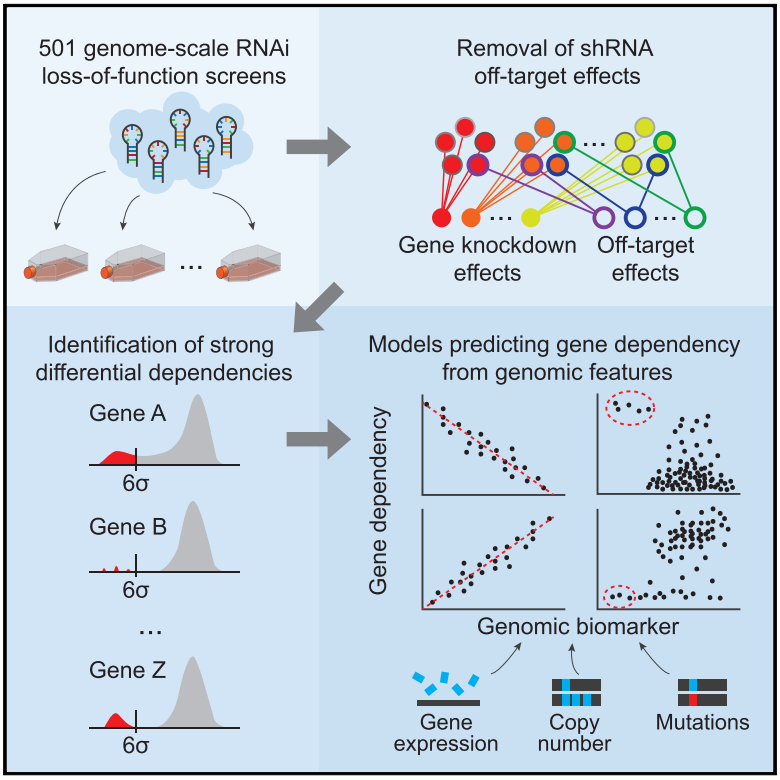
Resource

Cell

Defining a Cancer Dependency Map

Graphical Abstract



Highlights

- The DEMETER computational model segregates on- from off-target effects of RNAi
- 769 strong differential dependencies were identified in 501 cancer cell lines
- Predictive models for 426 dependencies were found using 66,646 molecular features
- This cancer dependency map facilitates the prioritization of therapeutic targets

Authors

Aviad Tsherniak, Francisca Vazquez, Phil G. Montgomery, ..., Todd R. Golub, Jesse S. Boehm, William C. Hahn

Correspondence

william_hahn@dfci.harvard.edu

In Brief

A large-scale analysis of 501 cancer cell lines reveals new vulnerabilities that will help prioritize therapeutic targets

Tsherniak et al., 2017, Cell *170*, 564–576
July 27, 2017 © 2017 Elsevier Inc.
http://dx.doi.org/10.1016/j.cell.2017.06.010

CellPress

Cell

Defining a Cancer Dependency Map

Aviad Tsherniak,[1,6] Francisca Vazquez,[1,2,6] Phil G. Montgomery,[1] Barbara A. Weir,[1,2] Gregory Kryukov,[1,2] Glenn S. Cowley,[1] Stanley Gill,[1,2] William F. Harrington,[1] Sasha Pantel,[1] John M. Krill-Burger,[1] Robin M. Meyers,[1] Levi Ali,[1] Amy Goodale,[1] Yenarae Lee,[1] Guozhi Jiang,[1] Jessica Hsiao,[1] William F.J. Gerath,[1] Sara Howell,[1] Erin Merkel,[1] Mahmoud Ghandi,[1] Levi A. Garraway,[1,2,3,4,5] David E. Root,[1,7] Todd R. Golub,[1,2,4,5,7] Jesse S. Boehm,[1,7] and William C. Hahn[1,2,3,4,7,8,*]

[1]Broad Institute of Harvard and MIT, 415 Main Street, Cambridge, MA, USA
[2]Dana-Farber Cancer Institute, 450 Brookline Avenue, Boston, MA, USA
[3]Department of Medicine, Brigham and Women's Hospital, 75 Francis Street, Boston, MA, USA
[4]Harvard Medical School, 25 Shattuck Street, Boston, MA, USA
[5]Howard Hughes Medical Institute, 4000 Jones Bridge Road, Chevy Chase, MD, USA
[6]These authors contributed equally
[7]Senior author
[8]Lead Contact
*Correspondence: william_hahn@dfci.harvard.edu
http://dx.doi.org/10.1016/j.cell.2017.06.010

SUMMARY

Most human epithelial tumors harbor numerous alterations, making it difficult to predict which genes are required for tumor survival. To systematically identify cancer dependencies, we analyzed 501 genome-scale loss-of-function screens performed in diverse human cancer cell lines. We developed DEMETER, an analytical framework that segregates on- from off-target effects of RNAi. 769 genes were differentially required in subsets of these cell lines at a threshold of six SDs from the mean. We found predictive models for 426 dependencies (55%) by nonlinear regression modeling considering 66,646 molecular features. Many dependencies fall into a limited number of classes, and unexpectedly, in 82% of models, the top biomarkers were expression based. We demonstrated the basis behind one such predictive model linking hypermethylation of the *UBB* ubiquitin gene to a dependency on *UBC*. Together, these observations provide a foundation for a cancer dependency map that facilitates the prioritization of therapeutic targets.

INTRODUCTION

Multiple genetic or epigenetic changes are required to program the malignant state. Although we now have an initial view of the landscape of genetic alterations that occur in cancers, our understanding of the biological impact of these features and how they conspire to induce specific tumor vulnerabilities is largely incomplete. As a result, the use of genetic information from tumors to enable cancer precision medicine is limited.

One approach to identifying genes essential for cancer cell proliferation/survival is to perform systematic loss of function screens in a large number of well-annotated cell lines representing the heterogeneity of tumors. We and others have demonstrated that these experiments are feasible (Aguirre et al., 2016; Cheung et al., 2011; Cowley et al., 2014; Luo et al., 2008; Marcotte et al., 2012, 2016), and the interrogation of single or multiple lineages has identified new oncogenes and genes essential for cell proliferation or the activity of specific signaling pathways (Aguirre et al., 2016; Barbie et al., 2009; Cheung et al., 2011; Cowley et al., 2014; Luo et al., 2008; Marcotte et al., 2012, 2016). However, these RNAi and CRISPR-Cas9 experiments have been limited by off-target effects of such reagents (Aguirre et al., 2016; Birmingham et al., 2006; Buehler et al., 2012b; Jackson and Linsley, 2004; Munoz et al., 2016) and also by an insufficient number of cell line models to adequately represent the full spectrum of the molecular complexity of cancer.

Here, we have integrated a large number of genome-scale RNAi-based loss-of-function screens to facilitate the interrogation of gene function. Using this dataset, we developed an analytical approach that quantifies on- and off-target effects of each RNAi reagent. By combining this information with a comprehensive genomic characterization of these cell lines, we systematically predicted cancer dependencies, thereby establishing an initial framework for a cancer dependencies map.

RESULTS

Overcoming Off-Target Effects of RNAi to Accurately Infer Cancer Dependencies

Although RNAi is a powerful technique, microRNA (miRNA) "seed"-based off-target effects have been reported to confound experimental interpretation (Birmingham et al., 2006; Buehler et al., 2012b; Jackson et al., 2006). We hypothesized that explicitly modeling on- and off-target effects induced by RNAi in a large set of cancer cell lines would provide the means to estimate the on-target effects of suppressing genes in these experiments. We first built on our previous study of 216 human cancer cell lines (Cowley et al., 2014) by screening an additional 285 cell lines. In brief, these screens consist of transducing each cell line

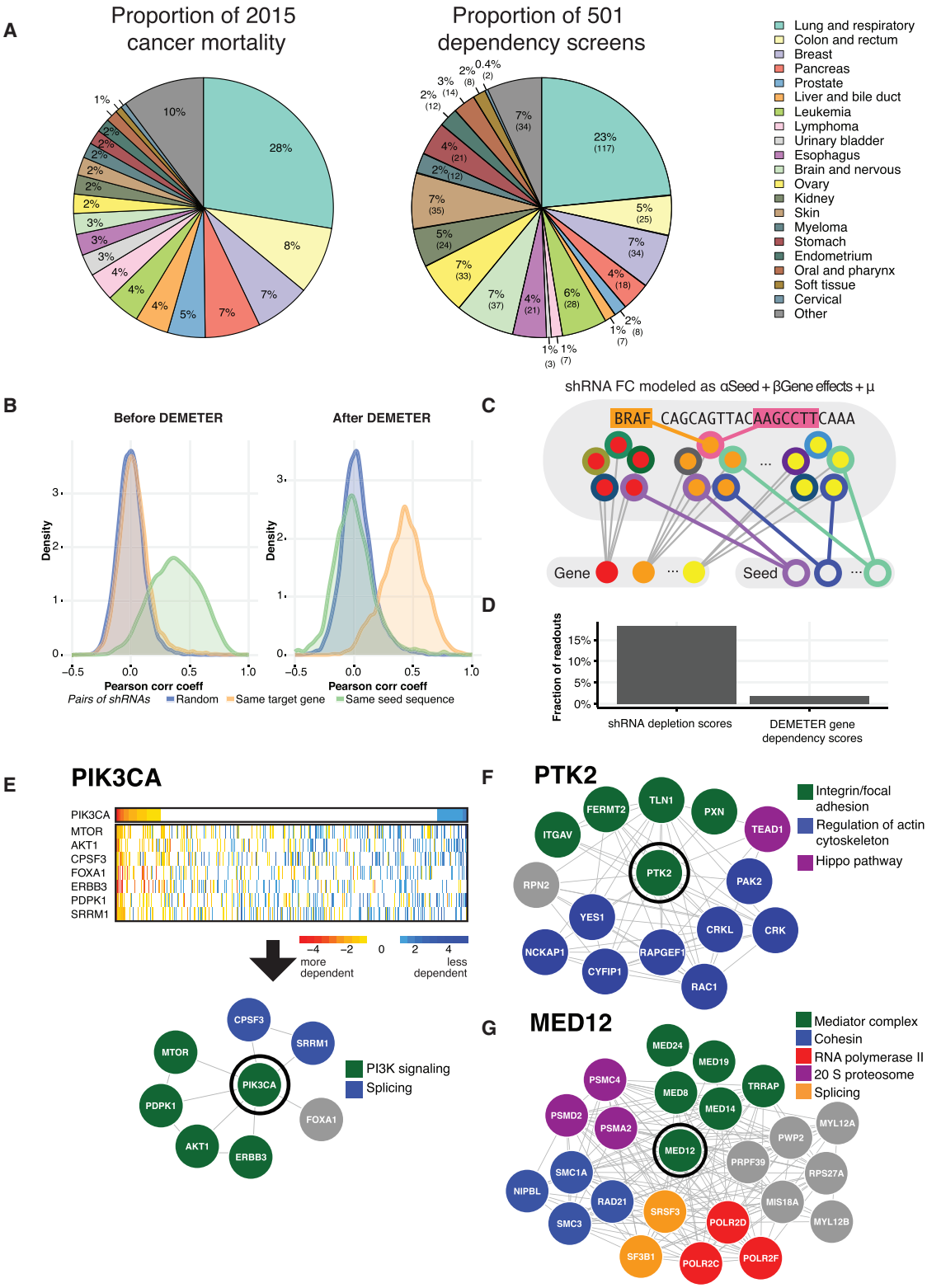


Figure 1. Computational Segregation of On- and Off-Target Effects of RNAi

(A) Tumor types by their contribution to cancer mortality (left) and their cancer cell line representation in the reported dataset (right).

(B) Distributions of Pearson correlation coefficients for pairs of shRNA viability profiles before (left) and after (right) removal of inferred seed effects and selection of effective shRNAs (n > 12,000) by DEMETER. Pairs of shRNAs selected randomly (blue lines), targeting the same gene (orange) and sharing a seed sequence (green).

(legend continued on next page)

with a genome-scale library of ~100,000 short hairpin RNAs (shRNAs) at low MOI in ~60 M cells for each of four replicates, so that each cell gets one shRNA, passaging the cells for 16 doublings, up to 40 days, and then assessing by massively parallel sequencing the depletion of each shRNA from the cell population versus its relative abundance in the original pooled library of shRNA plasmids. The genes targeted by the most depleted shRNAs are inferred to be most essential for proliferation/viability (see STAR Methods for details). The resulting compiled dataset of genome-scale screens in 501 cell lines includes a wide diversity of cancer types (Figure 1A; Table S1).

First, we empirically assessed the prevalence of off-target effects induced by RNAi. Essentially all shRNAs in the library (>99.3%) have a seed sequence that is shared by at least one other shRNA designed to target a different gene (average 12 shRNAs per seed). We found that shRNA depletion scores for pairs of shRNAs that share 7-mer miRNA-like seed sequences were significantly more correlated (mean Pearson correlation coefficient r = 0.37) than profiles of shRNAs targeting the same gene (mean r = 0.03; p value $<10^{-15}$, Mann-Whitney U test; Figures 1B and S1B). These observations confirm that miRNA-like seed effects are highly prevalent in RNAi in this dataset.

Both on-target and seed-based effects of RNAi are sequence specific. However, previous solutions to overcome seed effects have been incomplete as they focused on reduction of false-positive results using multiple shRNA constructs targeting each gene (Kampmann et al., 2013, 2015), inferring on-target effects by identifying shRNA constructs that induce strong concordant on-target effects (Shao et al., 2013), or identifying the seed-based effects (Buehler et al., 2012b; Yilmazel et al., 2014). The gespeR approach (Schmich et al., 2015) considers both on- and off-target effects but involves a computational prediction of seed targets for each reagent. We reasoned that explicitly modeling the combined on-target and seed-based effects directly from the empirical screen data would improve the estimates of the gene-knockdown effects. We, therefore, developed a computational method (DEMETER) that uses the depletion values induced by each shRNA construct to infer the effect of suppressing its intended target (on-target) and of expressing a given miRNA seed (off-target) in each screened cell line. It models each depletion value as a sum of two unobserved quantities: gene knockdown and seed-based effects. It then estimates these quantities by fitting the model to the full dataset. This is possible as the shRNA libraries we used contain multiple shRNAs designed to target each gene as well as multiple shRNAs harboring each seed sequence (Figure 1C; STAR Methods). We applied DEMETER to obtain, in each of 501 cell lines, gene-level differential dependency scores for 17,098 unique genes and seed-sequence effects for 15,142 unique 7-mer sequences (available at https://portals.broadinstitute.org/achilles), as well as performance metrics for each shRNA (Table S2). When we subtracted inferred seed effects from each shRNA and recomputed the correlation coefficient between shRNA constructs targeting the same gene, we found that gene-targeting shRNA pairs were now substantially more correlated (p value $<10^{-15}$, Mann-Whitney test; Figure 1B), validating our approach.

To determine whether DEMETER facilitates the use of RNAi to identify biological relationships, we assessed three parameters. First, we reasoned that non-expressed genes were unlikely to be required for viability. Indeed, the fraction of the highest (top 0.1%) DEMETER dependency scores that represented gene-cell line combinations where the gene was non-expressed was 9-fold lower than for the most dependent shRNA-level readouts (Figure 1D). This finding is consistent with our prediction that DEMETER effectively corrects for off-target effects of shRNAs. Second, we compared the dependency profiles corresponding to a subset of genes encoding physically interacting proteins and found a 43-fold increase in highly correlated (Pearson r *Z* score >3) dependency profiles among 20,466 pairs of gene products annotated to be in the same physical complex as compared to random gene pairs (p value $<10^{-15}$; Fisher's exact test; Figure S1A; STAR Methods). This represents a 3-fold improvement over the performance of a correlation-based method (Shao et al., 2013). Third, by extending this finding to members of the same pathway, we confirmed that we were able to discover known biological relationships directly from correlated dependency profiles (STAR Methods; Table S3). We note three representative examples: (1) *PIK3CA* dependency profiles were tightly correlated with known pathway members (*MTOR*, *PDPK1*, *AKT1*, and *ERBB3*) (Figure 1E), (2) cell lines that were more dependent on the expression of the *PTK2* tyrosine kinase were also more dependent on specific members of integrin/focal adhesion, and actin cytoskeleton regulating pathways (Figure 1F), and (3) cell lines dependent on *MED12* were correlated with members of the mediator complex (Figure 1G). These cases were among many other examples such as members of the PRC2, SWI/SNF complexes, and mitochondrial respiratory genes where one or more members of the complex were identified (Figure S1C).

Furthermore, we noted that cells that depend on *PIK3CA* also required the expression of the key splicing mediators *CPSF3* and *SRRM1* (Figure 1E), and cells that depend on *PTK2* required the transcription factor *TEAD*1 and the glycosyltransferase *RPN2* (Figure 1F). Finally, cells that required *MED12* also depended on specific members of the cohesin, splicing, 20S proteasome, and RNA polymerase complex (Figure 1G), suggesting that this

(C) Schematic representation of DEMETER and its computational model. Gene- and seed-related effects are estimated from shRNA depletion data. The color of inner circles represents the shRNA target gene, and the color of outer circles represents the shRNAs seed sequence.
(D) For the top 0.1% most depleted shRNA readouts and the top 0.1% DEMETER gene dependency scores across the whole dataset, the fraction of data points corresponding to a cell line not expressing the target gene.
(E) A heatmap depicts the dependency scores (rows) across 501 screened lines (columns) for PIK3CA and the seven genes that have significantly correlated dependency profiles (*Z* score >3). These data were used to plot a gene network, with each edge representing a significant correlation between a pair of dependency profiles.
(F and G) Genes are colored by functional classes. The same analysis was used to generate gene networks for PTK2 (F) and MED12 (G).
See also Figure S1 and Tables S1, S2, and S3.

approach also permits the discovery of new co-dependency relationships. Together, these observations demonstrate that DEMETER provides a rigorous approach to distinguishing on- and off-target effects of RNAi and facilitates the discovery of novel cancer dependencies and biology.

Systematic Identification of Differential Dependencies

We next undertook a census of cancer dependencies. To define those more likely to be cancer specific, we focused on genes with a robust differential dependency identified in a minority of the 501 cancer cell lines (DEMETER gene dependency scores that are multiple SDs beyond the mean) (Figure 2A).

The number of differential dependencies identified in this census is a function of both the magnitude of the differential dependency and its prevalence in our cell line collection (Figure 2B). Across the 501 cell lines, we identified a set of 769 strong differential gene dependencies for which the DEMETER scores of at least one cell line were six SDs (6σ) or greater from the mean across all cell lines (Table S4). Using a stringent threshold provides high confidence that these are true differential dependencies rather than false-positive results. We found that 92% of the cell lines (n = 460) harbored at least one such 6σ dependency (Figure 2C). Overall, these 769 genes represent many different classes of proteins including transcription factors and kinases (Figure 2D), and 20% of these (n = 152) have been annotated as potentially druggable (Figure 2E). Furthermore, 53 of the 6σ dependencies are common to at least 5% of the cell lines (n = 25). Consistent with these observations, we found that as few as 76 genes represent 6σ dependencies in 92% of the cell lines, and indeed we found multiple gene sets of this size. Similarly, sets of only ten genes captured 6σ dependencies in 58% the cell lines. This observation suggests that a modest number of therapeutic targets might be relevant across a disproportionately large number of tumors. Indeed, 74% of the cell lines had at least one 6σ dependency representing a readily druggable target (Figure 2F; Table S4).

Predicting Dependencies from Molecular Features

The ability to predict cancer dependencies from tumor features may provide insights into mechanism and opportunities for patient stratification. Thus, we next asked whether we could identify features that predict these 6σ dependencies. To achieve this goal, we developed a nonlinear regression model (ATLANTIS) that is based on conditional inference trees (Hothorn et al., 2006), an adaptation of the random forest model (see STAR Methods). We used it to create predictive models for gene dependency scores from 66,646 molecular features (somatic gene mutations, gene copy number, gene expression) measured at baseline as part of the Cancer Cell Line Encyclopedia (CCLE) project (Barretina et al., 2012) (see STAR Methods). We initially focused on non-hematopoietic cell lines because they represented the majority of the cell lines (455/501) and because they have substantially different gene expression patterns than hematopoietic cell lines (Barretina et al., 2012).

Using this approach, we generated predictive models (marker dependency pairs [MDPs]) with statistically significant accuracy (false discovery rate [FDR] <0.05; permutation test) for 289 (38%) of the 769 6σ dependencies (see STAR Methods; Figures 3A and 3B). An unbiased approach utilizing a large number of candidate predictive features (66,646) is useful for finding unexpected marker dependency relationships, but it also creates a very high bar for statistical significance. To address this, we also employed an alternative approach whereby the feature space was reduced based on prior biological knowledge. Specifically, for each target dependency, we used molecular features of genes representing direct physical interaction, membership in protein complexes, or membership in known signaling pathways (named collectively "related features," see below and STAR Methods). These metrics yielded 361 significant MDPs, of which 251 overlap with the unbiased approach (Figures 3A and 3B). Having discovered MDPs for high-confidence 6σ dependencies, we next applied them to 5,536 candidate dependencies at lower confidence levels (between a threshold of 2σ and 6σ from the mean). These additional analyses netted significant MDPs for 741 additional genes, a rate (13.4%) much lower than observed for 6σ dependencies (51.8%), reflecting the lower signal in this candidate dependencies set (Figures 3B and S2A).

We next examined the nature of the biomarkers that led to predictive models of dependency. Specifically, we asked whether DNA mutation, copy number, or RNA expression was particularly informative with respect to predicting dependencies. Surprisingly, the vast majority of predictable differential dependencies (82%) were best predicted by RNA expression levels, whereas DNA mutation accounted for only 16% and DNA copy number only 2% (Figure 3C). This observation is in concordance with the observation that small-molecule cancer dependencies are similarly most commonly predicted by gene expression (Seashore-Ludlow et al., 2015).

While these MDPs included many previously described relationships (Figures 3D and 6B), additional markers were discovered in most cases. For example, we found that mutations in *KRAS* or *BRAF* were anticorrelated with dependency on *PTPN11*, an activator of the RAS pathway (Figure 3D). Likewise, expression of known TP53 transcriptional targets (*RPS27L*, *CDKN1A*, and *EDA2R*) as well as the *ELMSAN1*, and *ACER2* genes predicted *MDM4* dependency, consistent with MDM4 functioning as a negative regulator of TP53. Novel biological relationships were also discovered, suggesting new mechanistic hypotheses. For instance, strong dependency on the actin-regulating *CYFIP1* gene was predicted by expression of integrin and membrane raft proteins (*ICAM4*, *ITGB4*, *MALL*) (Figure 3D). In many cases, multivariate predictive models, which use multiple features, held greater predictive power than those restricted to single features (Figure 3E). Together, these results support the notion that the ability to predict a cancer dependency provides helpful insight into the mechanistic underpinnings driving differential dependencies in cancer.

Classification of Differential Dependencies

Having found a large number of dependencies (many of which are accompanied by predictive biomarkers), we asked whether they could be classified into distinct biological classes.

One class of MDPs, where somatic mutation or copy number gain of a gene predicts a dependency on the same gene for

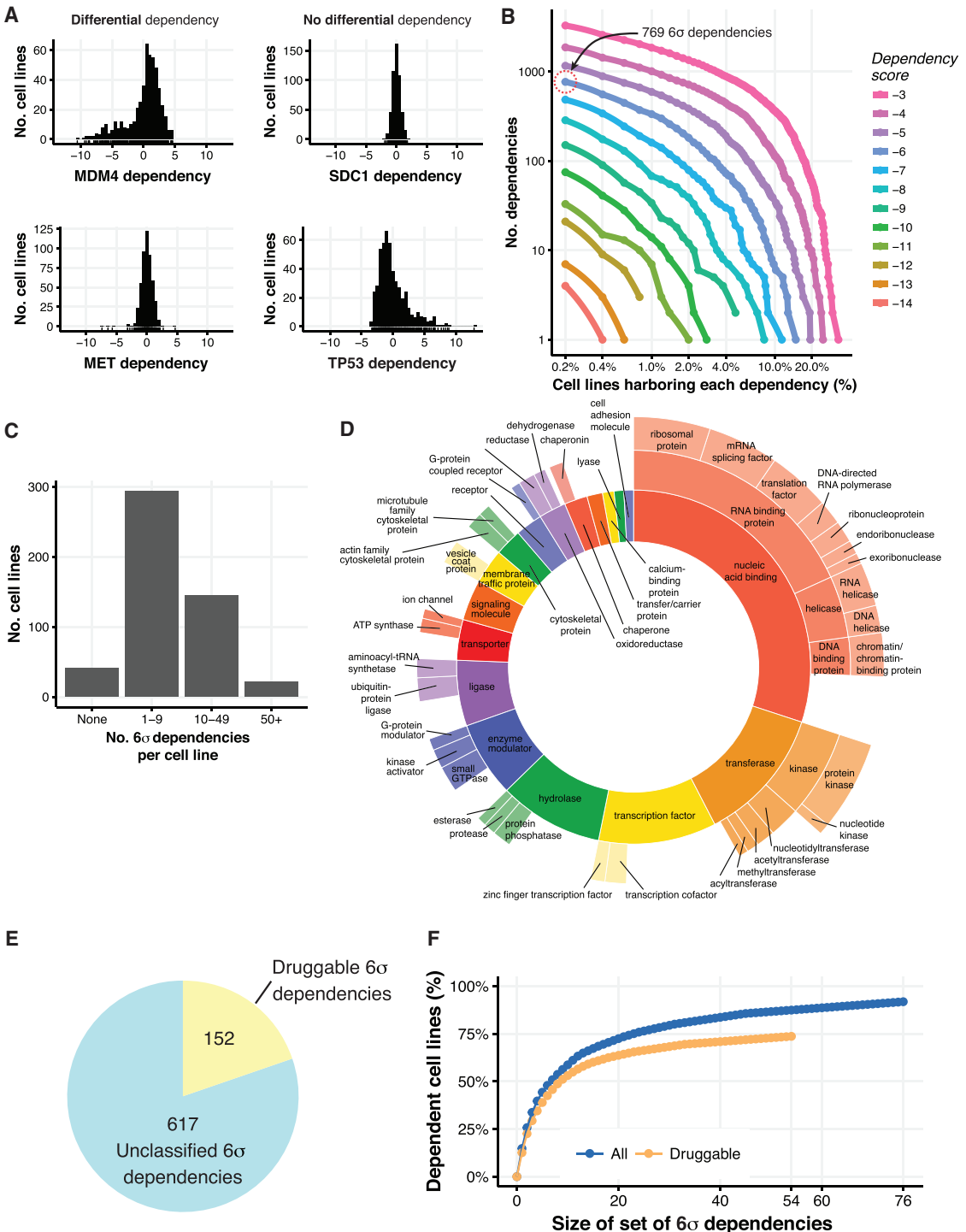


Figure 2. The Landscape of Genetic Dependencies in 501 Cancer Cell Lines

(A) Histograms of gene dependency scores for the indicated genes for all cell lines (x axis).

(B) For each differential dependency strength (line color), and for each number of cell lines (x axis), the number of genes that are differential dependencies is shown (y axis).

(C) Distribution of the number of 6σ dependencies per cell line.

(D) Distribution of 6σ dependencies by protein classes.

(E) The number of 6σ dependencies annotated as druggable by either being included in DGIdb or International Union of Basic and Clinical Pharmacology (IUPHAR)/British Pharmacological Society (BPS) Guide to Pharmacology.

(F) The fraction of cell lines (y axis) that have a 6σ differential dependency on at least one gene in a set of a given size (x axis). Blue line, considering all 6σ dependencies; orange line, considering only druggable ones.

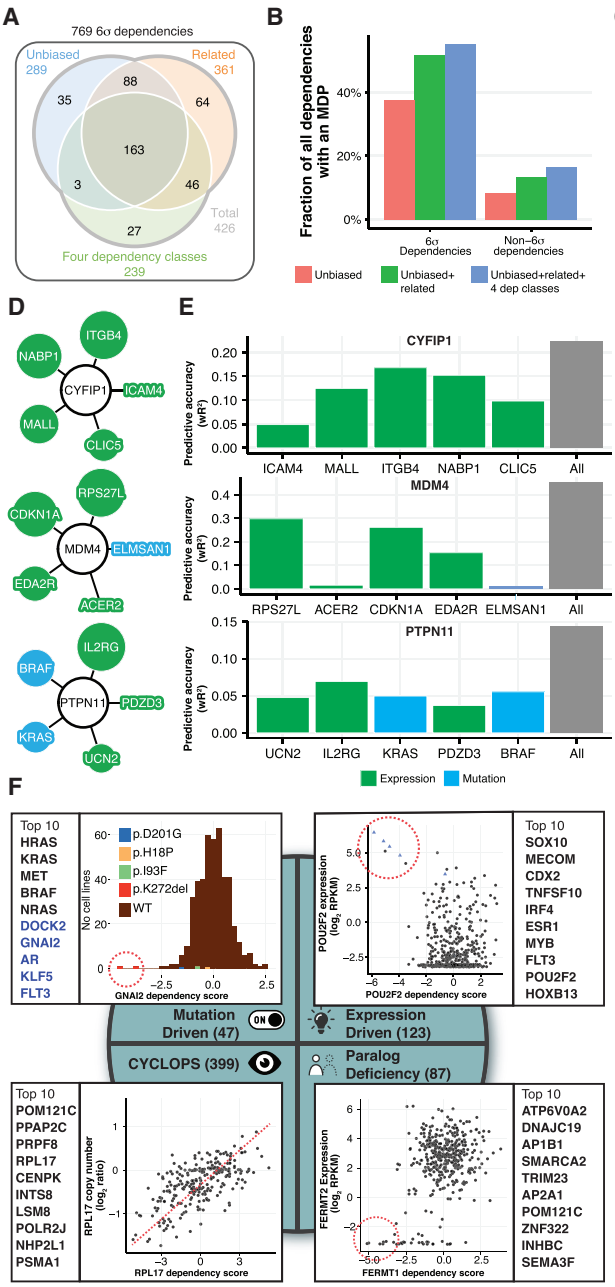


Figure 3. Prediction of Differential Dependencies Using Molecular Markers

(A) The number of 6σ dependencies with predictive models built using all features (Unbiased, blue), features of genes related to the dependency gene (Related, red), and those falling into one of the four identified dependency classes (green).

(B) Cumulative fraction of 6σ and non-6σ dependencies with predictive models (y axis) using all features (red bars), plus related features (green), plus those in the four dependency classes (blue).

(C) The proportion of the top predictive feature type (copy number, orange; expression, green; mutation, blue) in all unbiased models of 6σ dependencies.

(D) Top five features of predictive models for three gene dependencies in white circles. Circle size is proportional to the relative importance of each feature to the model's predictive power.

(E) Predictive accuracy of ATLANTIS models using only single features (black and colored bars) and using all features (gray bars).

(F) Four classes of MDPs, each with a representative example and the top ten predictable dependencies. Red dotted circles highlight the most dependent cell lines. (top left) A histogram of GNAI2 dependency scores (x axis). The two cell lines most dependent on GNAI2 harbor the same indel mutation. (top right) POU2F2 dependency scores (x axis) and expression levels (y axis). Cell lines overexpressing POU2F2 are the most dependent lines. (bottom left) RPL17 dependency (x axis) and copy number (y axis) illustrating a CYCLOPS dependency. (bottom right) FERMT1 dependency (x axis) and FERMT2 expression levels (y axis) for cell lines with low expression of FERMT3 (log_2RPKM <3). Cell lines most dependent on FERMT1 do not express FERMT2.

See also Figure S2 and Tables S4, S5, and S6.

survival, includes known oncogenes. To identify such MDPs, we attempted to build models that would predict each dependency using only the gene's own mutation and amplification features; however, we noted that in some cases few cell lines existed harboring each mutation, limiting our statistical power. Thus, for completeness, we also searched for cases in which cell lines differentially dependent on a gene were enriched for mutations in that gene (Table S5; see STAR Methods). In total, we discovered 47 such mutation-driven MDPs, including 18 corresponding to 6σ dependencies (Figure 3F; Table S6).

While these dependencies included the known oncogenes *KRAS*, *NRAS*, *HRAS*, *BRAF*, *PIK3CA*, *MET*, *MCL1*, *MDM2*, and *ESR1*, they also included multiple novel dependencies including *SOX10*, *DOCK2*, and *GNAI2*. Interestingly, the two diffuse large B cell lymphoma cell lines with a 6σ dependency on the small GTPase *GNAI2* (Morin et al., 2013) harbored the same in-frame deletion (p.K272del), suggesting that such mutations are activating and that targeting GNAI2 in *GNAI2* mutant cancers might be an effective therapeutic strategy (Figure 3F, top left).

By contrast, 399 (30%) of the dependencies with biomarkers, including 184 6σ dependencies, represented genes for which hemizygous copy number loss and/or reduction in expression levels were predictive of increased dependency. These findings extend our previous report describing this class of cancer dependencies that we termed CYCLOPS genes (Nijhawan et al., 2012) (Figure 3F, bottom left; Table S6; STAR Methods). This class of MDPs includes the previously validated dependencies *PSMC2* (Nijhawan et al., 2012) and *POLR2A* (Liu et al., 2015) as well as novel candidates such as members of the kinetochore

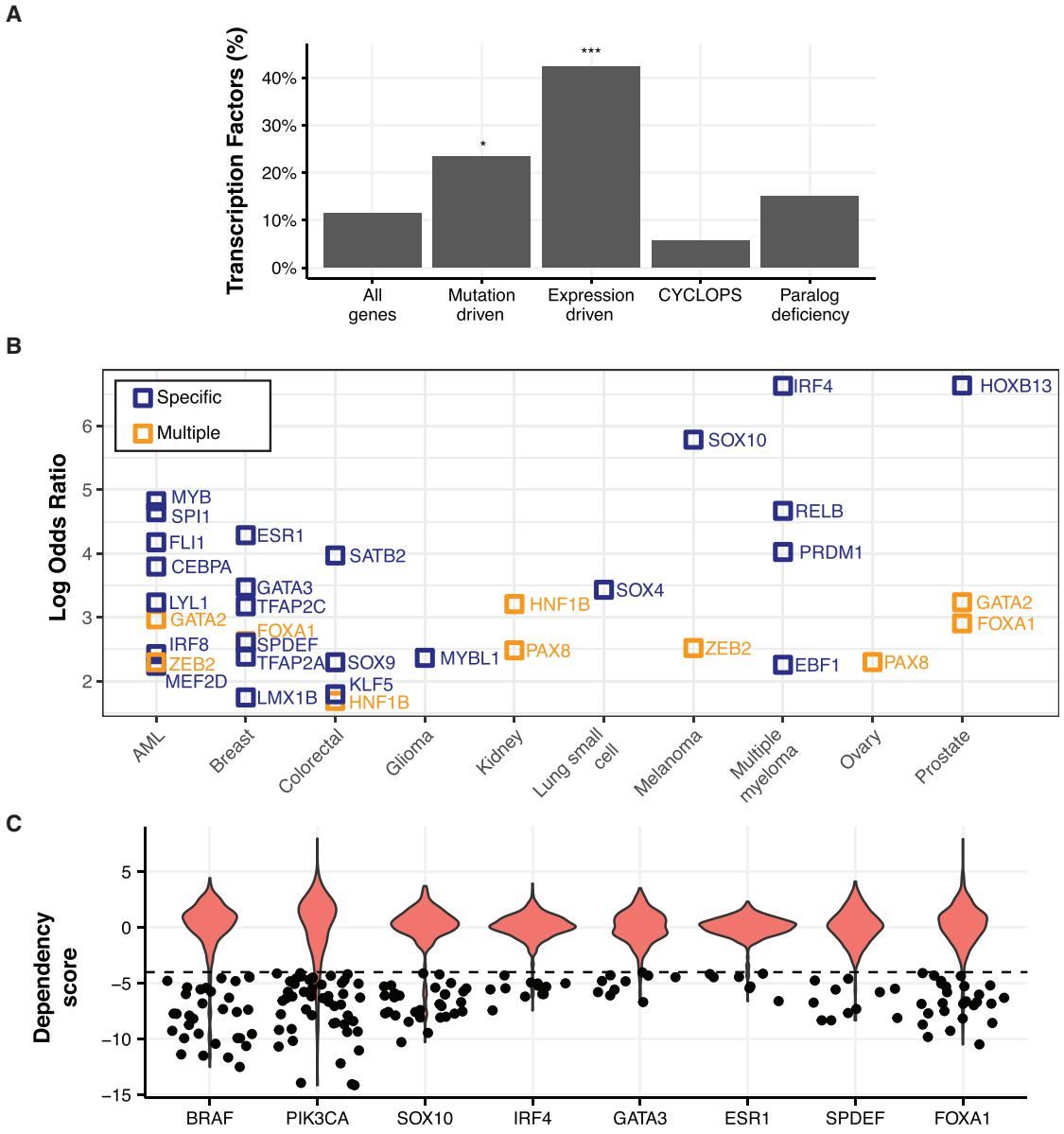


Figure 4. Oncogene and Expression Addiction MDPs Are Enriched in Lineage-Specific Transcription Factors

(A) Percentage of transcription factors (TF) among all genes and the four dependency classes. *p value <0.05, ***p value $<10^{-15}$, Fisher's exact test.

(B) Lineage enrichment (odds ratio; y axis) of mutation- and expression-driven TF dependencies (n = 50) for lineages (x axis) with significant enrichment (Fisher's exact FDR <0.05) in a single (blue) or multiple (orange) lineages.

(C) Distributions of 6σ TF dependencies overrepresented in non-essential lineages (ovary, breast, prostate, multiple myeloma, and melanoma) compared to known mutation-driven dependencies (BRAF, PIK3CA); dots depict dependency scores greater than 4σ.

associated complex (*SKA1*), SET1 complex (*WDR82*), or mediator complex (*MED9*).

We next evaluated a third distinct class of MDPs, representing genes whose elevated expression is associated with dependency. Such expression-driven dependencies include lineage-specifying transcription factors such as *SPDEF*, *NKX2-1*, and *PAX8* (Buchwalter et al., 2013; Cheung et al., 2011; Weir et al., 2007). In all, we discovered 123 (9%) such dependencies, including 33 6σ dependencies (see STAR Methods). Indeed, 49 (45%) of such dependencies were transcription factors (Figure 4A), many known to act as master regulators in the specification and survival of particular tissue lineages (Buchwalter et al., 2013; Laury et al., 2011).

We next investigated in greater detail the relationships between specific cancer types and master transcription factor dependencies. Since targeting such transcription factors may also induce cell death in normal tissues expressing those factors, we paid particular attention to transcription factor dependencies

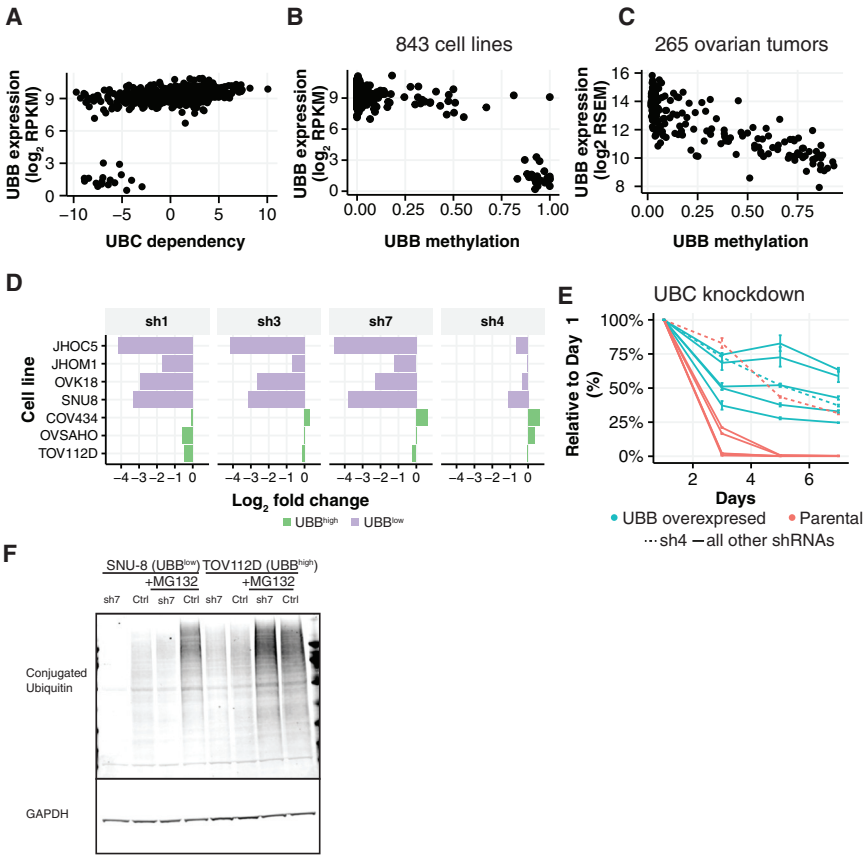


Figure 5. *UBB/UBC* as a Paralog Deficiency MDP in Ovarian Cancer Cell Lines

(A) *UBC* dependency scores (x axis) versus *UBB* expression levels (y axis).

(B and C) UBB expression (y axis) versus promoter methylation (x axis; Fraction) in (B) ovarian cell lines (CCLE data) and (C) tumors (TCGA data).

(D) GFP viability competition assay in UBB^{low} and UBB^{high} ovarian cell lines using four shRNAs targeting *UBC*. Log_2 fold change of shUBC-expressing cells relative to negative controls is shown.

(E) Time course of relative viability upon *UBC* suppression with or without ectopic expression of monoubiquitin (UBB) in a UBB^{low} cell line (SNU8). Data represent fold change relative to day 1 normalized to pLKO_TRC005 nullT. Error bars represent SD.

(F) Levels of conjugated ubiquitin upon *UBC* suppression in UBB^{low} (SNU8) and UBB^{high} (TOV112D) cell lines.

See also Figures S3 and S4.

restricted to specific cell lineages. Indeed, while multiple lineages were dependent on transcription factors such as *TEAD1*, several cancer lineages were specifically dependent on particular master transcription factors (Figure 4B), including *ESR1*, *TFAP2C*, *GATA3*, *SPDEF*, and *FOXA1* in breast cancer and *HOXB13* in prostate cancer, as previously reported (Buchwalter et al., 2013; Marcotte et al., 2016; Pomerantz et al., 2015), as well as novel candidates including *SATB2* in colorectal cancer and *LYL1* in acute myeloid leukemia (AML). Particularly interesting among these lineage-related dependencies are those involved in cell types or organs that are not essential for adult survival (e.g., prostate, breast, thyroid, ovary, melanocytes, plasma cells). Examples of 6σ dependencies in dispensable lineages include *ESR1*, *FOXA1*, *GATA3*, *IRF4*, *SOX10*, and *SPDEF*; the strength of such dependencies was comparable to mutation-driven dependencies (Figure 4C). Together, these observations suggest that these strong lineage-specific cancer dependencies represent potential cancer targets as evidenced by the success of estrogen receptor inhibitors in breast cancer.

Finally, we observed a fourth prominent class of 87 dependencies (7%), including 27 6σ ones, for which the functional loss of one paralog is associated with a dependency on another. While previous reports have noted examples of such paralog deficiency dependencies (Aksoy et al., 2014; D'Antonio et al., 2013; Helming et al., 2014; Muller et al., 2012; Wilson et al., 2014), here we systematically identified over 80 such dependencies using ATLANTIS (Table S6). For example, we identified low *FERMT2* expression as a marker for *FERMT1* dependency, a gene involved in integrin and cytoskeleton regulation (Figure 3F, bottom right). Focusing only on solid tumor lineages, where *FERMT2* is mostly expressed (Figure S2B), we found that very few cell lines expressed either FERMT1 or FERMT2, and the subset of cells with no *FERMT2* expression was exquisitely dependent on *FERMT1* (Figures S2C and S2D). These results indicate that epithelial cells require either *FERMT1* or *FERMT2* for survival.

Together these observations demonstrate that a large fraction (45.8%) of the dependencies, for which a predictive model was found, fall into at least one of these four classes (Figures 3A and 3F). Moreover, mutation-driven dependencies represented only a small minority of these dependencies, suggesting that there exists a large number of unexpected, strong differential dependencies that may serve as therapeutic targets.

Mechanistic Investigation of *UBC* Dependency

Dependency on the *UBC* ubiquitin gene was one of the most highly predictable 6σ paralog deficiency dependencies (Table S6), with low expression of *UBB* as the top marker (Figure S3A). Indeed, we found that all 20 cell lines (100%) with low expression of the *UBB* ubiquitin gene were highly dependent on the *UBC* ubiquitin gene (Figure 5A).

Since *UBB* expression is uniform across the majority of normal tissues (Figure S3B; GTEx), we hypothesized that somatic loss of *UBB* expression in cancer occurred through gene deletion or epigenetic silencing. While no relationship was observed with copy number (Figure S3C), we found that loss of *UBB* expression and *UBB* promoter hypermethylation was frequent in ovarian and uterine tumors (Figures S3D and S3E). *UBB* expression was correlated with promoter hypermethylation, as assessed

by reduced-representation bisulfite sequencing (RRBS) in both cell lines and ovarian tumors (Figures 5B and 5C).

We next validated the DEMETER-inferred *UBC* dependency in ovarian cancer cell lines. Indeed, four cell lines expressing low levels of *UBB* were highly dependent on *UBC* in contrast to three cell lines expressing average *UBB* levels (p value <0.03, Mann-Whitney U test; Figure 5D). As expected, the degree of *UBC* effect inversely correlated with DEMETER gene values (Figures S4A). Moreover, RNAi reagents that contained matched seed sequences but do not target *UBC* failed to induce cell death (see STAR Methods [C911 controls]; Figure S4B) (Buehler et al., 2012a), confirming that the observed effects were due to on-target activities of these shRNAs.

We further explored the *UBB*-*UBC* dependency relationship. First, we found that *UBB* and *UBC* are co-regulated, since cancer cell lines that express low levels of *UBB* expressed higher levels of *UBC* (Figure S4D). We also found that *UBC* suppression induced *UBB* expression (Figure S4C). Exogenous expression of monoubiquitin from a *UBB* ORF in cell lines with low *UBB* levels alleviated the requirement for *UBC* expression (p value = 0.026, F-test) (Figures 5E and S4E). Finally, we found that suppression of *UBC* expression resulted in a decrease in total levels of conjugated ubiquitin in UBB^{low} but not UBB^{high} cell lines (Figures 5F and S4F).

Taken together, these results confirm that cells require either *UBB* or *UBC* for survival, suggesting that these proteins may functionally buffer each other. The recent elucidation of protein degradation as the mechanism by which lenalidomide induces cell death in myeloma suggests that targeting this and other MDPs may prove useful (Krönke et al., 2014; Lu et al., 2014). In addition, these observations demonstrate that MDPs may not only have diagnostic potential, but also facilitate rapid insights into the mechanistic basis of dependencies in cancer.

Progress toward a Cancer Dependency Map

A consensus visualization of the results described above produced an initial map of cancer dependencies and predictive power (Figures 6A and S5A–S5F). As a final step, we took two complementary approaches to determine the completeness of this map. First, we curated a list of 39 oncogene addictions from the literature, including validated drug responses (Table S7; see STAR Methods). Our dataset identified a differential dependency on 33 (85%) of these genes and returned the "concordant" marker in 20 (51%) instances (Figure 6B). For the other 13 cases (33%), either distinct, yet biologically meaningful markers were discovered (five) or the dataset did not include cell lines that harbored the validated marker (six).

In six (15%) of the remaining cases, the dataset did not include cell lines that harbored the validated marker. Accordingly, we successfully derived predictive models for 86% of the 6σ dependencies present in over 20 cell lines, but only 45% of the 6σ dependencies present in only one cell line (Figure 6C). These observations suggest that more cellular contexts are needed to both observe and predict each dependency.

Leveraging these concepts, we performed a down-sampling analysis to evaluate how scaling the number of cell line contexts relates to the ability to observe dependencies. In this analysis, we first determined the sensitivity of smaller datasets to observe dependencies discovered in the complete dataset (Figure 6D, blue line). These results show an inflection point in the rate of 6σ dependency discovery at a dataset size of 200–300 cell lines. While exact extrapolation is difficult due to cell line contexts that are completely absent, these results are consistent with a prediction that approximately 1,000 cell lines may be needed to observe most 6σ dependencies in cancer at least once. However, given the result that observing a dependency in >20 cell lines is required to predict >80% of 6σ dependencies (Figure 6C), we noted that at least an order of magnitude increase in scale beyond the present 501 cell lines (>5,000) is likely to be needed to fully predict most cancer dependencies from cell features (Figure 6D, green and red lines).

DISCUSSION

Using RNAi-based, loss-of-function genetic screens in 501 cancer cell lines, we identified genes whose expression is required for the proliferation or survival of subsets of these cell lines and developed an approach to identifying features that predict these gene dependencies. This cancer dependency map provides an approach to defining and predicting genes that are essential for cell viability, thereby facilitating the identification of cancer targets. We have made all of these data and analysis results available at https://depmap.org/rnai.

The off-target effects of shRNAs have become increasingly recognized, and this has led to skepticism about the utility of RNAi-based screens. To the contrary, we show here that such off-target effects can be distinguished from on-target effects resulting in highly reproducible and biologically meaningful results. We previously reported the use of the ATARiS algorithm to integrate across often discordant measurements obtained from different shRNAs targeting the same gene (Shao et al., 2013). While somewhat effective, residual off-target shRNA effects remained. Related approaches to minimize off-target effects have similarly been described (Cheung et al., 2011; König et al., 2007; Marcotte et al., 2016; Zhang et al., 2011). The DEMETER method introduced here, however, leverages the observation that the majority of shRNA off-target effects are attributable to miRNA seed sequences. We hypothesized that explicitly modeling such seed effects would improve the performance of algorithms such as ATARiS, that are based solely on correlation. Indeed, DEMETER dramatically outperformed ATARiS in our analysis of 501 cancer cell lines. Notably, in contrast to other approaches that attempt to model RNAi seed effects (Schmich et al., 2015), DEMETER requires no prior knowledge of the off-target effects of a given shRNA; DEMETER automatically identifies seed effects for any collection of shRNAs.

An alternative way to address the off-target effects of shRNA is to use other loss-of-function approaches. Specifically, genome editing through the use of CRISPR-Cas9 technology has emerged as a promising complementary method to RNAi to identify essential genes. Although CRISPR-Cas9-mediated gene editing exhibits a high degree of specificity in gene targeting, we and others have recently reported that Cas9 endonuclease activity induces a gene-independent cell-cycle arrest, likely due to DNA damage (Aguirre et al., 2016; Munoz et al., 2016; Wang et al., 2015). In addition, we recently showed that

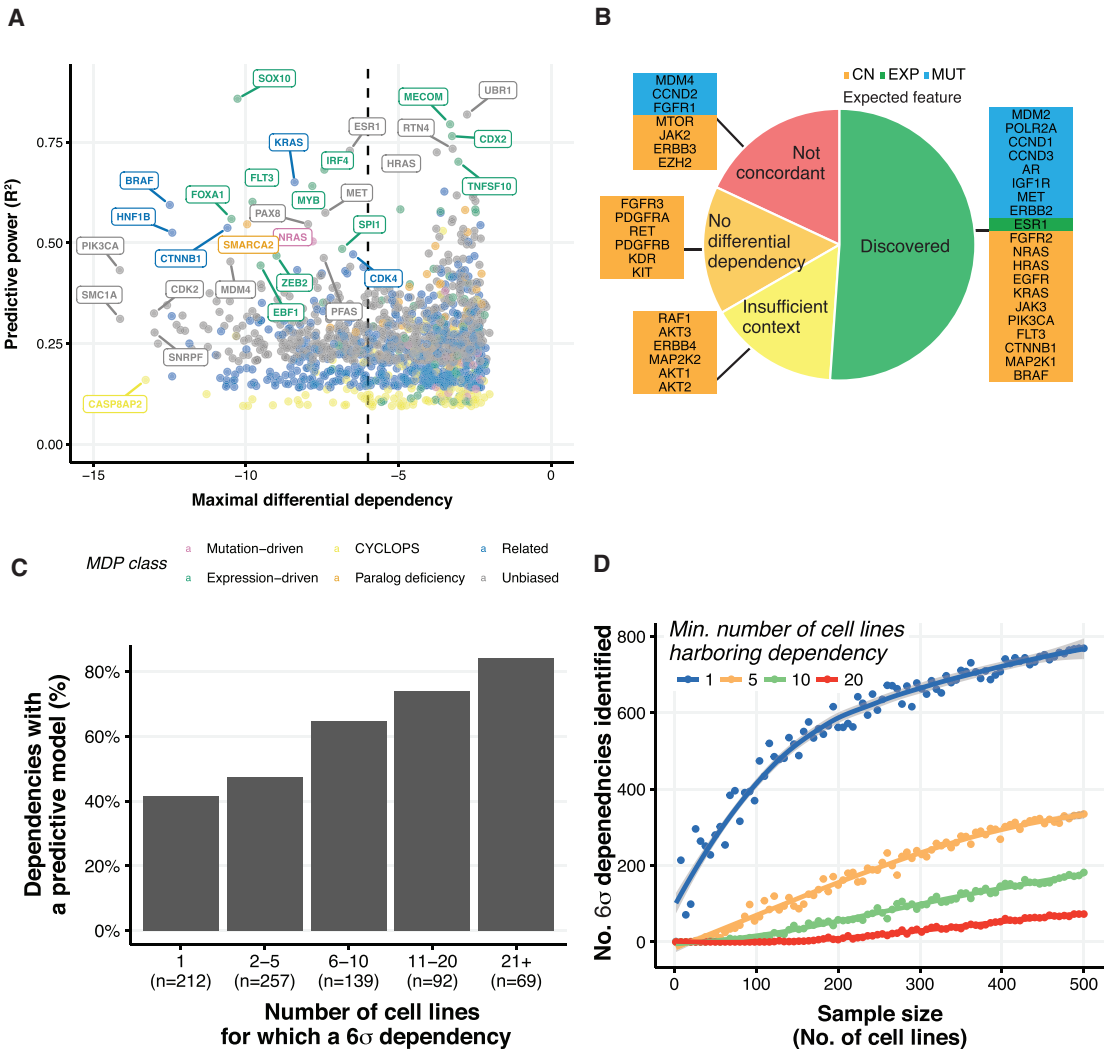


Figure 6. Effects of Scale on the Completeness of a Cancer Dependency Map

(A) For each differential dependency with a significant predictive model, the predictive power of the best model (y axis) and its MDP class (color) along with the strength of the dependency in the most dependent cell line (x axis).

(B) Discovery status of a curated set of 39 mutation- and expression-related dependencies in the dataset. We computed the correlations of each marker with all the differential dependencies and categorized them as (1) discovered, (2) not concordant (3) insufficient context, or (4) no differential dependency (see STAR Methods).

(C) Fraction of predictable 6σ dependencies, summarized by the number of 6σ-dependent cell lines.

(D) Results of a down-sampling analysis showing the number of 6σ differential dependencies identified (y axis) in randomly sampled subsets of the screened cell lines (x axis). The blue, orange, green, and red lines correspond to dependencies observed in at least 1, 5, 10, or 20 cell lines, respectively.

See also Table S7.

gene suppression rather than gene deletion permits the identification of gene dependencies, such as CYCLOPS genes (Rosenbluh et al., 2016). Taken together, these observations suggest that the information from CRISPR-Cas9 and RNAi screens are complementary.

The cancer dependencies identified in these studies represent targets for therapeutic efforts. Although this initial report allowed us to define several classes of gene dependencies, we recognize that this approach is focused on biological processes essential for cell-autonomous cell survival. Moreover, we defined cancer dependencies based on cell proliferation and survival. Future studies using analogous approaches will be necessary to interrogate cell-cell interactions and other cancer phenotypes, which may expand the number and types of cancer dependencies.

Although we identified both known and novel oncogenes, genes that are somatically mutated and/or focally amplified represent a minority of the cancer dependencies. Indeed, gene expression emerged as the molecular feature that best predicted differential dependency. Since most therapeutic targeting efforts have focused on mutated oncogenes, these efforts suggest that a large number of cancer targets remain to

be tested for efficacy when targeted therapeutically. Although defining and validating these dependencies will require substantial further validation, these observations suggest that targeting these gene dependencies may allow the identification of a larger set of cancer targets suitable for therapeutic targeting. Moreover, expanding these types of studies to a larger set of cancer cell lines and phenotypes provides a path to defining a comprehensive map of cancer dependencies as well as the context (genetic, cell-cell interactions, etc.) that drive these MDP relationships.

Our observations indicate that the comprehensive identification and prediction of dependencies will require a substantial increase in the number and diversity of cell lines analyzed (Figures 6C and 6D). Thus, we propose that a concerted, international effort should be launched to create a definitive cancer dependency map. Such a map would serve as a foundation for the entire field, leading to a blueprint for targeted therapeutic development, and to an acceleration of cancer precision medicine.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENTS AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Cell lines
- METHODS DETAILS
  - Screening and deconvolution using next-generation sequencing
  - Cloning of C911 shRNAs
  - Viral production
  - GFP competition assay
  - UBB rescue experiments
  - Western Blots
  - RT-PCR
  - DEMETER
  - Assessing shRNA performance
  - Data processing pipeline
  - Applying DEMETER to 501 RNAi screens
  - Benchmarking DEMETER against ATARiS
  - Correlation of dependency profiles (Figures 1E-G and S1B)
  - Differential dependencies and 6σ dependencies
  - RNASeq
  - ATLANTIS
  - Identifying dependency classes (Table S6)
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - GFP competition assay (Figure 5D)
  - UBB rescue (Figure 5E)
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes five figures and seven tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2017.06.010.

AUTHOR CONTRIBUTIONS

A.T. conceived the project, designed the analytical framework, developed computational methods, and performed bioinformatic analysis, interpreted results, and wrote the manuscript. F.V. conceived the project, designed the analytical framework, designed and analyzed experimental work, provided management support for RNAi screens, interpreted results, and wrote the manuscript. P.G.M. developed computational methods and performed bioinformatic analysis and wrote the manuscript. B.A.W. prepared the RNAi datasets and contributed to the analytical framework and to writing the manuscript. G.K. designed and performed bioinformatic analysis of the CCLE datasets and contributed to the design of the analytical framework. G.S.C. designed and provided project management for all the RNAi screens. S.G. designed and performed and analyzed experimental work. W.F.H. performed and analyzed experimental work. S.P. performed RNAi screens. J.M.K.-B. contributed to the design and performance of bioinformatic analysis, performed statistical analysis for experimental work, and contributed to writing the manuscript. R.M.M. contributed to the bioinformatic analysis. L.A., A.G., Y.L., G.J., J.H., and W.F.J.G. performed RNAi screens. E.M. contributed to experimental work. S.H. and M.G. contributed to the bioinformatic analysis. L.A.G. contributed to project design. D.E.R. conceived and designed the project, directed the RNAi screens, and interpreted results. T.R.G. and J.S.B. conceived and designed the project and wrote the manuscript. W.C.H. conceived and designed the project, directed and interpreted the results, and wrote the manuscript.

ACKNOWLEDGMENTS

We thank Andrew Tubelli for help with graphical design. Screening data can be accessed at https://portals.broadinstitute.org/achilles. All analysis results and software are available at https://depmap.org/rnai. This work was funded by NIH U01 CA176058 (to W.C.H.), NIH ICBP grant U54 CA112962 (to T.R.G.), and The Carlos Slim Foundation in Mexico through the Slim Initiative for Genomic Medicine (to T.R.G.). W.C.H. and L.A.G. report receiving commercial research grants from Novartis and is a consultant/advisory board member for the same. W.C.H. is a advisory board member for KSQ Therapeutics.

Received: January 12, 2017
Revised: April 9, 2017
Accepted: June 7, 2017
Published: July 27, 2017

REFERENCES

Aguirre, A.J., Meyers, R.M., Weir, B.A., Vazquez, F., Zhang, C.Z., Ben-David, U., Cook, A., Ha, G., Harrington, W.F., Doshi, M.B., et al. (2016). Genomic copy number dictates a gene-independent cell response to CRISPR/Cas9 targeting. Cancer Discov. *6*, 914–929.

Aksoy, B.A., Demir, E., Babur, Ö., Wang, W., Jing, X., Schultz, N., and Sander, C. (2014). Prediction of individualized therapeutic vulnerabilities in cancer from genomic profiles. Bioinformatics *30*, 2051–2059.

Ashton, J.M., Balys, M., Neering, S.J., Hassane, D.C., Cowley, G., Root, D.E., Miller, P.G., Ebert, B.L., McMurray, H.R., Land, H., and Jordan, C.T. (2012). Gene sets identified with oncogene cooperativity analysis regulate in vivo growth and survival of leukemia stem cells. Cell Stem Cell *11*, 359–372.

Barbie, D.A., Tamayo, P., Boehm, J.S., Kim, S.Y., Moody, S.E., Dunn, I.F., Schinzel, A.C., Sandy, P., Meylan, E., Scholl, C., et al. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature *462*, 108–112.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603–607.

Birmingham, A., Anderson, E.M., Reynolds, A., Ilsley-Tyree, D., Leake, D., Fedorov, Y., Baskerville, S., Maksimova, E., Robinson, K., Karpilow, J., et al.

(2006). 3′ UTR seed matches, but not overall identity, are associated with RNAi off-targets. Nat. Methods *3*, 199–204.

Brenner, J.C., Ateeq, B., Li, Y., Yocum, A.K., Cao, Q., Asangani, I.A., Patel, S., Wang, X., Liang, H., Yu, J., et al. (2011). Mechanistic rationale for inhibition of poly(ADP-ribose) polymerase in ETS gene fusion-positive prostate cancer. Cancer Cell *19*, 664–678.

Buchwalter, G., Hickey, M.M., Cromer, A., Selfors, L.M., Gunawardane, R.N., Frishman, J., Jeselsohn, R., Lim, E., Chi, D., Fu, X., et al. (2013). PDEF promotes luminal differentiation and acts as a survival factor for ER-positive breast cancer cells. Cancer Cell *23*, 753–767.

Buehler, E., Chen, Y.C., and Martin, S. (2012a). C911: A bench-level control for sequence specific siRNA off-target effects. PLoS ONE *7*, e51942.

Buehler, E., Khan, A.A., Marine, S., Rajaram, M., Bahl, A., Burchard, J., and Ferrer, M. (2012b). siRNA off-target effects in genome-wide screens identify signaling pathway members. Sci. Rep. *2*, 428.

Cheung, H.W., Cowley, G.S., Weir, B.A., Boehm, J.S., Rusin, S., Scott, J.A., East, A., Ali, L.D., Lizotte, P.H., Wong, T.C., et al. (2011). Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. Proc. Natl. Acad. Sci. USA *108*, 12372–12377.

Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. *31*, 213–219.

Cowley, G.S., Weir, B.A., Vazquez, F., Tamayo, P., Scott, J.A., Rusin, S., East-Seletsky, A., Ali, L.D., Gerath, W.F., Pantel, S.E., et al. (2014). Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. Sci. Data *1*, 140035.

D'Antonio, M., Guerra, R.F., Cereda, M., Marchesi, S., Montani, F., Nicassio, F., Di Fiore, P.P., and Ciccarelli, F.D. (2013). Recessive cancer genes engage in negative genetic interactions with their functional paralogs. Cell Rep. *5*, 1519–1526.

DeLuca, D.S., Levin, J.Z., Sivachenko, A., Fennell, T., Nazaire, M.D., Williams, C., Reich, M., Winckler, W., and Getz, G. (2012). RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics *28*, 1530–1532.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: Ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Helming, K.C., Wang, X., Wilson, B.G., Vazquez, F., Haswell, J.R., Manchester, H.E., Kim, Y., Kryukov, G.V., Ghandi, M., Aguirre, A.J., et al. (2014). ARID1B is a specific vulnerability in ARID1A-mutant cancers. Nat. Med. *20*, 251–254.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. J. Comput. Graph. Stat. *15*, 651–674.

Jackson, A.L., and Linsley, P.S. (2004). Noise amidst the silence: Off-target effects of siRNAs? Trends Genet. *20*, 521–524.

Jackson, A.L., Burchard, J., Schelter, J., Chau, B.N., Cleary, M., Lim, L., and Linsley, P.S. (2006). Widespread siRNA "off-target" transcript silencing mediated by seed region sequence complementarity. RNA *12*, 1179–1187.

Kampmann, M., Bassik, M.C., and Weissman, J.S. (2013). Integrated platform for genome-wide screening and construction of high-density genetic interaction maps in mammalian cells. Proc. Natl. Acad. Sci. USA *110*, E2317–E2326.

Kampmann, M., Horlbeck, M.A., Chen, Y., Tsai, J.C., Bassik, M.C., Gilbert, L.A., Villalta, J.E., Kwon, S.C., Chang, H., Kim, V.N., and Weissman, J.S. (2015). Next-generation libraries for robust RNA interference-based genome-wide screens. Proc. Natl. Acad. Sci. USA *112*, E3384–E3391.

König, R., Chiang, C.Y., Tu, B.P., Yan, S.F., DeJesus, P.D., Romero, A., Bergauer, T., Orth, A., Krueger, U., Zhou, Y., and Chanda, S.K. (2007). A probability-based approach for the analysis of large-scale RNAi screens. Nat. Methods *4*, 847–849.

Krönke, J., Hurst, S.N., and Ebert, B.L. (2014). Lenalidomide induces degradation of IKZF1 and IKZF3. OncoImmunology *3*, e941742.

Lage, K., Karlberg, E.O., Størling, Z.M., Olason, P.I., Pedersen, A.G., Rigina, O., Hinsby, A.M., Tümer, Z., Pociot, F., Tommerup, N., et al. (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat. Biotechnol. *25*, 309–316.

Lage, K., Hansen, N.T., Karlberg, E.O., Eklund, A.C., Roque, F.S., Donahoe, P.K., Szallasi, Z., Jensen, T.S., and Brunak, S. (2008). A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. Proc. Natl. Acad. Sci. USA *105*, 20870–20875.

Laury, A.R., Perets, R., Piao, H., Krane, J.F., Barletta, J.A., French, C., Chirieac, L.R., Lis, R., Loda, M., Hornick, J.L., et al. (2011). A comprehensive analysis of PAX8 expression in human epithelial tumors. Am. J. Surg. Pathol. *35*, 816–826.

Liu, Y., Zhang, X., Han, C., Wan, G., Huang, X., Ivan, C., Jiang, D., Rodriguez-Aguayo, C., Lopez-Berestein, G., Rao, P.H., et al. (2015). TP53 loss creates therapeutic vulnerability in colorectal cancer. Nature *520*, 697–701.

Lu, G., Middleton, R.E., Sun, H., Naniong, M., Ott, C.J., Mitsiades, C.S., Wong, K.K., Bradner, J.E., and Kaelin, W.G., Jr. (2014). The myeloma drug lenalidomide promotes the cereblon-dependent destruction of Ikaros proteins. Science *343*, 305–309.

Luo, B., Cheung, H.W., Subramanian, A., Sharifnia, T., Okamoto, M., Yang, X., Hinkle, G., Boehm, J.S., Beroukhim, R., Weir, B.A., et al. (2008). Highly parallel identification of essential genes in cancer cells. Proc. Natl. Acad. Sci. USA *105*, 20380–20385.

Marcotte, R., Brown, K.R., Suarez, F., Sayad, A., Karamboulas, K., Krzyzanowski, P.M., Sircoulomb, F., Medrano, M., Fedyshyn, Y., Koh, J.L., et al. (2012). Essential gene profiles in breast, pancreatic, and ovarian cancer cells. Cancer Discov. *2*, 172–189.

Marcotte, R., Sayad, A., Brown, K.R., Sanchez-Garcia, F., Reimand, J., Haider, M., Virtanen, C., Bradner, J.E., Bader, G.D., Mills, G.B., et al. (2016). Functional genomic landscape of human breast cancer drivers, vulnerabilities, and resistance. Cell *164*, 293–309.

Mi, H., Poudel, S., Muruganujan, A., Casagrande, J.T., and Thomas, P.D. (2016). PANTHER version 10: Expanded protein families and functions, and analysis tools. Nucleic Acids Res. *44*(D1), D336–D342.

Morin, R.D., Mungall, K., Pleasance, E., Mungall, A.J., Goya, R., Huff, R.D., Scott, D.W., Ding, J., Roth, A., Chiu, R., et al. (2013). Mutational and structural analysis of diffuse large B-cell lymphoma using whole-genome sequencing. Blood *122*, 1256–1265.

Muller, F.L., Colla, S., Aquilanti, E., Manzo, V.E., Genovese, G., Lee, J., Eisenson, D., Narurkar, R., Deng, P., Nezi, L., et al. (2012). Passenger deletions generate therapeutic vulnerabilities in cancer. Nature *488*, 337–342.

Munoz, D.M., Cassiani, P.J., Li, L., Billy, E., Korn, J.M., Jones, M.D., Golji, J., Ruddy, D.A., Yu, K., McAllister, G., et al. (2016). CRISPR screens provide a comprehensive assessment of cancer vulnerabilities but generate false-positive hits for highly amplified genomic regions. Cancer Discov. *6*, 900–913.

Nijhawan, D., Zack, T.I., Ren, Y., Strickland, M.R., Lamothe, R., Schumacher, S.E., Tsherniak, A., Besche, H.C., Rosenbluh, J., Shehata, S., et al. (2012). Cancer vulnerabilities unveiled by genomic loss. Cell *150*, 842–854.

Pomerantz, M.M., Li, F., Takeda, D.Y., Lenci, R., Chonkar, A., Chabot, M., Cejas, P., Vazquez, F., Cook, J., Shivdasani, R.A., et al. (2015). The androgen receptor cistrome is extensively reprogrammed in human prostate tumorigenesis. Nat. Genet. *47*, 1346–1351.

Ramos, A.H., Lichtenstein, L., Gupta, M., Lawrence, M.S., Pugh, T.J., Saksena, G., Meyerson, M., and Getz, G. (2015). Oncotator: Cancer variant annotation tool. Hum. Mutat. *36*, E2423–E2429.

Rosenbluh, J., Mercer, J., Shrestha, Y., Oliver, R., Tamayo, P., Doench, J.G., Tirosh, I., Piccioni, F., Hartenian, E., Horn, H., et al. (2016). Genetic and proteomic interrogation of lower confidence candidate genes reveals signaling networks in beta-catenin-active cancers. Cell Syst. *3*, 302–316.

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.W. (2010). CORUM: The comprehensive resource of mammalian protein complexes–2009. Nucleic Acids Res. *38*, D497–D501.

Saunders, C.T., Wong, W.S., Swamy, S., Becq, J., Murray, L.J., and Cheetham, R.K. (2012). Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics *28*, 1811–1817.

Schmich, F., Szczurek, E., Kreibich, S., Dilling, S., Andritschke, D., Casanova, A., Low, S.H., Eicher, S., Muntwiler, S., Emmenlauer, M., et al. (2015). gespeR: A statistical model for deconvoluting off-target-confounded RNA interference screens. Genome Biol. *16*, 220.

Seashore-Ludlow, B., Rees, M.G., Cheah, J.H., Cokol, M., Price, E.V., Coletti, M.E., Jones, V., Bodycombe, N.E., Soule, C.K., Gould, J., et al. (2015). Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. Cancer Discov. *5*, 1210–1223.

Shao, D.D., Tsherniak, A., Gopal, S., Weir, B.A., Tamayo, P., Stransky, N., Schumacher, S.E., Zack, T.I., Beroukhim, R., Garraway, L.A., et al. (2013). ATARiS: Computational quantification of gene suppression phenotypes from multisample RNAi screens. Genome Res. *23*, 665–678.

Southan, C., Sharman, J.L., Benson, H.E., Faccenda, E., Pawson, A.J., Alexander, S.P., Buneman, O.P., Davenport, A.P., McGrath, J.C., Peters, J.A., et al.; NC-IUPHAR (2016). The IUPHAR/BPS Guide to PHARMACOLOGY in 2016: Towards curated quantitative interactions between 1300 protein targets and 6000 ligands. Nucleic Acids Res. *44*(D1), D1054–D1068.

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Stein, T.I., Nudel, R., Lieder, I., Mazor, Y., et al. (2016). The GeneCards Suite: From gene data mining to disease genome sequence analyses. Curr. Protoc Bioinformatics *54*, 1.30.1–1.30.33.

Stewart, S.A., Dykxhoorn, D.M., Palliser, D., Mizuno, H., Yu, E.Y., An, D.S., Sabatini, D.M., Chen, I.S., Hahn, W.C., Sharp, P.A., et al. (2003). Lentivirus-delivered stable gene silencing by RNAi in primary cells. RNA *9*, 493–501.

Strobl, C., Boulesteix, A.L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. BMC Bioinformatics *9*, 307.

Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A., and Luscombe, N.M. (2009). A census of human transcription factors: Function, expression and evolution. Nat. Rev. Genet. *10*, 252–263.

Wagner, A.H., Coffman, A.C., Ainscough, B.J., Spies, N.C., Skidmore, Z.L., Campbell, K.M., Krysiak, K., Pan, D., McMichael, J.F., Eldred, J.M., et al. (2016). DGIdb 2.0: Mining clinically relevant drug-gene interactions. Nucleic Acids Res. *44*(D1), D1036–D1044.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.

Wang, T., Birsoy, K., Hughes, N.W., Krupczak, K.M., Post, Y., Wei, J.J., Lander, E.S., and Sabatini, D.M. (2015). Identification and characterization of essential genes in the human genome. Science *350*, 1096–1101.

Weir, B.A., Woo, M.S., Getz, G., Perner, S., Ding, L., Beroukhim, R., Lin, W.M., Province, M.A., Kraja, A., Johnson, L.A., et al. (2007). Characterizing the cancer genome in lung adenocarcinoma. Nature *450*, 893–898.

Wilson, B.G., Helming, K.C., Wang, X., Kim, Y., Vazquez, F., Jagani, Z., Hahn, W.C., and Roberts, C.W. (2014). Residual complexes containing SMARCA2 (BRM) underlie the oncogenic drive of SMARCA4 (BRG1) mutation. Mol. Cell. Biol. *34*, 1136–1144.

Yilmazel, B., Hu, Y., Sigoillot, F., Smith, J.A., Shamu, C.E., Perrimon, N., and Mohr, S.E. (2014). Online GESS: Prediction of miRNA-like off-target effects in large-scale RNAi screen data by seed region analysis. BMC Bioinformatics *15*, 192.

Zhang, X.D., Santini, F., Lacson, R., Marine, S.D., Wu, Q., Benetti, L., Yang, R., McCampbell, A., Berger, J.P., Toolan, D.M., et al. (2011). cSSMD: Assessing collective activity for addressing off-target effects in genome-scale RNA interference screens. Bioinformatics *27*, 2775–2781.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Mouse monoclonal Ubiquitin	Cell Signaling	Cat#3936, RRID: AB_331292
Monoclonal rabbit GAPDH	Cell Signaling	Cat#5174, RRID: AB_10622025
Anti-mouse IR secondary antibody	LICOR	Cat#926-68020, RRID: AB_10706161
Anti-rabbit IR secondary antibody	LICOR	Cat#926-32210, RRID: AB_621842
Chemicals, Peptides, and Recombinant Proteins		
Puromycin	Sigma-Aldrich	Cat#P9620
Blasticidin	Life Technologies	Cat#A11139-03
Polybrene	Sigma-Aldrich	Cat#H9268
RIPA Lysis Buffer	Sigma-Aldrich	Cat#R0278-500ML
EDTA-free Protease Inhibitor Cocktail	Roche	Cat#11873580001
Sodium Orthovanadate	New England BioLabs	Cat#P0758
Sodium Fluoride	New England Biolabs	Cat#P0759
MG-132	Selleck Chemicals	Cat#S2619
Critical Commercial Assays		
Cell Titer Glo	Promega	Cat#G7573
Pierce BCA assay kit	Thermo Fisher Scientific	Cat#23225
Thermo Fisher Superscript III First-Strand Synthesis System	Thermo Fisher Scientific	Cat#18080051
Power SYBR Green Master Mix	Thermo Fisher Scientific	Cat#4368706
Deposited Data		
shRNA log fold change values for 216 cell lines screened with 54k library	(Cowley et al., 2014)	https://portals.broadinstitute.org/achilles v2.4.6
shRNA log fold change values for 285 cell lines screened with 98k library	This paper	https://portals.broadinstitute.org/achilles v2.19.2
DNA copy number assayed by Affymetrix SNP6.0 arrays	(Barretina et al., 2012)	https://portals.broadinstitute.org/ccle
Gene expression data for CCLE lines derived from RNaseq data	This paper	https://portals.broadinstitute.org/ccle
Mutation calls for CCLE lines derived from RNaseq data	This paper	https://portals.broadinstitute.org/ccle
DEMETER gene solutions	This paper	https://depmap.org/rnai
DEMETER seed solutions	This paper	https://depmap.org/rnai
shRNA performance scores	This paper	https://depmap.org/rnai
List of transcription factors	(Vaquerizas et al., 2009)	https://doi.org/10.1038/nrg2538 Table S2
List of druggable genes from DGIDB	(Wagner et al., 2016)	http://dgidb.genome.wustl.edu/
List of druggable genes from IUPHAR/BPS Guide to PHARMACOLOGY	(Southan et al., 2016)	http://www.guidetopharmacology.org/
Protein classifications from PantherDB	(Mi et al., 2016)	http://www.pantherdb.org/
Experimental Models: Cell Lines		
See Table S1 for a list of cell lines	N/A	N/A
Oligonucleotides		
Sequencing Primer (5′ with barcode) AATGATACGGCGACCAC CGACCGTAACTTGAAAGTATTTCGATTTCTTGGCTTTATATAT CNNNNNNAAAGG*A*C	(Cowley et al., 2014)	N/A
Sequencing Primer (3′) CAAGCAGAAGACGGCATACGAGCTCTTCCGATCTTGTGGATG AATACTGCCATTTGTCTC	(Cowley et al., 2014)	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Custom sequencing Primer GAGAAAGTATTTCGATTTCTTGGCTTTATATATCTTGTGGA	(Cowley et al., 2014)	N/A
PCR *UBB* Forward GGTCCTGCGTCTGAGAGGT	This paper	N/A
PCR *UBB* Reverse GCCTTCACATTTTCGATGGTGT	This paper	N/A
PCR *UBC* Forward GGAGCCGAGTGACACCATTG	This paper	N/A
PCR *UBC* Reverse CAGGGTACGACCATCTTCCAG	This paper	N/A
PCR *ACTB* Forward CGCGAGAAGATGACCCAGAT	(Brenner et al., 2011)	N/A
PCR 2 *ACTB* Reverse GAGTCCATCACGATGCCAGT	(Brenner et al., 2011)	N/A
Recombinant DNA		
psPAX2	Didier Trono	Addgene 12260
pCMV-VSVG	(Stewart et al., 2003)	Addgene 8454
pLKO_047 (GFP expressing only)	GPP, Broad Institute	pLKO_047
pLKO_TRC005	GPP, Broad Institute	pLKO_TRC005
pLKO_TRC005-nullT	GPP, Broad Institute	TRCN0000241923
pLX-TRC304-UBB(Ubiquitin-V5)	GPP, Broad Institute	ccsbBroad304_14873
shUBB-1	GPP, Broad Institute	TRCN0000011102
shUBB-2	GPP, Broad Institute	TRCN0000007735
shUBC-1	GPP, Broad Institute	TRCN0000011109, TRCN0000011107
hUBC-3	GPP, Broad Institute	TRCN0000011111
shUBC-4	GPP, Broad Institute	TRCN0000011110
shUBC-5	GPP, Broad Institute	TRCN0000423348
shUBC-6	GPP, Broad Institute	TRCN0000426019
shUBC-7	GPP, Broad Institute	TRCN0000011108
shPSMD2	GPP, Broad Institute	TRCN0000066072
C911-shUBC-1 CGAGAACCAGAAAGCAAAGAT	This paper	N/A
C911-shUBC-3 GAGGTTGTAGTTTGCCGGAAA	This paper	N/A
C911-shUBC-4 AGGTTGAAGATTGCTGGGAAA	This paper	N/A
C911-shUBC-7 GCAAAGAAGGAAGACAAGGAA	This paper	N/A
shGFP	GPP, Broad Institute	TRCN0000072181
shLuciferase	GPP, Broad Institute	TRCN0000072256
shRFP	GPP, Broad Institute	TRCN0000072209
Software and Algorithms		
DEMETER	This paper	https://github.com/cancerdatasci/demeter
ATLANTIS	This paper	https://github.com/cancerdatasci/atlantis

CONTACT FOR REAGENTS AND RESOURCE SHARING

As Lead Contact, William Hahn (william_hahn@dfci.harvard.edu) is responsible for all reagent and resource requests.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell lines

Cell lines were obtained from the Cancer Cell Line Encyclopedia (https://portals.broadinstitute.org/ccle/home) unless otherwise indicated. Cell line information, including source is listed in Table S1. Information on tissue, tumor type and growth media conditions, (used to grow the cells and also for screening) were obtained from the CCLE project or source laboratory and are listed in Table S1. All cell lines were fingerprinted multiple times using one of two genotyping platforms, Sequenom or Fluidigm.

METHODS DETAILS

Screening and deconvolution using next-generation sequencing

We extended our previous study of 216 cell lines (Cowley et al., 2014) by performing genome-wide pooled loss of function screening on additional 285 cancer cell lines across approximately 100k shRNAs (final files include 107,523 shRNA values in Achilles_v2.19.2 to produce 17,098 DEMETER gene solutions in Achilles_v2.20.2). Each cell line was infected with the shRNA pool by lentivirus, in quadruplicate and propagated for at least 16 population doublings or 40 days, whichever came first. To determine the viral volume needed to achieve the desired transduction rate of ~40%, each cell line was titrated with 6 volumes of virus (0-500 uL) in a 12 well plate at a concentration of 3E6 cells/well. Then cells were cultured in the presence or absence of puromycin in 6 well dishes before infection rates were determined. Cells were expanded for infection in quadruplicate with a target of 3.7E7 infected cells. Before infection, cells were filtered through a 40 um cell strainer to remove clumps, then resuspended in media containing 4 ug/ml polybrene, and the appropriate volume of 98K library lentivirus to achieve a cell concentration of 1.5E6 cells/ml. This cell suspension was seeded into 12 well plates at 2 mL/well and centrifuged for 2 hr at 930xg at 30°C. After the spin infection, 2 mL of fresh media was added to each well. After 24 hr, the cells from each replicate infection were pooled into T225 flasks with 60ml medium containing puromycin. To provide an in-line assessment of transduction rate, 150k of infected and uninfected cells were cultured in 6 well dishes in the presence or absence of puromycin. After 96 hr, both the in-line assay wells and the screen replicates were trypsinized. The infection rate was determined by calculating the number of viable cells selected in puromycin divided by the number of viable cells without puromycin selection.

Screening was continued if the infection rates were within the range of 30%–65% so that the selected cells were nearly all MOI = 1 and so that there was a sufficient number of cells to provide adequate representation of each shRNA. For each of the replicates, 6E7 cells were plated into new T225 flasks in 60ml of media with puromycin. For the remaining passages, only 3E7 cells per replicate were carried over, and the remaining cells were spun down and resuspended in PBS for genomic DNA isolation. Passaging for each cell line was continued for at least 16 population doublings or 28 days, whichever was longer. Puromycin selection was maintained until day 7. At the end of passaging, genomic DNA from the screen endpoints were used to measure the abundance of shRNAs in comparison to the initial DNA plasmid pool. Samples were sequenced using a custom sequencing primer using standard Illumina conditions. Deconvolution was performed similar to that described in Ashton et al. (Ashton et al., 2012) and all steps are described more completely in Cowley et al. (Cowley et al., 2014), with the following alterations. A total of 280 μg gDNA was used as template for PCR from each replicate. Thermal cycler PCR conditions consisted of heating samples to 95 °C for 5 min; 28 cycles of 94 °C for 30 s, 53 °C for 30 s, and 72 °C for 20 s; and 72 °C for 10 min. PCR reactions were then pooled per sample. After PCR and additional of sample barcodes, 20 replicates were multiplexed into a single Illumina sample, and run on multiple lanes to achieve a minimum of 2^7 reads per replicate. PCR sequences are listed in Key Resources Table. Cell line specific information is listed in Table S1. Cell doubling time was calculated from the lentivirally infected cells during the course of the screens. Days in culture represent the days from the day of infection until the date of the harvest. Passage number represents the number of cell splits during during the screen and refer to the time point of the sample that was used for data collection specific to each cell line.

Cloning of C911 shRNAs

C911 shRNAs were designed by changing the nucleotides at positions 9 through 11 of the corresponding experimental shRNA to their complement base and appending an AgeI recognition site at the 5′ end and an EcoRI recognition site at the 3′ end with appropriate overhang sequences. Oligonucleotides were purchased from Integrated DNA Technologies. Complementary oligos were annealed and ligated to the pLKO_TRC005 vector cut with restriction enzymes AgeI and EcoRI. Ligation products were transformed into DH5a chemically competent cells (Invitrogen) according to manufacturer's instructions and plated on agar plates containing 100ug/mL carbenicillin incubated for 16 hr at 37°C. Single colonies were used for DNA preparation (QIAGEN). All clones were verified by sequencing.

Viral production

293T cells were seeded in 96 well plates at $2.2*10^4$ per well (100uL volume) 24 hr pre-transfection. Transfection was performed using TransIT-LTI Transfection Reagent (Mirus). Briefly, two solutions were prepared in different 96-well plates for each construct. One solution contained 0.6uL of LT1 diluted in 10uL of Opti-Mem (Corning) for each well incubated at room temperature for 5 min. For the second solution, a master mix that contained 100ng/well psPAX2 (Addgene 12260), 10ng/well pCMV-VSVG (Addgene 8454), and Opti-MEM for a total volume of 10uL/well was added to a plate that contained 100ng of the transfer vector diluted in 10uL of sterile

water. The two final solutions were combined and incubated at room temperature for 30 min. The transfection mixture was then added to the plate of cells and incubated at standard cell culture conditions (37°C, 5% CO2) until the following morning. At least 18 hr post transfection, media on the cells was changed to 170uL high-BSA growth media (DMEM + 10% FBS + 1% BSA). Virus was harvested 24 hr after the media change, the media was replenished, and a second harvest occurred at 48 hr after the media change. Virus from both harvests was pooled, aliquoted, and stored at −80°C until use in the experiments.

GFP competition assay

All infections were performed by centrifuging freshly seeded plates containing cells with lentiviral particles and 4μg/mL polybrene for two hours at 2000 rpm.

Cell lines stably expressing GFP were generated using a lentiviral expression vector (pLKO_047). shRNAs were introduced to non-GFP expressing cells in duplicate and selected for 2-3 days with 3-6μg/ml of puromycin before starting the co-culture. Co-cultures were created by mixing GFP expressing cells with shRNA-infected non-GFP cells at a ratio of 75 GFP negative to 25 GFP positive. Time-points quantifying the ratio of GFP to non-GFP population were taken using flow cytometry (BD Biosciences BD Accuri C6) each time the co-culture was split (every three to four days) for 9-12 days post selection. Log2 fold change of percentage GFP negative cells remaining for each experimental construct compared to the average of the percentage GFP negative in negative controls (pLKO_TRC005-nullT, shGFP, shRFP, shLuciferase) was calculated for each time point. Since different cell lines grow at different rates, for comparison between cell lines the time-point of maximal depletion (median of shUBC-1, 3 and 7) was selected per cell line. Results are representative of two independent experiments.

UBB rescue experiments

An exogenous ORF fragment from UBB (NM_018955.3, 844-1083) encoding for ubiquitin-V5 (ccsbBroad304_14873) was overexpressed in SNU8 cells using lentivirus. Ubiquitin overexpressing or parental cells were seeded in a 96-well plate at 1000 cells/well and infected on the same day with lentivirus expressing shUBC, shGFP, shPSMD2, shRPS6 or pLKO_TRC005 nullT in individual wells. Viability was measured 24h after infection and every 48h over a 7 day time-course using CellTiterGlo (Promega) on a Perkin Elmer EnVision. Three separate infection replicates were used for each time point. Average raw luminescent signal for each condition was normalized to the average of the pLKO_TRC005 nullT signal. Fold-change to day 1 was calculated from the normalized signal. Data are representative of two independent experiments.

Western Blots

Cells were infected with lentivirus expressing shUBC-3, shUBC-7, or pLKO_TRC005 nullT and selected with puromycin at a concentration of 4ug/mL for 48-72hr or until all uninfected cells were dead. Cells were stored as pellets at −80°C. MG-132 treated samples were incubated with 50μM MG-132 for 2 hr prior to harvest. Whole cell lysates were prepared using RIPA buffer (Sigma-Aldrich) supplemented with EDTA-free Protease Inhibitor Cocktail (Roche), 1mM Sodium Orthovanadate (NEB), and 5mM Sodium Fluoride (NEB). Protein levels were quantified using the Pierce BCA assay kit (Thermo Fisher Scientific #23225). Immunoblots were run using 4%–12% Bis-Tris Pre-Cast gels (Thermo Fisher Scientific NuPAGE Novex #NP0335) and transferred to a membrane using the iBlot 2 system (Thermo Fisher Scientific). Ubiquitin levels were detected using a monoclonal mouse anti-Ubiquitin Antibody at 1:1000 dilution (Cell Signaling P4D1 #3936) and a LICOR-compatible anti-mouse IR secondary antibody (LICOR #926-68020) at 1:5000 dilution. GAPDH levels were detected using a monoclonal rabbit GAPDH antibody (Cell Signaling 14C10 #2118) at 1:1000 and a LICOR-compatible anti-rabbit IR secondary antibody (#926-32211) at 1:5000 dilution. Western blots shown are representative of two independent experiments.

RT-PCR

COV434 cells were infected with lentivirus expressing shRNAs targeting UBB, UBC or shLuciferase and selected with puromycin at a concentration of 4ug/mL for 48 hr. Cells were stored as pellets at −80°C. Total RNA was isolated using QIAGEN RNeasy Plus Mini Kits (QIAGEN #74134). Reverse transcription for RNA samples was performed using Thermo Fisher Superscript III First-Strand Synthesis System (Thermo Fisher #18080-051). RT-PCR was performed on the QuantStudio 6 Flex (Applied Biosystems) using Thermo Fisher Power SYBR Green Master Mix (Thermo Fisher # 4367659) with probes against UBB, UBC and Actin (See Key Resources Table). Each measurement was taken in triplicate. Comparative CT (Delta Delta CT) was used for quantification analysis. Actin was used as reference for normalization. Results are representative of two independent experiments.

DEMETER

The main goal of DEMETER is to infer gene knockdown viability effects (“gene dependency scores”) for each gene and cell line screened by an shRNA (or siRNA) library containing multiple reagents designed to target the same gene. Given the observed phenotypic effects produced by shRNAs and knowledge of which shRNAs share a common ‘seed sequence’ and which target a common gene, DEMETER deconvolves the effects of each shRNA into a linear combination of the effects due to knockdown of the target gene and the effects associated with the seed sequence. In addition, we expect a batch effect due to variation in the initial abundance of shRNA in each library. We remove that batch effect by modeling those gene and seed effects as relative to the mean for each batch.

We assign two seed sequences to each shRNA – positions 1-7 and 2-8 on the antisense strand (corresponding to positions 12-18 and 11-17 on the sense strand). These two regions were chosen as those that maximized intra-group correlation of fold-change

depletion when grouping the shRNAs by any 7-mer subsequence (Figure S1C). The seed sequences present in shRNA i are denoted as $seed(i)$. Similarly, we assign one or more genes targeted by each shRNA by aligning the sequence to the reference genome. The genes targeted by shRNA i are denoted as $gene(i)$.

Given a dataset consisting of p shRNAs and n cell lines, we define an observation matrix H, where each element H_{ij} represents the readout resulting from perturbing cell line j ($j = 1, 2, \ldots, n$) by shRNA i ($i = 1, 2, \ldots, p$). We decompose H_{ij} into, G_{lj}, the effect of knocking down gene l in cell line j, and S_{kj}, the effect of an shRNA with seed k on cell line j. Both effects are relative to the mean readout for shRNA i within each batch b_j, denoted as μ_{ib_j}. Relative effects were sufficient because we focused on discovering differential dependencies. Non-differential dependencies have the potential to be generally essential and non-selective.

Formally, the DEMETER model for each observed data point H_{ij} is defined as:

$$H_{ij} = \sum_{k \in seed(i)} \alpha_{ik} S_{kj} + \sum_{l \in gene(i)} \beta_{il} G_{lj} + \mu_{ib_j} + \varepsilon$$

subject to

$$0 \leq \alpha_{ik},\ \beta_{il} \leq 1$$

In addition to the effects discussed above, the coefficients α_{ik} and β_{il} scale the seed effect, S_{kj}, of seed k on cell line j and the gene effect, G_{lj}, of gene l on cell line j for the specific shRNA i.

We only fit gene, G_{kj}, and seed effects, S_{kj}, supported by two or more measurements. We explicitly remove those corresponding G_{kj} and S_{kj} terms from the objective function that are only used to compute a single H_{ij}. This can occur when a gene or seed is supported by a single shRNA or when all but one shRNA for that gene in the cell line are missing values. Additionally, H may have missing values for an shRNA across all cell lines screened in a particular library due to that shRNA being only included in another library.

After all parameters have been fit, we make gene effects comparable to one another by dividing G_{lj} by $\max_i \beta_{il} | l \in gene(i)$. Since the objective function only includes the product of $\beta_{il} G_{lj}$, and not G_{lj} we can apply an arbitrary scale to β_{il} as long as we also divide G_{lj} by that scale. As a result, the scaled elements in β can be thought of as the strength of the gene effect relative to the shRNA with the strongest gene effect.

The objective function

To fit the parameters for this model, we formulate the following optimization problem:

$$\min_{S,G,\alpha,\beta,\mu} \sum_{ij} \left(H_{ij} - \hat{H}_{ij} \right)^2 + p_{reg} + p_{con}$$

where $\hat{H}_{ij}$ is the prediction of the effect of perturbing cell line j by reagent i:

$$\hat{H}_{ij} = \sum_{k \in seed(i)} \alpha_{ik} S_{kj} + \sum_{l \in gene(i)} \beta_{il} G_{lj} + \mu_{ib_j}$$

We regularize the model parameters by the penalty p_{reg}:

$$p_{reg} = \lambda_s \sum_{kj} (S_{kj})^2 + \lambda_g \sum_{kj} (G_{kj})^2 + \lambda_\alpha \sum_{ik} (\alpha_{ik})^2 + \lambda_\beta \sum_{il} (\beta_{il})^2$$

and penalize by p_{con} to enforce constraints $\alpha \geq 0$, $\beta \geq 0$.

$$p_{con} = -\lambda_p \sum_i \min(0, \alpha_i) - \lambda_p \sum_i \min(0, \beta_i)$$

Stochastic gradient descent was used to minimize the objective function.

Initial solution for gradient descent

To determine the initial parameter values from which the gradient descent starts, we compute $\overline{\mu}_b$ as the mean of all measurements for cell lines in batch b.

$$\overline{\mu}_{ib} = \operatorname*{mean}_j H_{ij} \,\Big|\, j \in b$$

Then, S and G are computed as the marginal means of H after subtracting $\overline{\mu}_{b_j}$ where $\overline{\mu}_{b_j}$ is the mean for the batch that contains cell line j.

$$\overline{S}_{kj} = \operatorname*{mean}_i H_{ij} - \overline{\mu}_{ib_j} | k \in seed(i)$$

And

$$\overline{G}_{lj} = \operatorname*{mean}_k H_{ij} - \overline{\mu}_{ib_j} | l \in gene(i)$$

Finally, to determine an initial α and β, we fit the linear model for each shRNA i across all cell lines:

$$H_{ij} \sim \sum_{k \in seed(i)} \alpha_{ik} \overline{S}_{kj} + \sum_{l \in gene(i)} \beta_{il} \overline{G}_{lj} + \overline{\mu}_{ib_j}$$

Update step for stochastic gradient descent

Computing the gradient for a given H_{ij} we get:

$$\varepsilon = H_{ij} - \widehat{H}_{ij}$$

$$\nabla_{G_{ki}} = 2\left(\lambda_g G_{ki} - \epsilon \beta_{ik}\right)$$

$$\nabla_{S_{ki}} = 2\left(\lambda_s S_{ki} - \epsilon \alpha_{ik}\right)$$

$$\nabla_{\alpha_{ik}} = 2\left(\lambda_\alpha \alpha_{ik} - \epsilon S_{ki}\right)$$

$$\nabla_{\beta_{ik}} = 2\left(2\lambda_\beta \beta_{ik} - \epsilon G_{ki}\right)$$

$$\nabla_\mu = -2\epsilon$$

We update each parameter by the gradient, scaling by a learning rate γ:

$$x_{n+1} = x_n - \gamma \nabla_x$$

We iterate through the elements H_{ij} in random order, performing the update for each element. We chose a learning rate of $\gamma = 0.005$ for all parameters. To strongly discourage constraint violations we set $\lambda_p = 10$. To choose the remaining hyperparameters, we randomly sampled parameters, and chooses those that minimized the mean out-of-sample RMSE based on three rounds of cross validation, where 1% of the elements in H_{ij} are held out in each round. After the hyperparameters were chosen, we re-ran DEMETER on all of the data, iterating through the elements in H_{ij} the same number of passes required to achieve the minimum out-of-sample RMSE during the cross-validation procedure.

Assessing shRNA performance

We assess individual shRNA performance by looking at the variance explained by the contribution of the gene effect and seed effect per shRNA. We computed the variance explained, $R^2(y, f) = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \overline{y}_i)^2}$, using only the contribution of the seed or gene to predict the observed values. That is to say s_c and g_c, the shRNA's seed effect and gene effect contribution respectively were computed as:

$$s_c = R^2\left(H_{ij}, \sum_k^{seed(i)} \alpha_{kj} S_{ki} + \mu_{ib_j}\right), \quad g_c = R^2\left(H_{ij}, \sum_k^{gene(i)} \beta_{ik} \overline{G}_{kj} + \mu_{ib_j}\right)$$

Data processing pipeline

Raw Illumina reads were normalized across replicates to alleviate the variable read depth of each replicate. Normalized shRNA value = $\log_2$([(Raw read value for shRNA)/(Total raw read value for replicate) x $1e^6$] +1). Normalized and $\log_2$ transformed read counts were processed in a GenePattern pipeline separately each shRNA library dataset, starting with modules that remove undesirable shRNAs and failing QC replicate samples ('FilterLowshRNAs', 'shRNAremoveOverlap' and 'removeSamples'). Fold change values are next calculated ('shRNAfoldChange') using an appropriate pDNA reference sample, based on both shRNA library (55k, 98k) and sequencing chemistry kits (cBotV7/sbsv2, cBOTv8/sbsv3) and then quantile normalized per replicate cell line ('NormLines'). Replicate cell lines values are then collapsed ('shRNAcollapseReps') and shRNAs are mapped to the newest gene transcriptome mapping / HUGO gene symbols ('shRNAmapGenes'). The previous 55k library data (Achilles_2.4.3, (Cowley et al., 2014)) was also remapped using these newest gene mappings and subsequently renamed Achilles_v2.4.6. Gene summarization was performed using the DEMETER algorithm (next section), which also combined the data from each shRNA library dataset (55k library: Achilles_v2.4.6, 216 cell lines and 98k library: Achilles_v2.19.2, 285 cell lines) to produce the final gene level data (Achilles_v2.20.2, 501 cell lines). All steps, including quality control steps and sample fingerprinting are described in detail in Cowley et al. (Cowley et al., 2014) and GP modules are available from the GenePattern Archive: http://gparc.org/. Data can be downloaded from the Project Achilles Portal (https://portals.broadinstitute.org/achilles).

Applying DEMETER to 501 RNAi screens

DEMETER was run separately on the Achilles data divided into three batches: the Achilles v2.4.6 lines divided into a batch for cell lines processed with cBotV7/sbsv2 kits and a batch for the cBOTv8/sbsv3 kit, and a final batch containing all of the lines comprising Achilles v2.19.2.

Sixty-five shRNAs, those targeting more than 10 genes, were removed because we suspected the interactions would be too complex to derive meaningful information from those shRNAs. Those shRNAs whose gene label starts with "NO_CURRENT" are not known to target any gene, but are present in the library due to the reference genome changing after the library was designed. Even without a targeted gene, these shRNAs were included because they contributed to the estimation of seed effects.

A pair of genes targeted by identical shRNAs cannot be distinguished from one another and left untreated would result in half of the total gene effect being attributed each gene. Therefore, we created a "gene family" for which the total effect is derived. Overall, 399 genes were collapsed into 172 such families. After the deconvolution was complete, the estimated effect for a gene family was reported for each gene in the family.

Next, hyperparameter optimization was performed by random search and $\lambda_\alpha = \lambda_\beta = 0.9$ and $\lambda_g = \lambda_s = 4e-5$ were chosen. These parameters achieved a mean out-of-sample RMSE of 0.67 and in-sample RMSE of 0.53.

Afterward, DEMETER was run on the full data, without holding any data out, resulted in an in-sample RMSE of 0.54. DEMETER next transformed elements in G into z-scores using the global mean and standard deviation of G. The final set of z-scores values was obtained after expanding the gene families and removing the records corresponding to labels prefixed with "NO_CURRENT." In addition, performance metrics for each shRNA are summarized in table S2.

Benchmarking DEMETER against ATARiS

In comparing the performance of DEMETER and ATARiS, we limited ourselves to data which could be processed by both methods. ATARiS does not support multiple batches, so we only used data from largest batch, the Achilles 98k library containing 285 cell lines. Also, ATARiS does not produce a gene solution for every gene, so we limited ourselves to the 9,348 genes that had a solution from ATARiS. If ATARiS produced multiple solutions, only the first solution was considered.

We assume that knocking down genes participating in the same protein complex should be enriched for similar dependency profiles. The CORUM database was used to to associate 2,505 genes with 1,749 protein complexes. Separately for ATARiS and DEMETER, we computed the distribution of Pearson correlation coefficients between pairs of profiles from genes that participated in the same protein complex. Then, to compare the two distributions, we normalized by z-scoring the correlations, using the standard deviation from the distribution of correlations between random pairs of profiles (fig. S1A).

Correlation of dependency profiles (Figures 1E-G and S1B)

Pearson correlations of DEMETER gene dependency scores were computed across cell lines (N = 501) for all pairs of variable genes that share overlap in cell lines (N = 6,300). The resulting gene similarity matrix was converted to a discrete adjacency matrix by converting correlation coefficients to standard scores and adding edges only between pairs of genes with standard scores ≥ 3. The networks in Figures 1E-G show the connected neighbors of a selected gene. The heatmap in Figure 1E shows DEMETER gene scores as colors/values, but only genes connected to PIK3CA in the adjacency matrix are shown and ordered by decreasing correlation coefficient.

Differential dependencies and 6σ dependencies

The 17,098 unique genes in the DEMETER dataset were filtered for genes for which at least one cell line's dependency score is -2 or below and expression of the gene in the most dependent cell line is above –2 $\log_2$ RPKM, resulting in 6,305 dependency profiles representing potential differential dependencies. Of these, 6σ dependencies were defined as genes where at least one cell line is dependent on them at a level of six "global" standard deviations (i.e., computed using scores for all genes in all cell lines) from the mean of each gene. This resulted in 769 6σ dependencies.

RNASeq

Library construction and sequencing

RNA sequencing: library construction and sequencing Non-strand specific RNA sequencing was performed using large-scale, automated variant of the Illumina TruSeq RNA Sample Preparation protocol. Oligo dT beads were used to select polyadenylated mRNA. The selected RNA was then heat fragmented and randomly primed before cDNA synthesis. To maximize power to detect fusions insert size of fragments was set to 400nt. The resultant cDNA then went through Illumina library preparation (end-repair, base 'A' addition, adaptor ligation, and enrichment) using Broad designed indexed adapters for multiplexing. Sequencing was performed on the Illumina HiSeq 2000 or HiSeq 2500 instruments, with sequence coverage of no less than 100 million paired 101 nucleotides-long reads per sample.

Expression data analysis

RNaseq reads were aligned to the B37 version of human genome using TopHat version 1.4. Gene and exon-level RPKM values were calculated using pipeline developed for the GTEx project (https://gtexportal.org/home/, (DeLuca et al., 2012)).

Calling substitutions

Variant calling and annotation: Nucleotide substitutions were detected with MuTect (Cibulskis et al., 2013) (http://www.broadinstitute.org/cancer/cga/MuTect). MuTect program was run in the mode that does not require matching normal DNA and thus identifies all variants that differ from a reference genome. Variants were annotated using the Oncotator (Ramos et al., 2015) and AnnoVar software (Wang et al., 2010) (http://annovar.openbioinformatics.org).

Variant filtration

The allelic fraction was calculated for each detected variant per cell line as a fraction of reads that supported an alternative allele (e.g., different from the reference) among reads overlapping the position. Only reads with allelic fractions above 0.25 were used in the downstream sensitivity prediction analysis.

Variant filtration by exclusion of common germline variants: Variants for which the global allele frequency (GAF) in dbSNP134 or allele frequency in the NHLBI Exome Sequencing Project (http://evs.gs.washington.edu/EVS, data release ESP2500) was higher than 0.1% were excluded from further analysis.

Variant filtration by exclusion of variants observed in a panel of normals: Variants detected in a panel of 278 whole exomes sequenced at the Broad as part of the 1000 Genomes Project were excluded from further analysis. Beyond removal of additional germline variation, this step also allowed elimination of common false positives that originate predominantly from alignment artifacts.

Calling indels

For indel calling RNASeq data were realigned using STAR (Dobin et al., 2013) and indels were called using Strelka (Saunders et al., 2012).

ATLANTIS

We developed ATLANTIS, a nonlinear regression modeling method, to find molecular markers that are predictive of DEMETER dependency scores. The predictive features were derived from CCLE's molecular characterization of the cell lines and the target learned was the dependency scores reported by DEMETER. ATLANTIS, our tool for finding and characterizing predictive biomarker-dependency models uses the R package "party" to build an ensemble of conditional inference (Strobl et al., 2008). This method was chosen for its ability to capture nonlinear relationships, accommodating both categorical and continuous features in the same model, and its ability to accommodate missing values.

After learning a model with ATLANTIS, we record the out-of-bag weighted R^2 as the goodness-of-fit metric. We next prune the feature list used by that model to present a shorter list of candidate biomarkers. First, we compute the variable importance using the party package's "varImp" function for each feature used in the model. To prune poorly chosen features, we drop any features whose variable importance was either negative or absolute variable importance was in the bottom 0.01 quantile. We then train a new model, using only those features remaining, and again do another round of pruning dropping only features with a negative variable importance. The remaining features are reported along with their final variable importance in the ATLANTIS reports.

Compensating for few dependent lines

We were most interested in ATLANTIS capturing the difference between dependent and insensitive lines. However, it was difficult to model as a classification problem when we did have a clear threshold on dependency score which we could use to define the dependent and insensitive classes. Also, there may be times where we might be able to predict the variance in the sensitive class, so we opted to instead keep it as a regression problem, but refer to lines whose z-scored dependency score is less than -2 as "dependent." At -2 standard deviations from the mean, we may have some lines that are within the noise around the mean and not truly dependent, but we expect those lines are at least enriched for truly dependent lines.

The dependent lines were a small fraction of the lines assayed for each gene, but were demonstrating the behavior we wanted to predict. To encourage the model to distinguish between "dependent" and "nondependent" lines, we biased the sampling when selecting samples to build each tree to enrich for dependent lines. First, we sampled the potentially dependent lines, those with a dependency score < -2, picking each with a probability of 80%. Then the remaining samples were uniformly sampled from the non-dependent lines. Even after biasing the sampling, the "dependent" lines were far fewer than "nondependent" lines in the training set for each tree, so we used non-uniform weighting to make the two classes more balanced. Weights for each sample were assigned to the dependent and the non-dependent cell lines such that the sum of weights were equal for both classes, but capping the maximum weight of any one line at 5%.

To improve runtime and avoid pathological splits, the smallest bucket the tree was allowed to be three times the weight of a single dependent line. For each model, we removed any features consisting of a single distinct value for all, or all but one of the cell lines. In addition, we dropped any cell lines missing values for all features. Once this pre-filtering was complete, the decision tree ensemble was constructed by the "cforest" method in the "party" R package.

We assessed the goodness-of-fit of each model by computing the square of the out-of-bag weighted Pearson correlation coefficient. However, any model with a negative weighted correlation was given a score of 0. To compute p values testing whether the model's goodness-of-fit could have arisen by chance, a global null distribution was computed by 50k iterations of selecting a random gene, shuffling the dependency scores, and fitting and scoring a model with the procedure described above. Finally, to correct for multiple hypothesis testing, q-values were computed from the p values across all models fit for a given MDP class via Benjamini & Hochberg's method.

Identifying dependency classes (Table S6)

Mutation-driven dependencies

To identify putative oncogene addictions, we considered any hotspot mutations, missense mutations, and the copy number of the gene whose sensitivity we were modeling. Those genes whose model had the best biomarker negatively correlated with the dependency score were classified as putative oncogene addictions. To avoid ATLANTIS modeling any of the variation in the non-dependent portion of the distribution, we additionally generated a second model based on replacing all sensitivities > −2 with zero. In addition, we enforced that each tree could only threshold on a feature at most once per leaf to only capture behavior for extremes of a feature, and avoid modeling a prediction for an interval of a feature.

We note that many p53 wild-type cell lines gain a growth advantage following *TP53* suppression, leading *TP53* to be identified as a mutation-driven dependency (as mutated cell lines show stronger "dependency" compared to wild-type ones). We have therefore manually excluded *TP53* from this MDP class.

Expression-driven dependencies

The same method was used to identify gene addictions, with the exception that also gene expression was considered as a potential predictive feature. Those models that also had the strongest biomarker negatively correlated with the dependency score were classified as gene addictions.

CYCLOPS

For CYCLOPS, the gene expression and copy number of the modeled gene were used as predictive features. We continued to only allow a single split per feature, but only ran ATLANTIS once, predicting the gene's dependency scores. Among those models, those where the best biomarker was positively correlated with the dependency score prediction were classified as CYCLOPS.

Paralog deficiency dependencies

To identify instances where a gene dependency emerges due to loss of function of a paralogous gene, we run ATLANTIS using missense and damaging mutations, copy number and gene expression of all genes which were reported as sequence paralogs by GenesLikeMe. Again, here we produce two models, one with the original dependency data and one with values > −2 replaced with zero.

We note that RPL17 and RPL17-C18orf32 were identified as a paralog deficiency pair but they in fact represent the same gene and hence we manually excluded them from this MDP class.

Related features

The "related" MDP models were trained by limiting the features based on the gene whose dependency we were trying to predict. For each dependency being predicted, we limit the features only those of genes which were either reported as having a protein-protein interaction according to InWeb with a confidence score greater than 0.1 (Lage et al., 2008; Lage et al., 2007), associated with one another according to GenesLikeMe with a super-pathway score greater than 0.3 (Stelzer et al., 2016), or any gene which shares a complexes with the dependent gene according to CORUM (Ruepp et al., 2010).

We note that the dependency profile of MAP4K4 was removed from these analyses as we found it to suffer from strong off-target (non-seed-based) effects, causing it to mimic the profile of NRAS.

Table summarizing the definitions of the MDP classes

	Features	Models predict
Mutation-driven dependencies	Hotspot mutations,	One model predicts z-scored sensitivity.
	Missense mutations,	One model predicts z-scored sensitivity where values > −2 are replaced with zero
	Copy number	
Expression-driven dependencies	Hotspot mutations,	One model predicts z-scored sensitivity.
	Missense mutations,	One model predicts z-scored sensitivity where values > −2 are replaced with zero
	Copy number,	
	Gene expression	
CYCLOPS	Copy number, Gene expression	One model predicts z-scored sensitivity.
Paralog deficiency dependencies	Missense and damaging mutations, Copy number and Gene expression of all sequence paralog genes.	One model predicts z-scored sensitivity.
		One model predicts z-scored sensitivity where values > −2 are replaced with zero
Related	Missense and damaging mutations, Copy number and Gene expression of associated genes via PPI, CORUM or GenesLikeMe's super-pathways.	One model predicts z-scored sensitivity.
		One model predicts z-scored sensitivity where values > −2 are replaced with zero

Mutation enrichment analysis in mutation-driven dependencies (related to Table S5)

For each gene identified as a potential differential dependency (N = 6,305), cell lines were split into two groups, MUT and WT, based on presence or absence of an RNA missense mutation in the gene. Enrichment p values were calculated by further splitting the MUT and WT groups into dependent and non-dependent groups by discretizing the DEMETER gene scores at a particular threshold and performing a one-sided Fisher Exact test. Instead of using a single threshold of −2, as was done with the lineage enrichment of TF dependencies, a Fisher exact test was performed using the DEMETER score of each MUT cell line, −2 or below, as the dependency threshold. The multiple p values that result per gene from this process were Bonferroni corrected and the most negative threshold with $p < 0.001$ was selected to represent the gene.

A global null was built by performing 10 million permutations of cell line labels and compiling the minimum thresholds given the fisher criteria for all genes. Empirical p values were determined for each gene by counting number of times the null threshold was less than the true threshold for the gene. Empirical p values were corrected using Benjamini Hochberg method.

Lineage enrichment of transcription factor dependencies (related to Figure 4B)

For each lineage context with at least 7 cell lines (N = 20), an enrichment score was computed for dependency on each transcription factor (TF) included in the mutation- and expression-driven MDP classes (N = 49). The enrichment score is calculated by discretizing the DEMETER gene dependency scores (GS) for each TF into dependent (GS ≤ -2) and non-dependent (GS > -2) cell lines. Recall that a GS of −2 represents a dependency that is 2 standard deviations more dependent than the mean across all the cell lines. Dependent and non-dependent groups of cell lines are further split into a two-by-two contingency table based on membership in the specified lineage. p values are assigned to each (TF, lineage) pair based on one-sided Fisher's exact tests and converted to q-values using the Benjamini Hochberg method to correct for multiple hypothesis testing. TFs that are significantly enriched (q-value ≤ 0.05) in a single lineage are labeled 'Specific', whereas TFs that are significantly enriched in multiple lineages are labeled 'Multiple'. The y axis in Figure 4B is an odds ratio (OR), which is calculated as follows:

	Lineage	Non-lineage
Dependent	a	b
Non-dependent	c	d

$$OR = \frac{a+0.5}{c+0.5} \Big/ \frac{b+0.5}{d+0.5}$$

Benchmarking curated dependency-biomarker pairs (related to Figure 6B)

To determine the performance of DEMETER, a curated list of dependency-biomarker pairs was created based on literature reviews and experimental validation. We computed the Pearson correlation coefficients for each marker with each of the 6,305 identified dependency profiles. Dependencies were categorized as (1) Discovered, if the dependency scored in the top 100, (2) Not discovered, if the dependency did not score in the top 100 and could not be explained by having insufficient context, (3) Insufficient context, if the dependency did not score in the top 100 and the marker was a mutation and there were fewer than 3 cell lines with hotspot mutations (4) No differential dependency, if fewer than 3 cell lines with a dependency score of less than −2.

QUANTIFICATION AND STATISTICAL ANALYSIS

GFP competition assay (Figure 5D)

For each cell line (N = 7), mean fraction of GFP negative cells was calculated for UBC hairpins (shUBC: 1,3,7) and negative controls (pLKO_TRC025-nullT, shGFP, shRFP, shLuciferase). shUBC-4 was excluded from this analysis since DEMETER assigned a low gene-score. Values were converted to log2 fold-change of mean UBC targeting hairpins versus mean negative control and the fold-changes were compared between UBB high expressing (N = 3) and UBB low expressing (N = 4) using a one-sided Mann-Whitney test.

UBB rescue (Figure 5E)

The log2 fold-changes for hairpins targeting UBC (excluding shUBC-4 which has low predicted on-target activity) were averaged for each time point (Day: 1,3,5,7) and each group (parental, UBB overexpressed). A linear model was fit to the log2 fold-change response vector using only time point as the predictor (p value = 0.1538, F-statistic). A second model was fit with the additional group variable as a second predictive feature (p value = 0.0262, F-statistic). The additional contribution of the group variable to prediction is measured by comparing the two models using an F-test included in R 'stats' package anova function (v3.2.1).

DATA AND SOFTWARE AVAILABILITY

The shRNA data generated in this study are publically available at https://portals.broadinstitute.org/achilles. All analysis results are available at https://depmap.org/rnai, and code for DEMETER and ATLANTIS is available at https://github.com/cancerdatasci. Cell line molecular features can be downloaded from https://portals.broadinstitute.org/ccle/. See also Key Resources Table.

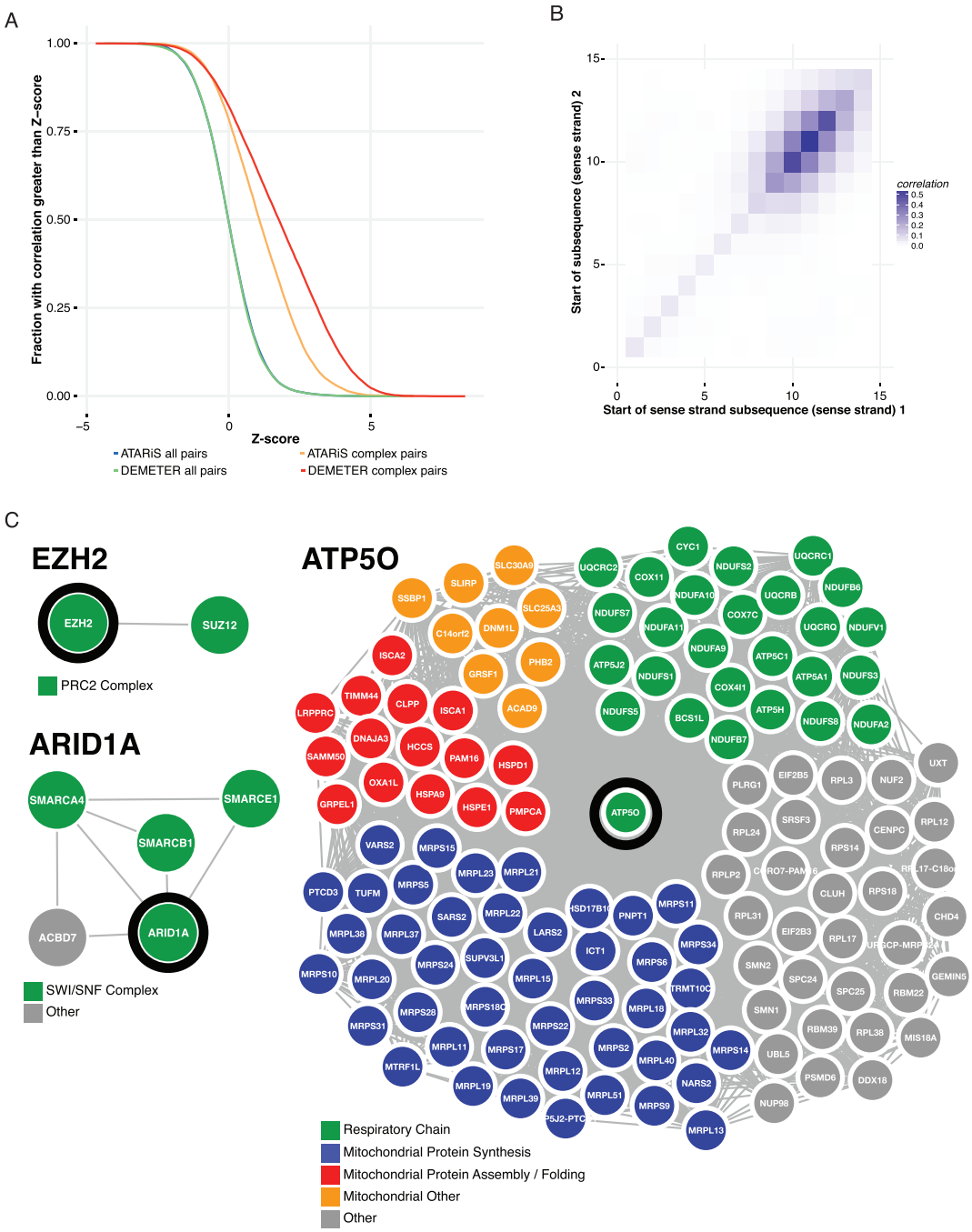


Figure S1. Correlation Analysis for Dependency Profiles, Related to Figure 1

(A) Correlation analysis for dependency profiles of genes associated with protein complexes. Cumulative distributions of z-scored Pearson correlation coefficients for pairs of dependency profiles generated by ATARiS and by DEMETER. Distributions generated from pairs of dependency profiles corresponding to proteins participating in the same complex (as annotated by the CORUM database) are in orange (ATARiS) and red (DEMETER). Those generated from random pairs of dependency profiles are in blue (ATARiS) and green (DEMETER).

(B) Co-dependency networks for EZH2, ARID1A and ATP5O. Edges represent significant Pearson correlation (z-score > 3) between a pair of gene dependency profiles across 501 screened lines. All genes with dependency profiles significantly correlated to the target gene's (circled in black) are shown, with colors representing functional classes.

(C) Correlation analysis for depletion profiles of shRNAs sharing 7-mer sequences. Correlation analysis for depletion profiles of shRNAs sharing 7-mer sequences. Pearson correlation coefficient was computed for the depletion profiles of pairs of shRNAs sharing the same 7-mer sequence, starting at positions as indicated by the x axis and the y axis. The color of each cell in the heatmap represents the average coefficient for all such pairs of shRNAs.

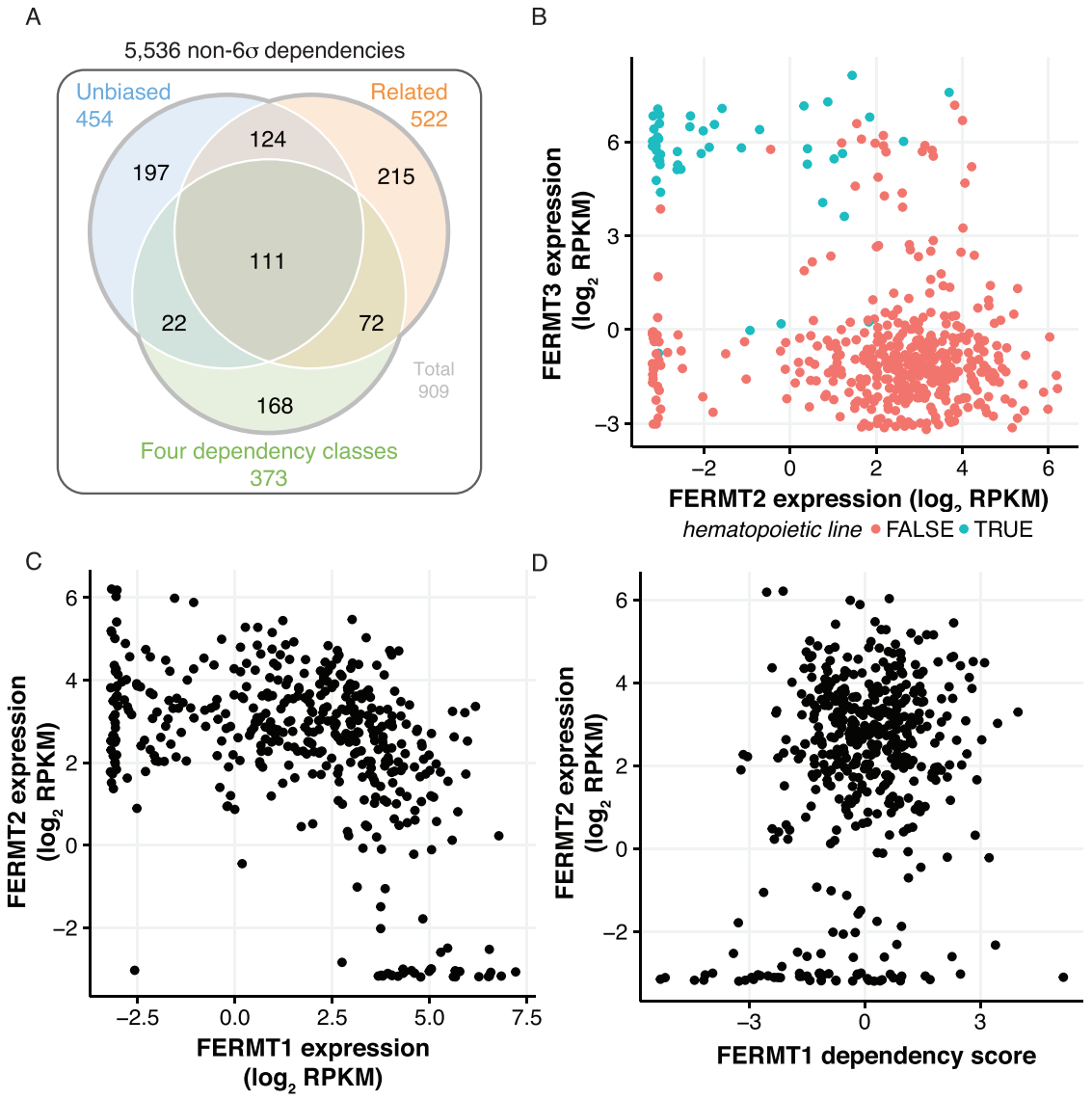


Figure S2. Predicting Dependencies from Molecular Features, Related to Figure 3

(A) Non-6σ dependencies with a predictive model. The number of non-6σ dependencies with predictive models built using all features (Unbiased, blue), features of genes related to the dependency gene (Related, red) and those falling into one of the four identified dependency classes (green).

(B) FERMT2 expression levels (x axis) are plotted against FERMT3 expression levels (y axis). Hematopoietic cell lines are colored in blue, all others are in red.

(C) FERMT1 expression levels (x axis) are plotted against FERMT2 expression levels (y axis).

(D) FERMT1 dependency (x axis) is compared to FERMT2 expression levels (y axis).

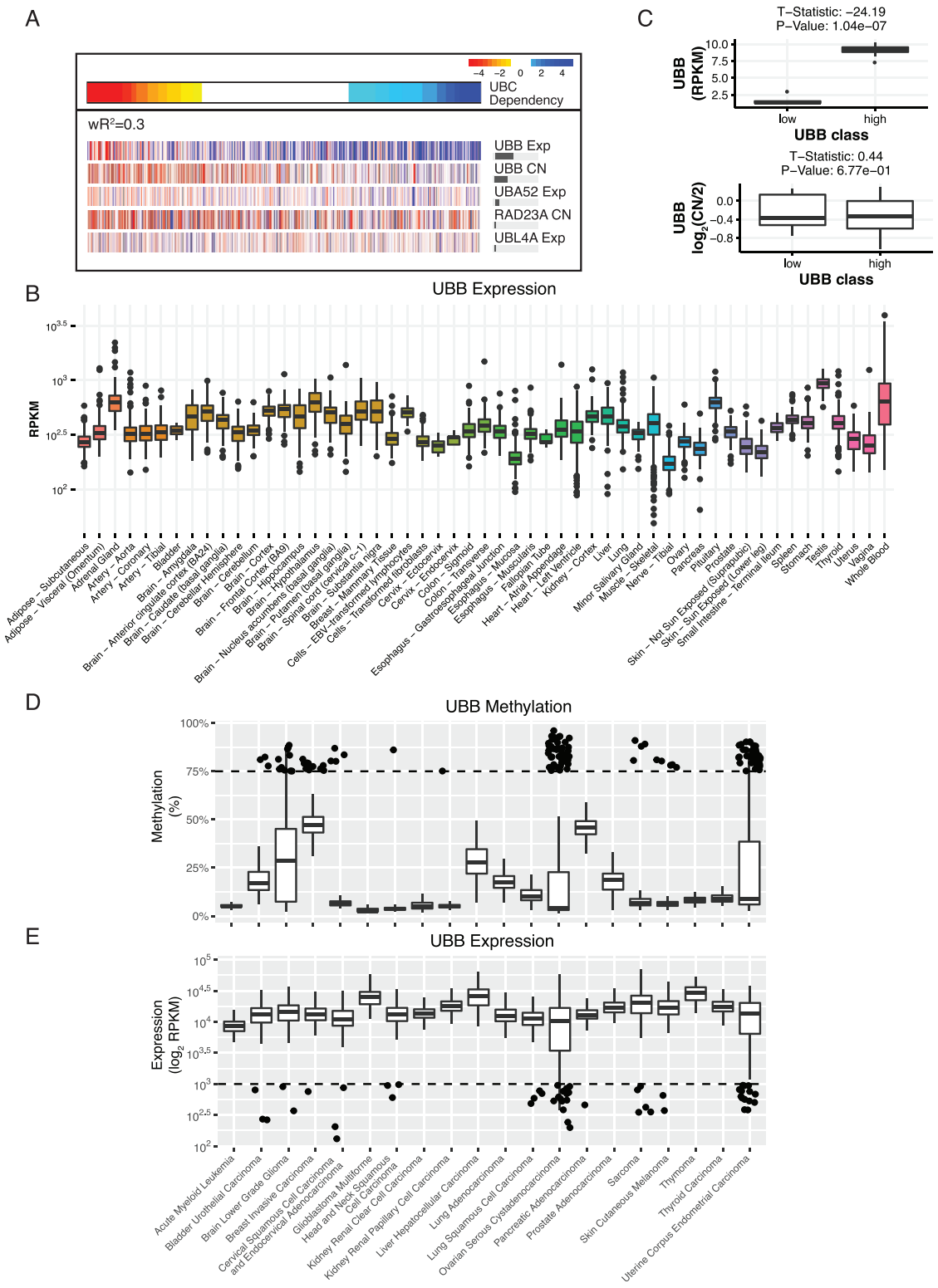


Figure S3. *UBC* Dependency Is Predicted by Low *UBB* Expression Levels, Related to Figure 5

(A) MDP paralog deficiency ATLANTIS model for *UBC*. *UBC* dependency is shown from most to least dependent cell line in columns (top panel, red to blue). Each lower panel shows the top five predictive markers used by that model; marker values are z-scores (high to low, red to blue). Horizontal bars on the right indicate the relative contribution to the model's out-of-bag R^2.

(legend continued on next page)

(B) *UBB* mRNA expression across tissues (data from GTEx).
(C) CCLE cell lines were classified as UBB^{high} or UBB^{low} based on whether *UBB* expression was greater or less than 10^5. The significance of the difference on *UBB* CN levels between these two classes was calculated by a two-tailed t test.
(D) *UBB* methylation and expression (y axis; RNaseq log_2RPKM) (E) across tumors (data from TCGA).

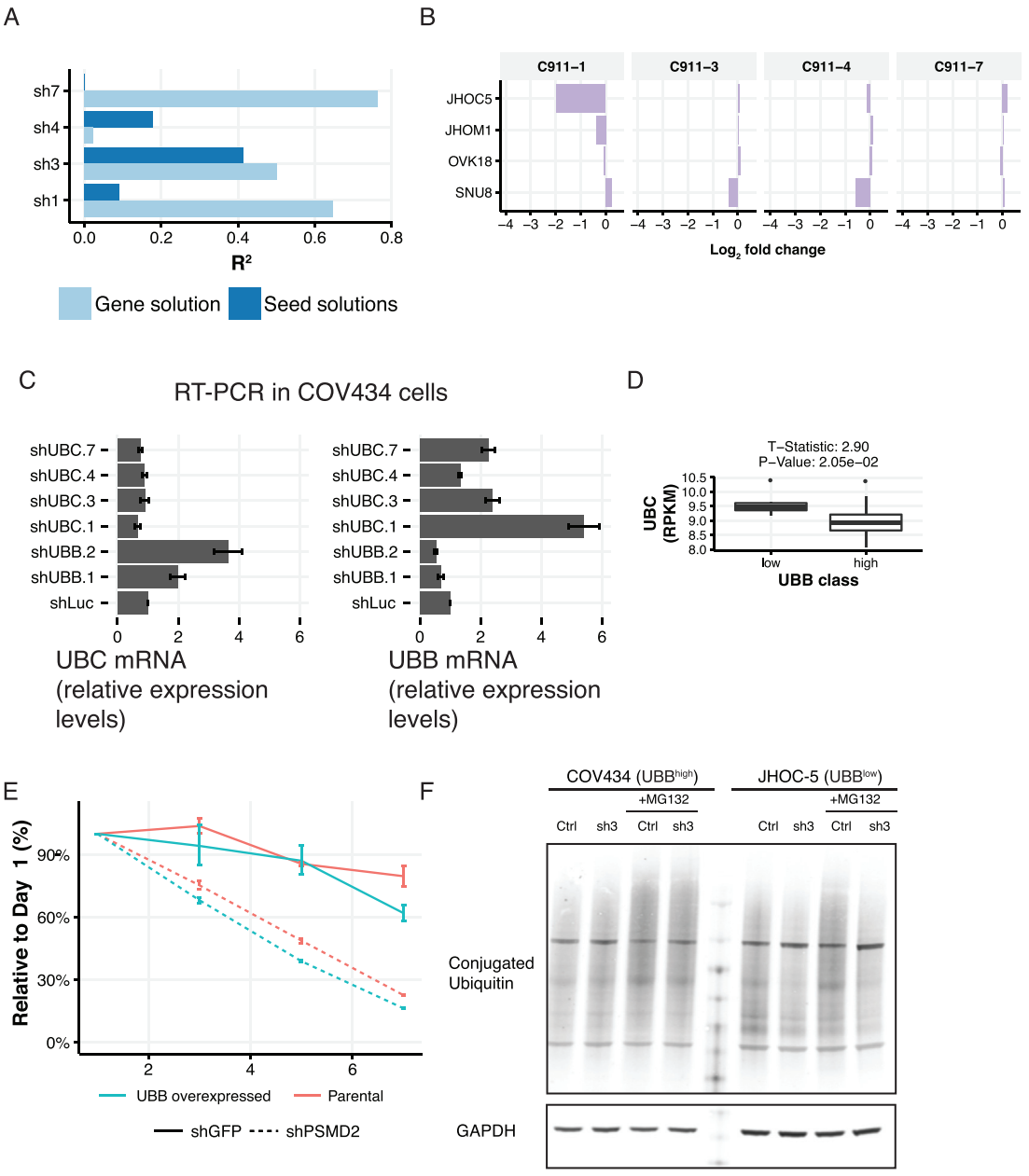


Figure S4. *UBC* and *UBB* Are Redundant Dependencies, Related to Figure 5

(A) DEMETER gene solutions and seed solutions for UBC shRNAs.

(B) GFP viability competition assay in ovarian cell lines with the indicated C911 shRNA controls for UBC shRNAs. Data represent log2 fold change compared to the average shRNAs negative controls.

(C) Expression levels of *UBC* or *UBB* in COV434 cells upon suppression of UBC expression with the indicated shRNAs or a control shRNA (shLuc). Data represent fold change relative to shLuc and error bars represent standard deviation of 3 technical replicates.

(D) CCLE cell lines classified as UBB^{high} or UBB^{low} based on whether *UBB* expression was greater or less than 10^5. The significance of the difference in *UBC* expression levels between these two classes was calculated by a two-tailed t test.

(E) Time course of relative viability upon PSMD2 suppression with or without ectopic expression of monoubiquitin (UBB) in a UBB^{low} cell line (SNU8). Data represent fold change relative to day 1 normalized to pLKO_TRC005 nullT. Error bars represent SD.

(F) Levels of conjugated ubiquitin upon *UBC* suppression in a UBB^{low} (JHOC5) and a UBB^{high} (COV434) cell line.

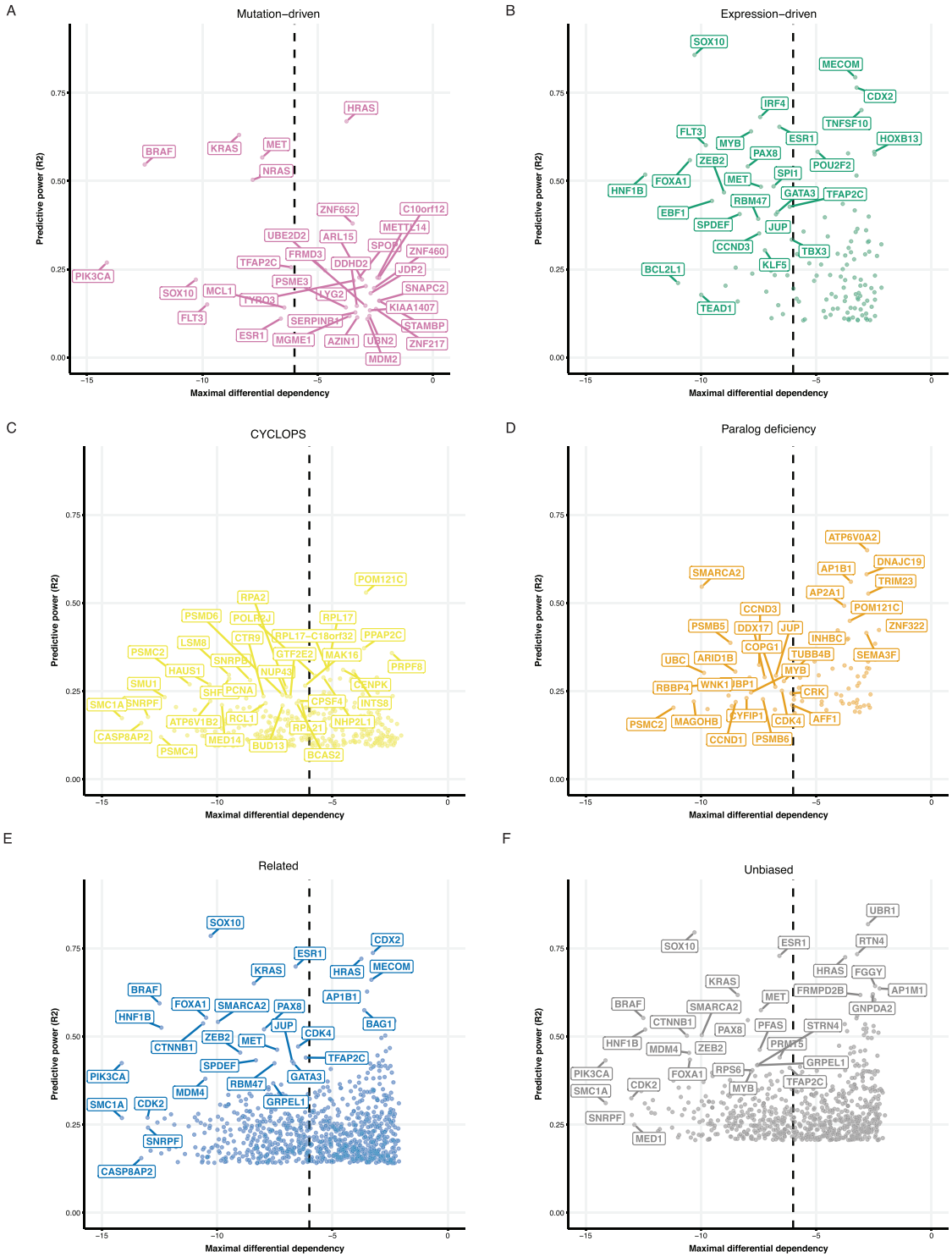


Figure S5. Summary of Differential Dependencies by MDP Class, Related to Figure 6

(A–F) For each differential dependency with a significant predictive model, the model's predictive power (y axis) and the strength of the dependency in the most dependent cell line (x axis) are indicated. MDP classes shown: (A) mutation-drive, (B) expression-driven, (C) CYCLOPS, (D) Paralog deficiency, (E) related features, and (F) unbiased (all) features.