35A expression.

**Secondary siRNA screens identify potential therapeutic targets.** The genes identified in the focused shRNA screen were subjected to further analysis to reveal whether short-term depletion of the gene would provide a similarly discriminatory effect between tumor and normal cells as that observed with the longer shRNA treatment. To this end, the effects of siRNA knockdown of the 44 selected genes were assessed in HCT-116, MDA-MB-231, PC-3, and BJ-hTERT cells. The cells were transfected with pooled siRNAs (GE Healthcare) in 96 well format, followed by quantification of the cell number five days after transfection (Supplemental Table S9). The siRNA depletion of 3 genes - ARCN1, DDX24, and RPL35A - significantly inhibited the growth of all three cancer cell lines but not normal cells. The growth inhibitory effects of pooled siRNAs were confirmed by the depletion of the genes with different individual siRNAs from another source (Qiagen) (Figs 3a,b, 4a and 5a). ARCN1 was identified as a group 1 gene by the analysis of shRNAs screening, and DDX24 and RPL35A were identified as group 3 genes (Supplemental Table S7).

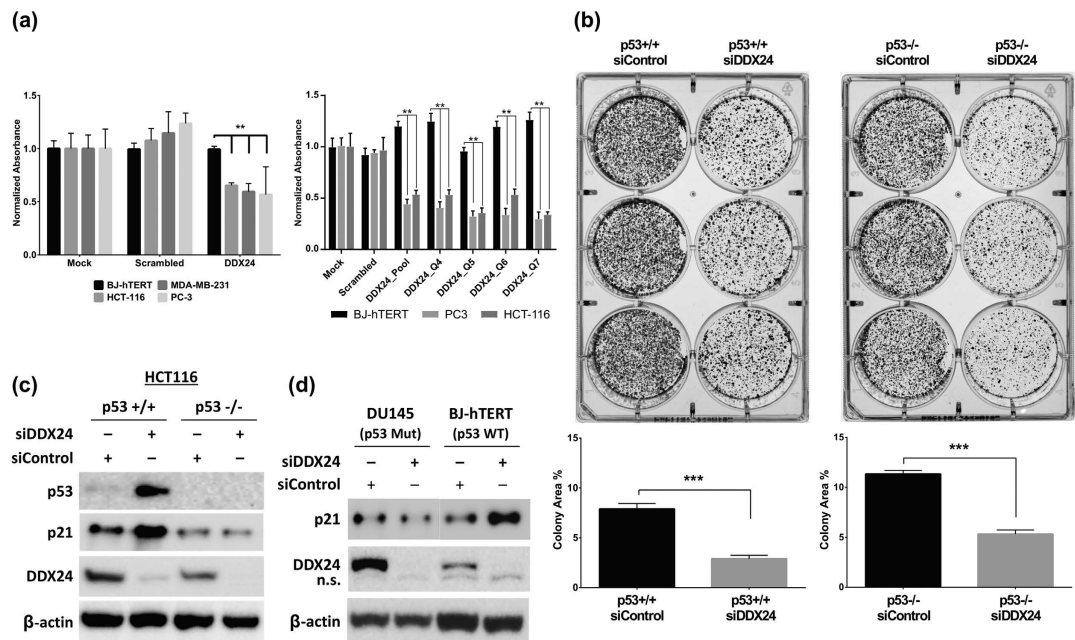
**Depletion of RPL35A preferentially inhibits growth of tumor cells and its expression correlates with poor survival of cancer patients.** RPL35A is a part of the large ribosomal subunit. The depletion of RPL35A with both pooled and four individual siRNAs preferentially inhibited growth of tumor cell lines (Fig. 3a,b). The cytotoxic effects observed in cancer and normal cells for 2 of the individual siRNAs (Qiagen) could be attributed to off-target effects. Analysis of gene expression profiles deposited in GEO shows that high expression level of RPL35A strongly correlated with poor survival of breast, ovarian, and lung cancer patients (Fig. 3c). Furthermore, in the majority of cancer datasets available in TCGA, the expression level of RPL35A mRNA is higher than in corresponding normal tissues (Supplemental Fig. 7).

**Effects of ARCN1 depletion on growth and Golgi integrity.** The ARCN1 gene encodes the  $\delta$ -COP protein which is an integral component of the COPI complex<sup>10,11</sup>. Previously we showed that depletion of another subunit of the COPI complex,  $\zeta$ 1-COP protein, caused Golgi disruption followed by cell death in the majority of tumor cell lines but not in normal cells<sup>8</sup>. To compare the mechanisms of selective growth inhibition of tumor cells





**Figure 4. ARCNI is a potential target for cancer therapy.** (a) BJ-hTERT and HCT-116 were transfected with 5 nM individual siRNA (Qiagen) (Upper Panel). BJ-hTERT, MDA-MB-231, HCT-116, and PC3 cells were transfected with 2.5 nM of pooled siRNAs (GE Healthcare) (Lower Panel). Cell number, representation, and statistical tests were done as in Fig. 3. P-values are <math><0.0063</math> and All <math><0.0001</math> for Upper and Lower Panels respectively. All siRNA transfections performed in 4 replicates. (b) ARCNI and COPZ1 expression were analyzed by WB in MDA-MB-231 and BJ-hTERT cells. Cells were transfected with control, ARCNI, or COPZ1 siRNAs (2.5 nM, GE Healthcare pooled siRNAs). Cells were lysed 72 hrs post-transfection followed by WB analysis of  $\beta$ -Actin, ARCNI, or COPZ1 proteins with corresponding antibodies. (c) BJ-hTERT and MDA-MB-231 cells were transfected with COPZ1, ARCNI, and control siRNA (as in (b)). Golgi and cell nuclei were visualized by IF with anti-GM130 and DAPI. Depletion of ARCNI results in Golgi disruption (no distinct tubular structure) in both normal (BJ-hTERT) and cancer cells (MDA-MB-231), while depletion of COPZ1 by siRNA results in selective disruption of the Golgi in cancer cells without affecting the Golgi in normal cells. (d) Association of ARCNI expression with survival of gastric, breast, and ovarian cancer patients. Kaplan–Meier analysis performed as described in Fig. 3.

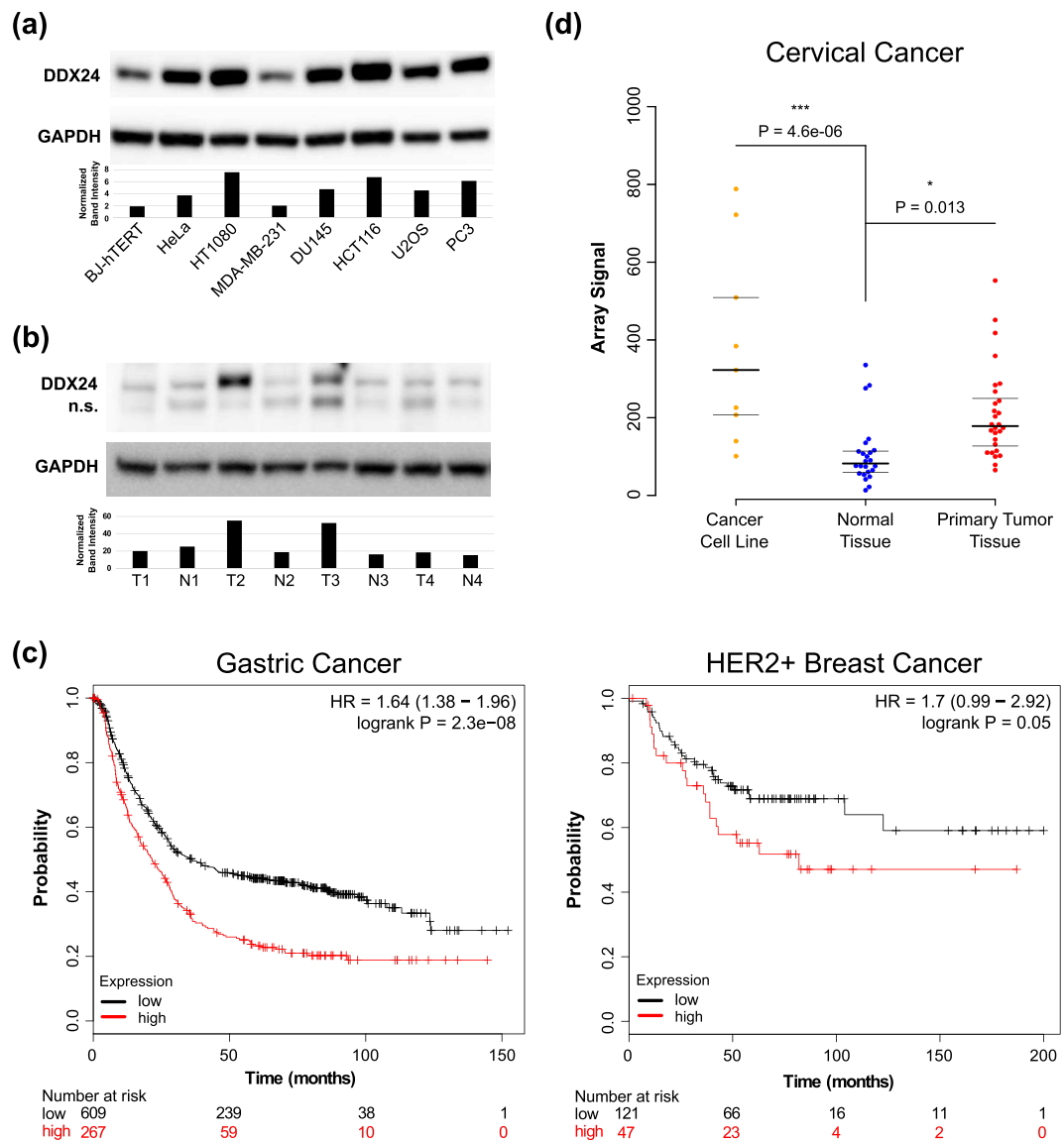


**Figure 5.** DDX24 mediated inhibition of cancer cell growth is independent of p53. ((a) Left panel) BJ-hTERT, HCT116, MDA-MB-231, and PC-3 cell lines were transfected with 2.5 nM pooled DDX24 siRNAs (GE Healthcare) in quadruplicates. Cell number, representation, and statistical tests as in Fig. 3 (P-values: 0.0026, 0.0004, and 0.0002, respectively). ((a) Right panel) BJ-hTERT, HCT116, and PC3 cells were transfected with 5 nM of pooled DDX24 siRNA (GE Healthcare) and 4 individual siRNAs (Qiagen) in 6 replicates. Cell number, representation, and statistical tests as in Fig. 3 (P-values < 0.0001). (b) HCT-116 with p53 wild-type (p53+/+) and p53 null (p53-/-) were transfected with 5 nM of control or DDX24 siRNAs and plated for colony formation assay. Cells were transfected in triplicate and plated at 1,000 cells per well followed by quantification of total colony area. Bars represent average  $\pm$  1 SD (P-values: 0.0005 and < 0.0001, respectively). (c) Western blot analysis of p53, p21, and DDX24 expression in wild-type and p53-/- HCT-116 cells. Cells were transfected with 5 nM of siRNAs and lysed 72 hrs post-transfection followed by WB with corresponding antibodies. (d) Western blot analysis of p21 and DDX24 expression in DU145 and BJ-hTERT cells. Cells were transfected with 5 nM of pooled siRNAs and lysed 72 hrs post-transfection followed by WB with corresponding antibodies.

we analyzed the results of  $\delta$ -COP depletion in MDA-MB-231 breast carcinoma cells and BJ-hTERT fibroblasts (Fig. 4b). Contrary to the results obtained from  $\zeta$ -COP depletion,  $\delta$ -COP siRNAs induced Golgi disruption in both MDA-MB-231 and immortalized fibroblasts (Fig. 4c, Supplemental Figs 5 and 6), while affecting the growth of cancer cells only (Fig. 4a). The disruption of Golgi is characterized by complete or partial disappearance of tubular structures identified by staining with Golgi-specific marker GM130 (Supplemental Fig. 6). Furthermore, the analysis of ARCN1 expression in human tumors shows that higher levels of ARCN1 expression correlates with lower survival rate of gastric, ovarian, and HER+ positive breast cancer patients (Fig. 4d). Interestingly, in the majority of cancer datasets available in TCGA the expression level of ARCN1 mRNA are not significantly changed relative to corresponding normal tissues (Supplemental Fig. 7) suggesting that the sensitivity of cancer cells to ARCN1 depletion may not be dependent on increased ARCN1 expression.

**Depletion of DDX24 inhibits growth of tumor cells regardless of p53 status.** DDX24 is a member of the DEAD-box family of putative RNA helicases<sup>12</sup> which have been implicated in diverse cellular functions<sup>13–18</sup>. However, to our knowledge, this is the first identification of DDX24 as a target for selective inhibition of tumor growth (Fig. 5a). Our initial results showed that depletion of DDX24 with pooled siRNAs (GE Healthcare) inhibits the growth of tumor cells with differing p53 status, including HCT116 with a normal level of the wild-type p53, MDA-MB-231 with mutated p53, and PC3 which are p53 null<sup>19</sup>. These results were confirmed with the set of four individual siRNAs (Qiagen) (Fig. 5a). Furthermore, siRNA-mediated DDX24 depletion in parental (p53+/+) and p53-/- HCT116 cells also inhibits the growth of both wild type and p53-/- cells with similar efficacy (Fig. 5b), while elevating p53 and p21 expression in p53+/+ cells and not p53-/- cells (Fig. 5c). Similar results were observed with DU145 cells expressing p53-223Leu and p53-274Phe mutant proteins<sup>19</sup> (Fig. 5d). Finally, DDX24 knockdown induced p21 expression in BJ-hTERT normal fibroblasts (Fig. 5d) without having significant effect on cell growth (Fig. 5a).

The level of DDX24 protein is elevated in all but one of the tumor cell lines we tested relative to BJ-hTERT fibroblasts (Fig. 6a). Moreover, DDX24 protein expression is significantly upregulated in two out of four samples of colon cancer tissues relative to normal counterparts (Fig. 6b). Analysis of expression of DDX24 in human tumors suggests that high level of the gene expression correlates with a lower survival rate of Gastric and HER2+ positive breast cancer patients (Fig. 6c). Furthermore, the level of DDX24 expression is elevated in cervical cancer



**Figure 6. DDX24 expression is upregulated in cancer and correlates with survival.** (a) Western blot analysis of DDX24 expression in BJ-hTERT, HeLa, HT1080, MDA-MB-231, DU145, HCT116, U2OS, and PC3 cell lines. Bars represent normalized DDX24/GAPDH intensity. (b) Western blot analysis of DDX24 expression in 4 pairs of matched tumor and normal colon tissue. Bars represent normalized DDX24/GAPDH intensity. (c) Association of DDX24 expression with survival of gastric and HER2 positive breast cancer patients. The Kaplan-Meier analysis was performed as described in Fig. 3. (d) The expression level of DDX24 in cervical cancer cell lines, primary tumors, and benign tissues<sup>75</sup> (microarray data deposited in GEO database). P-values, calculated using single tailed Welch's t-test with Holm-Sidak correction for multiple testing, are indicated for significant differences between the groups.

tissues and cell lines, relative to normal tissues (Fig. 6d). In the majority of cancer datasets available in TCGA the expression level of DDX24 mRNA is not significantly changed relative to corresponding normal tissues (Supplemental Fig. 7). The conflicting RNA-seq (TCGA) and Western blotting results for DDX24 suggests a post-transcriptional regulation of DDX24 expression in cancer cell lines and tumor tissues.

## Discussion

We have developed a screening and analytical discovery protocol to identify novel genes required for growth and survival of tumor cells but not normal cells. The procedure employs (i) screening of a full genome GSE library, (ii) validation of the selection results with a designed shRNA library, and (iii) additional validation with gene-specific siRNAs. All screens were performed on a panel of tumor cell lines of different origin (prostate, breast, colon, and connective tissues) and normal fibroblasts. The negative selection pressure against elements of the focused shRNA library was ascertained through NGS-based monitoring of shRNA-specific reads over the course of multiple passages. To statistically analyze the effects of several shRNAs (6 per gene) over 5 passages per cell line in biological replicates, we developed a procedure utilizing a linear regression approach. During the



development of this paper, an alternative approach that adapts the edgeR software<sup>20</sup> for differential gene expression analysis to perform this type of analysis was reported<sup>21</sup>.

Re-sequencing and analysis of previously performed GSE library selection<sup>8</sup> identified 221 genes targeted by growth inhibitory GSEs. Genes were selected for follow-up if their GSEs were enriched more than 1.5 fold in two or more tumor cell lines but not in normal fibroblasts. Follow-up validation with a focused shRNA library confirmed 44 genes; 26 of these had been previously linked to cancer development and progression (Supplemental Table S6). The selective dependency of tumor cells on the expression of the genes was further tested with individual siRNAs. The experimental procedure for siRNA-based analysis consisted of a short, 5-day treatment which was strikingly different than the shRNA library selection (5 passages over 16 to 25 days). It is therefore not surprising that depletion of only three genes – ribosomal protein RPL35A, COPI subunit ARCN1, and RNA helicase DDX24 – selectively inhibited the growth of all three cancer cell lines without strong effects on BJ-hTERT cells. The shorter duration of siRNA depletion provides additional strength to the importance of these validated targets.

The RPL35A gene encodes a 110 aa protein that functions as a part of the large ribosomal subunit, which is composed of 47 proteins. The RPL35A protein is not a part of the ribosome core structural proteins conserved along all taxons<sup>22</sup>; however, it is found in archaea (RPL35Ae) and all eukaryotes from vertebrates to yeast (RPL33 in *Saccharomyces Cerevisiae*)<sup>23</sup>. The yeast RPL35A homologue is involved in rRNA processing, 60S subunit assembly, and 60S to 40S subunit joining in the final step of translational initiation<sup>24</sup>. The depletion of RPL33 in yeast drastically inhibited protein biosynthesis and growth, while the phenotype could be partially reversed by overexpression of tRNA-Met<sup>24</sup>. The depletion of RPL35A in zebrafish embryos by egg-injection with morpholino antisense oligos resulted in abnormal phenotypes during embryo development<sup>25</sup>. Haploinsufficiency and mutations of RPL35A are associated with Diamond-Blackfan anemia, an autosomal dominant bone marrow failure syndrome<sup>26–28</sup>. The example of Diamond-Blackfan anemia, where mutations of different ribosomal proteins specifically affect the growth of highly proliferative erythroid progenitor cells, is a natural model of selective sensitivity of highly proliferating tumor cells to RPL35A depletion. Elevated RPL35A expression was found in glioma cells relative to normal astrocytes<sup>29</sup>, while RPL35A overexpression in Jurkat cells (an acute T-cell leukemia derived cell line) produced a cell death resistant phenotype<sup>30</sup>. The resistance of Jurkat cells overexpressing RPL35A was not accompanied by changes in the expression levels of Bcl-2 or Bcl-xL, suggesting a more complex mechanism for resistance than simple alterations in Bcl-dependent apoptotic signaling. Moreover, RPL35A was shown to cooperate with c-Myc in the transformation of pre-B cells<sup>31</sup>. shRNA-mediated depletion of RPL35A inhibits growth and promotes apoptosis of leukemic cell lines<sup>26</sup>, which is similar to our results. The molecular mechanism of differential sensitivity of tumor and normal cells is yet to be identified; however, the restrictive effect of RPL35A depletion in zebrafish embryo as well as haploinsufficiency in humans described above suggests the existence of a protective mechanism. In general, tumor cells are more sensitive to impairment of ribosomal biogenesis and associated stress, which has been suggested as an exploitable dependency for the development of new cancer treatments<sup>32</sup>. These results combined with our identification of RPL35A as a potent growth inhibitor in three independent cancer cell lines and the correlation of the poor survival of ovarian, breast, and lung cancer patients with the high expression of RPL35A, suggest that RPL35A is a key factor for survival of cancer cells, while normal cells are better protected from the effects of RPL35A depletion.

ARCN1 encodes the  $\delta$ -subunit of the COPI complex. The COPI complex is composed of seven subunits with two of them,  $\zeta$  and  $\gamma$ , encoded by two highly homologous isoforms each. The complex is one of three major vesicles coating protein complexes, Clathrin, COPII, and COPI. COPI-coated vesicles shuttle cargo within the Golgi organelle and from the Golgi back to the ER<sup>33</sup> and are involved in autophagosome formation<sup>34</sup>. The disruption of COPI by inhibition of any of the unique subunits results in Golgi collapse and block of autophagy<sup>35,36</sup>. Interestingly, the deficiency of different COPI subunits translates to different phenotypes. Impairment of COPA leads to hereditary autoimmune-mediated lung disease and arthritis<sup>37</sup>, while ARCN1 mutations cause craniofacial abnormalities<sup>38</sup>. The mechanism of selective effects of inhibition of different subunits of the COPI complex needs to be identified. Previously we showed the mechanism of selective sensitivity of tumor cells to the inhibition of the  $\zeta$ -subunit of the COPI complex, encoded by two homologous genes, COPZ1 and COPZ2. The depletion of COPZ1 specifically kills tumor cells, in which COPZ2 is downregulated, while normal cells are much less sensitive to COPZ1 depletion due to the expression of COPZ2, which maintains COPI complex function in the absence of the  $\zeta$ 1-isoform<sup>8,35</sup>. Contrary to the depletion of COPZ1, where COPZ2 expression protects normal cells from Golgi disruption, depletion of ARCN1 disrupted the Golgi equally in normal and tumor cells. Therefore, the observed differential sensitivity of the normal fibroblasts and tumor cell lines to ARCN1 depletion, as well as the previously described sensitivity of non-small cell lung cancer cells bearing KRAS mutations<sup>39</sup>, is likely mediated by a different mechanism. It could possibly be attributed to differential dependence on COPI-dependent lipolysis<sup>40,41</sup>; however, extensive additional studies will be needed to identify the precise molecular mechanism of selective inhibition of tumor cell growth by ARCN1 depletion. The importance of ARCN1 for tumor progression is further confirmed by the association of high ARCN1 expression with the poor survival of gastric cancer patients and patients with HER2-positive breast cancer. Overall the results with ARCN1 confirmed our previous finding that targeting the COPI complex is a viable strategy for new cancer therapy.

DDX24 is the most promising target we identified due to potential druggability of the DDX24 protein. DDX24 belongs to a diverse family of DEAD-box RNA helicases (DDXs) that are present in all eukaryotic cells, as well as many bacteria and archaea. According to UniProt there are at least 80 members of the DDX family in human cells. When examined biochemically, family members possess coordinated ATPase and helicase activities and are extensively involved in RNA metabolism<sup>42</sup>. Specifically, DEAD-box RNA helicases appear to be involved in transcription, translation, ribosomal biogenesis, RNA degradation, and small RNA processing<sup>42,43</sup>. Recently, a small molecule inhibitor of DDX3 was developed and shown to attenuate HIV production<sup>44</sup>. In parallel, accumulating evidence is establishing the role of DDX3 and the other members of the family (DDX1, DDX2, and DDX58) as

positive regulators of cancer development and progression<sup>45,46</sup>. The small molecule inhibitor of DDX3 targets the ATPase and helicase activity of DDX3 and was reported to inhibit cell growth and promote tumor regression in a murine model of lung cancer<sup>47</sup>. DDX24 is involved in multiple biological processes including HIV-1 infection<sup>13,48</sup>, interferon signaling<sup>15,16</sup>, and rRNA processing<sup>14,49</sup>. Importantly, DDX24 interacts with MDM2, p300, and p53<sup>14,17,18,50,51</sup>. Our results along with another recent report<sup>18</sup> show that depletion of DDX24 activates p21 expression in a p53 dependent manner. However, we report here that siRNA depletion of DDX24 equally inhibited the growth of both p53 positive and p53 negative tumor cell lines. The mechanisms of growth inhibition by DDX24 depletion have yet to be identified and may involve defects in ribosomal RNA biogenesis<sup>14,49</sup>. In line with our results, inhibition of ribosomal biogenesis induces elevation of the p53 protein<sup>52</sup>. Analysis of publicly available expression profiles of human tumors showed that high level of DDX24 expression correlates with decreased survival of HER2 positive breast cancer and gastric cancer patients. Moreover, significant elevation of the level of DDX24 protein was detected in the majority of tumor cell lines of different origins and in 50% of colon cancer tissues we tested. These results are further confirmed by recent results from another group that show the elevation of DDX24 expression in many breast cancer cell lines<sup>18</sup>. Taken together these results suggest that DDX24 is a promising target for the development of therapy for a variety of tumors.

Collectively, our results show that the approach based on analysis of shRNA depletion in tumor cell lines relative to normal fibroblasts over multiple data points allows for the identification of genes necessary for cancer cell survival, and the identified genes are promising targets for further development of cancer therapeutics.

## Materials and Methods

**Cell lines and human tissues.** MDA-MB-231, HCT116, HT1080, HeLa, U2OS, DU145, 293FT, and PC3 cell lines were obtained from American Type Culture Collection. HCT116 p53<sup>-/-</sup> (clone 379.2)<sup>53</sup> line was a gift of Dr. B. Vogelstein (Johns Hopkins University). BJ normal foreskin fibroblasts, immortalized with human telomerase reverse transcriptase (line BJ1-hTERT) were obtained from Clontech Laboratories. HeLa, U2OS, 293FT, DU145, MDA-MB-231, HCT116, HCT116 p53<sup>-/-</sup>, HT1080, and PC3 cell lines were grown in DMEM with 10% FC-2. BJ fibroblasts were maintained in BJ medium (4:1 DMEM/M199) supplemented with 1 mmol/L of sodium pyruvate and 10% FBS. Flash frozen tumor and adjacent normal colon tissue samples were obtained from the Center for Colon Cancer Research Tissue Biorepository at the University of South Carolina (USC) with oversight and approval from the USC institutional review board. All samples were consented for use in biomedical research at the time of surgery. All the tissue samples and associated data obtained from the biorepository are fully deidentified, therefore materials are considered to be exempt from Human Subject regulations. As such, the presented experiments are not required to have IRB approval.

**Genetic suppressor element screen.** Construction of the genome-wide GSE library and BrdU suicide selection were described previously<sup>8</sup>. To identify genes which were positively selected in this procedure, CLC workbench was used to map GSEs to the reference genome (hg19 downloaded from <http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/>). Peaks were called using the CLC workbench ChIP-seq module (<https://www.qiagenbioinformatics.com/>). The closest gene to a significant (P-value < 0.05) peak was identified (if the peak was not located within a gene then the gene to which the peak was closest was chosen). The total number of reads in each peak were then normalized to the library size and enrichment was calculated by the ratio of cancer cell reads to normal cell reads. Genes with an enrichment greater than 1.5 in 2 or more tumor cell lines and/or targeted with 2 or more GSEs were chosen for further analysis.

**Design and construction of focused shRNA library.** The focused shRNA library was intended to contain 6 shRNAs against each target gene such that: (i) each shRNA was 19bp long, (ii) each duplex stability was between  $-32$  and  $-28$   $\Delta G$ , and (iii) each shRNA uniquely targeted the gene of interest (no other gene had greater homology) using published algorithm<sup>54</sup>. The resulting library contained 1,273 shRNAs, including 6 shRNA per gene for 209 genes, 5 shRNAs for one gene, 4 shRNAs for one gene, 3 shRNAs for two genes, 2 shRNAs for two genes, 1 shRNA for one gene (Supplemental Dataset S2). Each shRNA was assigned an error-correcting barcode sequence, which withstands sequencing errors designed as described in ref. 55. See Supplemental Fig. 2A for the scheme of the oligonucleotide design and Supplemental Dataset S2 for the complete sequences of the oligonucleotides. The shRNA and barcode-containing oligonucleotides were synthesized on-chip (Mycroarray, Ann Arbor MI), 25,000 oligonucleotides per chip followed by droplet PCR amplification with oligo-specific primers. The oligonucleotide mixture was digested with Bpi I (NEB, Ipswich, MA) and inserted into pRSI9-U6-(sh)-UbiC-TagRFP-2A-Puro lentiviral vector (Collecta, Mountain View, CA). The library DNA was amplified as described previously<sup>8</sup>.

**Lentiviral infections.** Lentiviral transduction was carried out as described<sup>56</sup> using pCMV- $\Delta 8.9$  and pVSV-G packaging constructs. The vector plasmids pCMV- $\Delta 8.9$ , pVSV-G, and DNA were mixed at 5:4:1 ratio and cotransfected into 293 FT cells using the polyethylenimine protocol<sup>57</sup>. Lentivirus-containing supernatants were harvested three times, at 24, 48, and 72 hours after transfection. Lentiviral library infection was carried out as previously described<sup>6</sup>.

**Selection of the shRNA library.** Recipient cell lines were infected with the shRNA library in duplicates with a multiplicity of infection (MOI) of 0.9. Infected cells were cultivated over the course of five passages. On each passage one fourth of the cells were re-plated and the rest of the cells were used for purification of genomic DNA. Genomic DNA was purified as previously described<sup>6</sup>. Barcode sequences were amplified from the DNA with three rounds of PCR (Supplemental Fig. 2B), 10 cycles each with the following primer pairs: FwdHTS/RevHTS, FwdGex/Rev-Ind (1:50 ratio), and FwGex/Seq2N-AD (See Supplemental Table S7 for primer sequences and Supplemental Table S8 for index sequences). The amplified products were purified with a PCR purification kit (Qiagen, Valencia, CA). Unique 8 bp indexes assigned to each library were incorporated in the second

round of PCR and Illumina sequencing adapters were incorporated in the third round of PCR amplification (Supplemental Fig. 2B).

**Sequencing.** 50 bp, single end (SE) sequencing was performed at Cornell Epigenomics Core Facility using Illumina HiSeq2000. Resulting reads consisted of shRNA specific barcodes in FASTQ format. FASTQ data were processed in R<sup>58</sup> using an in-house algorithm. Briefly, reads were split by index to produce FASTQ files containing sequences specific to a cell line, passage, and biological replicate. Reads were then assigned to a specific shRNA by identifying the barcode sequence within the read allowing for only a single mismatch.

**siRNA validation.** Validation of hits from GSE and shRNA screens was performed using four different siRNAs from Qiagen (Human genome siRNA Library v1.0) and 4 pooled siRNAs from GE Healthcare (ON-TARGET plus SMARTpool). Four DDX24 specific siRNAs (not included into v1.0 above), SI04189073, SI04138064, SI04132135, SI00361214 were obtained from Qiagen. siRNA transfection was performed using a reverse transfection method. Briefly, siLentFect (Bio-Rad, Hercules, CA) was premixed with Opti-MEM media (ThermoFisher Scientific, Waltham, MA) at 0.075  $\mu$ l siLentFect per 20  $\mu$ l of Opti-MEM according to manufacturer protocol. siRNAs were added such that the final concentration was 5 nM (Qiagen) or 2.5 nM (GE Healthcare) and the siLentFect/Opti-MEM/siRNA mixture was incubated at room temperature for 20–30 min. Cells were added to premixed siRNA/siLentFect complexes at a seeding density of 5,000 cells per well in 96 well plate. siRNA treated cells were grown for 5 days at 37 °C and 5% CO<sub>2</sub>.

**Cell number quantification.** Cell numbers were analyzed using the sulforhodamine B (SRB) assay as described<sup>59</sup>. Briefly, plated cells were fixed with 10% TCA at 4 °C, followed by washing with water. Fixed cells were stained with 0.4% SRB solution followed by washing with 1% acetic acid and drying of the SRB stained cells. SRB was solubilized in 20 mM Tris (pH 10) and quantified by measuring absorbance of SRB solution at 540 nm minus absorbance at 630 nm (background absorbance).

**Colony formation assay.** siRNA-transfected cells were plated at 2,500 cells per well in 6 well plates followed by cultivation for 10 days. Colonies were fixed and stained with crystal violet fixing solution (0.25% Crystal Violet, 3.7% formaldehyde, 10% H<sub>2</sub>O, 80% methanol) as described<sup>60</sup>. Stained plates were imaged with the ChemiDoc Touch Imaging System (Bio-Rad) and resulting images were quantified using the Colony Area Plugin<sup>61</sup> for ImageJ<sup>62</sup>.

**Western blot.** To prepare total protein extracts for Western blot analysis, cells ( $3 \times 10^5 - 10^6$  per sample) were lysed according to standard procedures in TNT lysis buffer (20 mM Tris HCl, pH 7.5, 200 mM NaCl, 5 mM EDTA, 1% Triton X-100), supplemented with 1 mM DTT and Pierce Protease Inhibitor cocktail (Thermo Scientific, Cat. No. 88266). Total protein extracts from flash frozen tissue (5 mg per sample) were crushed followed by homogenization in TNT buffer. Protein concentration was measured using the Pierce BCA protein assay kit (Thermo Scientific, Cat. No. 23227) according to the manufacturer's protocol. Protein expression (20  $\mu$ g per sample) was analyzed by Western Blotting with primary antibodies: p53 (DO-1, BD Biosciences, Cat. No. 554293), p21 (H-164, Santa Cruz Biotechnology, Cat. No. sc-756), DDX24 (Bethyl Laboratories, Cat. No. A300-698A-T), ARCN1 (GeneTex, Cat. No. 103252) GAPDH (Santa Cruz Biotechnology, sc-32233), and  $\beta$ -Actin (Thermo Scientific, Cat. No. MA5-15739) followed by corresponding (HRP)-conjugated secondary antibodies (Thermo Scientific, Cat. Nos 31460, 31430). Membranes were cut into strips containing different molecular weight areas to allow simultaneous analysis of multiple proteins. Each membrane fragment was probed separately and stripped and re-probed if necessary. The digital images were obtained using the ChemiDoc Touch Imaging System (Bio-Rad, Hercules, CA). Complete copies of the WBs are presented in Supplemental Fig. 8.

**Fluorescent microscopy.** Cells cultured on glass coverslips (Bellco Glass) were fixed with fresh 3.7% paraformaldehyde and permeabilized with 0.5% Triton X-100. Fixed and permeabilized cells were blocked with 3% (BSA) and incubated with mouse anti-GM130 primary antibody (BD Biosciences, 610822) and DAPI (Invitrogen). The secondary antibody was Cy3-conjugated goat anti-mouse IgG (Jackson, 711-165-152). Fluorescence images were acquired with a PlanApo/N 60/1.42 NA objective on an Olympus IX81 microscope. Hamamatsu C10600 camera gain and exposure time settings were controlled with Metamorph Basic. Processing of images (merging, brightness, and final size) was performed using Fiji software<sup>62,63</sup>.

**Gene ontology.** Gene ontology analysis was performed using the Gene Ontology Consortium website's enrichment analysis tool<sup>64,65</sup> or Kegg Mapper Pathways tool<sup>66</sup>.

**Association of gene expression with survival of cancer patients.** Association of expression of the genes in GEO deposited Affymetrix data sets with relapse-free survival of breast cancer patients, progression free survival of ovarian cancer patients, and overall survival of gastric and lung cancer patients was determined using KM-plotter online survival analysis tool (<http://kmplot.com/analysis/>) with "Auto select best cutoff" option<sup>67-70</sup>.

**RNA-seq data of the gene expression in tumor and normal tissues.** The expression of RPL35A, DDX24, and ARCN1 mRNAs was obtained from publicly available data from TCGA Firehose<sup>71</sup>, which represents RNA expression data from 38 cancer types and 19,125 samples.

**Statistical analysis of siRNA results.** For siRNA toxicity, we used 4 biological replicates and performed two sided Welch's t-tests with Holm-Sidak correction for multiple testing when necessary. T-tests and multiple testing corrections were performed with GraphPad Prism (GraphPad Software, La Jolla CA).

**Bootstrap LRT to test regression coefficients lying in a cone (shRNA screening analysis).** shRNAs of interest were measured in both normal and cancerous cell lines at different passages. Assuming shRNA measurements are independent and identically distributed normal random variables, a simple linear regression (SLR) is performed to estimate the rate of change in the mean shRNA reads over multiple passages for each type of cell populations. Let  $\hat{Y}_1$  and  $\hat{Y}_2$  denote the estimated mean normalized shRNA reads for the cancer and normal cell lines, respectively. Let  $X_1$  and  $X_2$  denote the covariate vectors that indicate passage number for the measurements in the experiment. Let  $n_1$  and  $n_2$  denote the number of unique shRNAs targeting a gene in cancer and normal cell lines respectively. For  $i = 1, 2, \dots, n_1$ , and  $j = 1, 2, \dots, n_2$ , the fitted values of the shRNA reads are

$$\hat{Y}_{1i} = \hat{\beta}_{10} + \hat{\beta}_{11}X_{1i}; \quad \hat{Y}_{2j} = \hat{\beta}_{20} + \hat{\beta}_{21}X_{2j}; \quad (1)$$

$$\hat{\beta}_{11} = \frac{S_{1,xy}}{S_{1,xx}}; \quad \hat{\beta}_{21} = \frac{S_{2,xy}}{S_{2,xx}}; \quad (2)$$

$$\hat{\beta}_{10} = \bar{Y}_1 - \hat{\beta}_{11}\bar{X}_1; \quad \hat{\beta}_{20} = \bar{Y}_2 - \hat{\beta}_{21}\bar{X}_2; \quad (3)$$

$$S_{1,xx} = \sum_{i=1}^{n_1} (X_{1i} - \bar{X}_1)^2$$

$$S_{2,xx} = \sum_{j=1}^{n_2} (X_{2j} - \bar{X}_2)^2$$

$$S_{1,xy} = \sum_{i=1}^{n_1} (Y_{1i} - \bar{Y}_1)(X_{1i} - \bar{X}_1)$$

$$S_{2,xy} = \sum_{j=1}^{n_2} (Y_{2j} - \bar{Y}_2)(X_{2j} - \bar{X}_2)$$

The goal of the study is to test whether the  $(\beta_{11}, \beta_{21})$ -pair lies within the region of interest, which is the region of  $\left[\frac{3\pi}{4}, \frac{5\pi}{4}\right]$  on a 2D plane formed by using  $\beta_{11}$  as the x-axis, and  $\beta_{21}$  as the y-axis. A bootstrap likelihood ratio test (BLRT) is constructed to compare slopes of the two regression lines in equation (1). We assume equal variances in both normal and cancerous shRNA measurements, denoted by  $\sigma^2$ .  $\sigma^2$ ,  $\beta_{10}$  and  $\beta_{20}$  are then considered as nuisance parameters in this BLRT and we can form the following hypothesis:

$$H_0: \{|\beta_{21}| > |\beta_{11}|, \beta_{11} < 0\} \cup \{\beta_{11} > 0\} \quad H_a: |\beta_{21}| \leq |\beta_{11}|, \beta_{11} < 0 \quad (4)$$

That is, under the null hypothesis the absolute value of the slope of the regression for the normal cell line ( $|\beta_{21}|$ ) is more extreme than the absolute value of the slope of the regression for the cancer cell line ( $|\beta_{11}|$ ) when the slope of the cancer cell line is less than 0 or the slope of the line for cancer cells is greater than 0 (indicating that the shRNA is more effective at reducing the survival/proliferation of normal cells than cancer cells or is ineffective at reducing the survival/proliferation of cancer cells).

Under the alternative hypothesis the absolute value of the slope of the regression for the normal cell line is less extreme than the absolute value of the slope of the regression for the cancer cell line when the cancer cell line slope is less than 0.

Let  $\Theta_r$  denote the restricted parameter space,

$$\Theta_r = \{(\sigma^2, \beta_{10}, \beta_{11}, \beta_{20}, \beta_{21}): \sigma^2 \geq 0, |\beta_{21}| \leq |\beta_{11}|, \beta_{11} < 0, \beta_{10}, \beta_{20} \in \mathbf{R}\}$$

Let  $\Theta$  denote the unrestricted parameter space,

$$\Theta = \{(\sigma^2, \beta_{10}, \beta_{11}, \beta_{20}, \beta_{21}): \sigma^2 \geq 0, \beta_{11}, \beta_{21}, \beta_{10}, \beta_{20} \in \mathbf{R}\}$$

The likelihood ratio  $\lambda(\mathbf{Y}_1, \mathbf{Y}_2)$  is then the likelihood of the null model over the likelihood of the unrestricted model,

$$\lambda(\mathbf{Y}_1, \mathbf{Y}_2) = \frac{\sup_{\Theta_r} L(\Theta | \mathbf{Y}_1, \mathbf{Y}_2)}{\sup_{\Theta} L(\Theta | \mathbf{Y}_1, \mathbf{Y}_2)} \quad (5)$$

where the likelihood of the data is

$$L(\Theta|\mathbf{Y}_1, \mathbf{Y}_2) = \prod_{i=1}^{n_1} (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{[Y_{1i} - (\beta_{10} + \beta_{11}X_{1i})]^2}{2\sigma^2}\right\} \prod_{j=1}^{n_2} (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{[Y_{2j} - (\beta_{20} + \beta_{21}X_{2j})]^2}{2\sigma^2}\right\}$$

In equations (1) and (2), we obtain unrestricted parameter estimators for the regression equations. The restricted estimators are obtained by maximizing the likelihood of the data over the restricted parameter space  $\Theta_r$ . The estimators are shown below

1. When the unrestricted Maximum Likelihood Estimates (MLEs) for  $\beta_{11}$  and  $\beta_{21}$  fall in the  $[\frac{3\pi}{4}, \pi]$  region, then the restricted MLEs of  $\beta_{11}$  and  $\beta_{21}$  satisfy the property that  $\hat{\beta}_{21}^r = -\hat{\beta}_{11}^r$ , that is, they reside on the upper boundary of the region of  $\Theta_r$ . As such the restricted estimates of all the parameters are given by

$$\begin{aligned} \hat{\beta}_{11}^r &= \frac{S_{1,xx}}{S_{1,xx} + S_{2,xx}} \hat{\beta}_{11} - \frac{S_{2,xx}}{S_{1,xx} + S_{2,xx}} \hat{\beta}_{21} \\ \hat{\beta}_{10}^r &= \frac{\sum_{i=1}^{n_1} Y_{1i}}{n_1} - \hat{\beta}_{11}^r \frac{\sum_{i=1}^{n_1} X_{1i}}{n_1} \\ \hat{\beta}_{21}^r &= -\hat{\beta}_{11}^r \\ \hat{\beta}_{20}^r &= \frac{\sum_{j=1}^{n_2} Y_{2j}}{n_2} - \hat{\beta}_{21}^r \frac{\sum_{j=1}^{n_2} X_{2j}}{n_2} \end{aligned}$$

2. When the unrestricted MLEs for  $\beta_{11}$  and  $\beta_{21}$  fall in the  $[\pi, \frac{5\pi}{4}]$  region, then the restricted MLEs of  $\beta_{11}$  and  $\beta_{21}$  satisfy the property that  $\hat{\beta}_{21}^r = \hat{\beta}_{11}^r$ , that is, they reside on the lower boundary of the region of  $\Theta_r$ . As such the restricted estimates of all the parameters are given by

$$\begin{aligned} \hat{\beta}_{11}^r &= \frac{S_{1,xx}}{S_{1,xx} + S_{2,xx}} \hat{\beta}_{11} + \frac{S_{2,xx}}{S_{1,xx} + S_{2,xx}} \hat{\beta}_{21} \\ \hat{\beta}_{10}^r &= \frac{\sum_{i=1}^{n_1} Y_{1i}}{n_1} - \hat{\beta}_{11}^r \frac{\sum_{i=1}^{n_1} X_{1i}}{n_1} \\ \hat{\beta}_{21}^r &= \hat{\beta}_{11}^r \\ \hat{\beta}_{20}^r &= \frac{\sum_{j=1}^{n_2} Y_{2j}}{n_2} - \hat{\beta}_{21}^r \frac{\sum_{j=1}^{n_2} X_{2j}}{n_2} \end{aligned}$$

We then derive the negative log of the LRT statistic using equation (5)

$$-2 \log [\lambda(\mathbf{Y}_1, \mathbf{Y}_2)] = 2(n_1 + n_2) \log \left( \frac{\hat{\sigma}_r^2}{\hat{\sigma}^2} \right) \tag{6}$$

where  $\hat{\sigma}_r^2$  is the estimated variance under the null and  $\hat{\sigma}^2$  is the estimated unrestricted variance, defined as

$$\begin{aligned} \hat{\sigma}_r^2 &= \frac{1}{n_1 + n_2} \left\{ \sum_{i=1}^{n_1} [Y_{1i} - (\hat{\beta}_{10}^r + \hat{\beta}_{11}^r X_{1i})]^2 + \sum_{j=1}^{n_2} [Y_{2j} - (\hat{\beta}_{20}^r + \hat{\beta}_{21}^r X_{2j})]^2 \right\} \\ \hat{\sigma}^2 &= \frac{1}{n_1 + n_2} \left\{ \sum_{i=1}^{n_1} [Y_{1i} - (\hat{\beta}_{10} + \hat{\beta}_{11} X_{1i})]^2 + \sum_{j=1}^{n_2} [Y_{2j} - (\hat{\beta}_{20} + \hat{\beta}_{21} X_{2j})]^2 \right\} \end{aligned}$$

Denote the value of the test statistic from a observed data sample  $T_0$ . To obtain a P-value for the LRT,  $T_0$ , we approximate the null sampling distribution of the likelihood ratio test statistic via bootstrapping. We do this via the following procedure:

1. Obtain SLR residuals  $e_{1i}, i = 1, 2, \dots, n_1$  and  $e_{2j}, j = 1, 2, \dots, n_2$  for both the normal and cancerous cell lines in the sample.  $e_{1i} = Y_{1i} - (\hat{\beta}_{10} + \hat{\beta}_{11} X_{1i}); e_{2j} = Y_{2j} - (\hat{\beta}_{20} + \hat{\beta}_{21} X_{2j})$ , where  $\hat{\beta}_{11}, \hat{\beta}_{10}, \hat{\beta}_{20}, \hat{\beta}_{21}$  are the unrestricted maximum likelihood estimates (MLE) of the regression parameters.
2. Combine vector  $\mathbf{e}_1 = (e_{11}, e_{12}, \dots, e_{1n_1})$  and  $\mathbf{e}_2 = (e_{21}, e_{22}, \dots, e_{2n_2})$ . So, the combined vector is  $\mathbf{e} = (\mathbf{e}_1, \mathbf{e}_2)$ .
3. Obtain two samples of bootstrap residuals of size  $n_1$  and  $n_2$  from  $\mathbf{e}$ . Call the two samples of residuals  $\mathbf{S}_1$  and  $\mathbf{S}_2$ .  $\mathbf{S}_1$  and  $\mathbf{S}_2$  each is a random sample from  $\mathbf{e}$ .
4. Using restricted estimates, create a bootstrap sample  $(\mathbf{Y}_1^*, \mathbf{Y}_2^*)$  by adding non-restricted residuals in step 1. For example, for  $i = 1, 2, \dots, n_1, \mathbf{S}_1 = (s_{11}, s_{12}, \dots, s_{1n_1})$ ,



$$Y_{i}^{*} = \hat{\beta}_{10}^{f} + \hat{\beta}_{11}^{f} X_{1i} + s_{1i} \quad (7)$$

5. Compute the test statistic as in equation (6), say  $T_i^*$  for sample  $(Y_1^*, Y_2^*)$  using the proposed negative log of the LRT.
6. Repeat steps 1–5, and compute a large number, say  $M = 1000$ , such test statistics.
7. The computed bootstrap p-value is given by  $\frac{\sum_{i=1}^M I(T_i^* \geq T_0)}{M}$ . This p-value is used to test the  $H_0$  vs the  $H_1$ .

This analysis was performed in R<sup>58</sup>, using the MASS<sup>72</sup>, reshape2<sup>73</sup>, and preprocessCore<sup>74</sup> packages. The code for the statistical analysis presented here is available for download at [https://github.com/doliv071/Focused\\_shRNA\\_Analysis.git](https://github.com/doliv071/Focused_shRNA_Analysis.git).

## References

1. Roninson, I. B. *et al.* Genetic suppressor elements: new tools for molecular oncology—thirteenth cornelius p. rhoads memorial award lecture. *Cancer research* **55**, 4023–8 (1995).
2. Primiano, T. *et al.* Identification of potential anticancer drug targets through the selection of growth-inhibitory genetic suppressor elements. *Cancer Cell* **4**, 41–53 (2003).
3. Sims, D. *et al.* High-throughput RNA interference screening using pooled shRNA libraries and next generation sequencing. *Genome Biol.* **12**, R104 (2011).
4. Brummelkamp, T. R., Bernards, R. & Agami, R. A system for stable expression of short interfering RNAs in mammalian cells. *Science* **296**, 550–553 (2002).
5. Zhou, Y. *et al.* High-throughput screening of a crispr/cas9 library for functional genomics in human cells. *Nature* **509**, 487–91 (2014).
6. Shtutman, M. *et al.* Function-based gene identification using enzymatically generated normalized shRNA library and massive parallel sequencing. *Proc. Natl. Acad. Sci. USA* **107**, 7377–7382 (2010).
7. Mohr, S. E., Smith, J. A., Shamu, C. E., Neumuller, R. A. & Perrimon, N. RNAi screening comes of age: improved techniques and complementary approaches. *Nat. Rev. Mol. Cell Biol.* **15**, 591–600 (2014).
8. Shtutman, M. *et al.* Tumor-specific silencing of copz2 gene encoding coatomer protein complex subunit zeta 2 renders tumor cells dependent on its paralogous gene copz1. *Proc Natl Acad Sci USA* **108**, 12449–54 (2011).
9. Rao, D. D., Senzer, N., Cleary, M. A. & Nemunaitis, J. Comparative assessment of siRNA and shRNA off target effects: what is slowing clinical development. *Cancer Gene Ther.* **16**, 807–809 (2009).
10. Faulstich, D. *et al.* Architecture of coatomer: molecular characterization of delta-COP and protein interactions within the complex. *J. Cell Biol.* **135**, 53–61 (1996).
11. Cosson, P., Demolliere, C., Hennecke, S., Duden, R. & Letourneur, F. Delta- and zeta-COP, two coatomer subunits homologous to clathrin-associated proteins, are involved in ER retrieval. *EMBO J.* **15**, 1792–1798 (1996).
12. Zhao, Y. *et al.* Cloning and characterization of human DDX24 and mouse Ddx24, two novel putative DEAD-Box proteins, and mapping DDX24 to human chromosome 14q32. *Genomics* **67**, 351–355 (2000).
13. Ma, J. *et al.* The requirement of the DEAD-box protein DDX24 for the packaging of human immunodeficiency virus type 1 RNA. *Virology* **375**, 253–264 (2008).
14. Tafforeau, L. *et al.* The complexity of human ribosome biogenesis revealed by systematic nucleolar screening of Pre-rRNA processing factors. *Mol. Cell* **51**, 539–551 (2013).
15. Li, J. *et al.* A short hairpin RNA screen of interferon-stimulated genes identifies a novel negative regulator of the cellular antiviral response. *MBio* **4**, e00385–00313 (2013).
16. Ma, Z., Moore, R., Xu, X. & Barber, G. N. DDX24 negatively regulates cytosolic RNA-mediated innate immune signaling. *PLoS Pathog.* **9**, e1003721 (2013).
17. Yamauchi, T., Nishiyama, M., Moroishi, T., Yumimoto, K. & Nakayama, K. I. MDM2 mediates nonproteolytic polyubiquitylation of the DEAD-Box RNA helicase DDX24. *Mol. Cell. Biol.* **34**, 3321–3340 (2014).
18. Shi, D., Dai, C., Qin, J. & Gu, W. Negative regulation of the p300-p53 interplay by DDX24. *Oncogene* **35**, 528–536 (2016).
19. Edlund, K. *et al.* Data-driven unbiased curation of the TP53 tumor suppressor gene mutation database and validation by ultradeep sequencing of human tumors. *Proc. Natl. Acad. Sci. USA* **109**, 9551–9556 (2012).
20. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
21. Dai, Z. *et al.* edgeR: a versatile tool for the analysis of shRNA-seq and CRISPR-Cas9 genetic screens. *F1000Res* **3**, 95 (2014).
22. Melnikov, S. *et al.* One core, two shells: bacterial and eukaryotic ribosomes. *Nat. Struct. Mol. Biol.* **19**, 560–567 (2012).
23. Lecompte, O., Ripp, R., Thierry, J. C., Moras, D. & Poch, O. Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. *Nucleic Acids Res.* **30**, 5382–5390 (2002).
24. Martin-Marcos, P., Hinnebusch, A. G. & Tamame, M. Ribosomal protein L33 is required for ribosome biogenesis, subunit joining, and repression of GCN4 translation. *Mol. Cell. Biol.* **27**, 5968–5985 (2007).
25. Uechi, T. *et al.* Ribosomal protein gene knockdown causes developmental defects in zebrafish. *PLoS One* **1**, e37 (2006).
26. Farrar, J. E. *et al.* Abnormalities of the large ribosomal subunit protein, Rpl35a, in Diamond-Blackfan anemia. *Blood* **112**, 1582–1592 (2008).
27. Wang, R. *et al.* Loss of function mutations in RPL27 and RPS27 identified by whole-exome sequencing in Diamond-Blackfan anaemia. *Br. J. Haematol.* **168**, 854–864 (2015).
28. Kuramitsu, M. *et al.* Extensive gene deletions in Japanese patients with Diamond-Blackfan anemia. *Blood* **119**, 2376–2384 (2012).
29. Kroes, R. A. *et al.* The identification of novel therapeutic targets for the treatment of malignant brain tumors. *Cancer Lett.* **156**, 191–198 (2000).
30. Lopez, C. D., Martinovsky, G. & Naumovski, L. Inhibition of cell death by ribosomal protein L35a. *Cancer Lett.* **180**, 195–202 (2002).
31. Wolf, I., Bouquet, C. & Melchers, F. cDNA-library testing identifies transforming genes cooperating with c-myc in mouse pre-B cells. *Eur. J. Immunol.* (2016).
32. Zhou, X., Liao, W. J., Liao, J. M., Liao, P. & Lu, H. Ribosomal proteins: functions beyond the ribosome. *J Mol Cell Biol* **7**, 92–104 (2015).
33. Faini, M., Beck, R., Wieland, F. T. & Briggs, J. A. Vesicle coats: structure, function, and general principles of assembly. *Trends Cell Biol.* **23**, 279–288 (2013).
34. Razi, M., Chan, E. Y. & Tooze, S. A. Early endosomes and endosomal coatomer are required for autophagy. *J. Cell Biol.* **185**, 305–321 (2009).
35. Shtutman, M. & Roninson, I. B. A subunit of coatomer protein complex offers a novel tumor-specific target through a surprising mechanism. *Autophagy* **7**, 1551–1552 (2011).
36. Lippincott-Schwartz, J. & Liu, W. Insights into COPI coat assembly and function in living cells. *Trends Cell Biol.* **16**, 1–4 (2006).

37. Watkin, L. B. *et al.* COPA mutations impair ER-Golgi transport and cause hereditary autoimmune-mediated lung disease and arthritis. *Nat. Genet.* **47**, 654–660 (2015).
38. Izumi, K. *et al.* ARCN1 Mutations Cause a Recognizable Craniofacial Syndrome Due to COPI-Mediated Transport Defects. *Am. J. Hum. Genet.* **99**, 451–459 (2016).
39. Kim, H. S. *et al.* Systematic identification of molecular subtype-selective vulnerabilities in non-small-cell lung cancer. *Cell* **155**, 552–566 (2013).
40. Kobayashi, H., Nishimura, H., Matsumoto, K. & Yoshida, M. Identification of the determinants of 2-deoxyglucose sensitivity in cancer cells by shRNA library screening. *Biochem. Biophys. Res. Commun.* **467**, 121–127 (2015).
41. Singh, S. R. *et al.* The lipolysis pathway sustains normal and transformed stem cells in adult *Drosophila*. *Nature* **538**, 109–113 (2016).
42. Linder, P. & Jankowsky, E. From unwinding to clamping - the DEAD box RNA helicase family. *Nat. Rev. Mol. Cell Biol.* **12**, 505–516 (2011).
43. Jarmoskaite, I. & Russell, R. DEAD-box proteins as RNA helicases and chaperones. *Wiley Interdiscip Rev RNA* **2**, 135–152 (2011).
44. Radi, M. *et al.* Discovery of the first small molecule inhibitor of human DDX3 specifically designed to target the RNA binding site: towards the next generation HIV-1 inhibitors. *Bioorg. Med. Chem. Lett.* **22**, 2094–2098 (2012).
45. Robert, F. & Pelletier, J. Perturbations of RNA helicases in cancer. *Wiley Interdiscip Rev RNA* **4**, 333–349 (2013).
46. Zhao, L. *et al.* Multifunctional DDX3: dual roles in various cancer development and its related signaling pathways. *Am J Cancer Res* **6**, 387–402 (2016).
47. Bol, G. M. *et al.* Targeting DDX3 with a small molecule inhibitor for lung cancer therapy. *EMBO Mol Med* **7**, 648–669 (2015).
48. Gautier, V. W. *et al.* *In vitro* nuclear interactome of the HIV-1 Tat protein. *Retrovirology* **6**, 47 (2009).
49. Andersen, J. S. *et al.* Nucleolar proteome dynamics. *Nature* **433**, 77–83 (2005).
50. Jung, S. Y., Malovannaya, A., Wei, J., O'Malley, B. W. & Qin, J. Proteomic analysis of steady-state nuclear hormone receptor coactivator complexes. *Mol. Endocrinol.* **19**, 2451–2465 (2005).
51. Niture, S. K., Doneanu, C. E., Velu, C. S., Bailey, N. I. & Srivenugopal, K. S. Proteomic analysis of human O6-methylguanine-DNA methyltransferase by affinity chromatography and tandem mass spectrometry. *Biochem. Biophys. Res. Commun.* **337**, 1176–1184 (2005).
52. Scala, F. *et al.* Direct relationship between the level of p53 stabilization induced by rRNA synthesis-inhibiting drugs and the cell ribosome biogenesis rate. *Oncogene* **35**, 977–989 (2016).
53. Bunz, F. *et al.* Requirement for p53 and p21 to sustain G2 arrest after DNA damage. *Science* **282**, 1497–1501 (1998).
54. Matveeva, O. V. *et al.* Optimization of duplex stability and terminal asymmetry for shRNA design. *PLoS ONE* **5**, e10180 (2010).
55. Hamady, M., Walker, J. J., Harris, J. K., Gold, N. J. & Knight, R. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat. Methods* **5**, 235–237 (2008).
56. Zufferey, R., Nagy, D., Mandel, R. J., Naldini, L. & Trono, D. Multiply attenuated lentiviral vector achieves efficient gene delivery *in vivo*. *Nat. Biotechnol.* **15**, 871–875 (1997).
57. Segura, M. M., Garnier, A., Durocher, Y., Coelho, H. & Kamen, A. Production of lentiviral vectors by large-scale transient transfection of suspension cultures and affinity chromatography purification. *Biotechnol. Bioeng.* **98**, 789–799 (2007).
58. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, <https://www.R-project.org/> (2016).
59. Vichai, V. & Kirtikara, K. Sulforhodamine B colorimetric assay for cytotoxicity screening. *Nat Protoc* **1**, 1112–1116 (2006).
60. Dassai, J. P. *et al.* Targeted inhibition of prostate cancer metastases with an RNA aptamer to prostate-specific membrane antigen. *Mol. Ther.* **22**, 1910–1922 (2014).
61. Guzman, C., Bagga, M., Kaur, A., Westermarck, J. & Abankwa, D. ColonyArea: an ImageJ plugin to automatically quantify colony formation in clonogenic assays. *PLoS One* **9**, e92444 (2014).
62. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).
63. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
64. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
65. Blake, J. A. *et al.* Gene Ontology Consortium: going forward. *Nucleic Acids Res.* **43**, D1049–1056 (2015).
66. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
67. Gyorffy, B. *et al.* An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast Cancer Res. Treat.* **123**, 725–731 (2010).
68. Gyorffy, B., Lanczky, A. & Szallasi, Z. Implementing an online tool for genome-wide validation of survival-associated biomarkers in ovarian-cancer using microarray data from 1287 patients. *Endocr. Relat. Cancer* **19**, 197–208 (2012).
69. Gyorffy, B., Surowiak, P., Budczies, J. & Lanczky, A. Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in non-small-cell lung cancer. *PLoS One* **8**, e82241 (2013).
70. Szasz, A. M. *et al.* Cross-validation of survival associated biomarkers in gastric cancer using transcriptomic data of 1,065 patients. *Oncotarget* (2016).
71. Broad Institute TCGA Genome Data Analysis Center. *Analysis-ready standardized TCGA data from Broad GDAC Firehose 2016\_01\_28 run*, [http://gdac.broadinstitute.org/runs/stddata\\_\\_2016\\_01\\_28](http://gdac.broadinstitute.org/runs/stddata__2016_01_28) (2016).
72. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S*, fourth edn. <http://www.stats.ox.ac.uk/pub/MASS4>, ISBN 0-387-95457-0 (Springer, New York, 2002).
73. Wickham, H. Reshaping data with the reshape package. *Journal of Statistical Software* **21**, 1–20, <http://www.jstatsoft.org/v21/i12/> (2007).
74. Bolstad, B. M. *preprocessCore: A collection of pre-processing functions*, <https://github.com/bmbolstad/preprocessCore>, R package version 1.34.0 (2016).
75. Scotto, L. *et al.* Identification of copy number gain and overexpressed genes on chromosome arm 20q by an integrative genomic approach in cervical cancer: potential role in progression. *Genes Chromosomes Cancer* **47**, 755–765 (2008).

## Acknowledgements

We thank Dr. Gary Schools for his assistance with microscopy procedures and analysis, and Dr. Elina Levina for comments that greatly improved the manuscript and for help with manuscript preparation. We are also grateful to Dr. Marina Aksenova for her help with siRNAs-related experiments. This study was supported by Summer Undergraduate Minority Research Fellowship through the Center for Colon Cancer Research at USC, NIH grant P20 RR17698 (A.Z.), AFPE Pre-Doctoral Fellowship (D.O.), NIGMS grant 1P20GM109091-01, NIH grant P30 GM103336 to the Center for Colon Cancer Research at the University of South Carolina and SC INBRE pilot award (M.S.), NIH Grant P20 RR-016461 (H.V.).

### Author Contributions

M.S., D.O., P.B., and A.G. conceived the experiments. D.O., H.J., A.G., E.G., S.L., L.P., and A.Z. conducted the experiments. D.O., P.L., E.P., M.C., E.B., H.V., M.W., and M.S. analyzed the results. M.S. and D.O. wrote the manuscript. All authors reviewed the manuscript.

### Additional Information

**Supplementary information** accompanies this paper at <http://www.nature.com/srep>

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Oliver, D. *et al.* Identification of novel cancer therapeutic targets using a designed and pooled shRNA library screen. *Sci. Rep.* 7, 43023; doi: 10.1038/srep43023 (2017).

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

© The Author(s) 2017