shinyDepMap, a tool to identify targetable cancer genes and their functional connections from Cancer Dependency Map data

Kenichi Shimada*, John A Bachman, Jeremy L Muhlich, and Timothy J Mitchison

Laboratory of Systems Pharmacology and Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA. *Corresponding Author and Lead Contact.

Contact Info: kenichi_shimada@hms.harvard.edu

Abstract

Individual cancers rely on distinct essential genes for their survival. The Cancer Dependency Map (DepMap) is an ongoing project to uncover these gene dependencies in hundreds of cancer cell lines. To make this drug discovery resource more accessible to the scientific community we built an easy-to-use browser, shinyDepMap (https://labsyspharm.shinyapps.io/depmap). shinyDepMap combines CRISPR and shRNA data to determine, for each gene, the growth reduction caused by knockout/knockdown and the selectivity of this effect across cell lines. The tool also clusters genes with similar dependencies, revealing functional relationships. shinyDepMap can be used to 1) predict the efficacy and selectivity of drugs targeting particular genes; 2) identify maximally sensitive cell lines for testing a drug; 3) target hop, i.e., navigate from an undruggable protein with the desired selectivity profile, such as an activated oncogene, to more druggable targets with a similar profile; and 4) identify novel pathways driving cancer cell growth and survival.

Introduction

Cancer is a disease of the genome. Hundreds, if not thousands, of driver mutations cause cancer in different patients (Bailey et al., 2018) and extensive collaborative efforts such as the Cancer Genome Atlas Program (TCGA) have helped discover them (The Cancer Genome Atlas Research Network, 2019). Targeted therapies, a type of precision medicine, aim to treat cancer by selectively killing cancer cells with a specific genotype and spectrum of driver mutations (Friedman et al., 2015). The underlying hypothesis is that cancers depend on essential genes that are not the same for all tissues, and that these conditionally essential genes constitute a druggable dependency—an “Achilles’ heel”—that can be exploited to develop targeted drugs with minimal toxicity. To achieve this goal, it is important to identify conditionally essential genes for all cancers. It is also important to group these conditionally essential genes into functionally related sets to maximize the chance of finding a druggable target within each set, such as a kinase or other enzyme.

The concept of essential vs non-essential genes arose largely from genetic research in model organisms. Traditionally, it was considered a binary distinction that held across any genotype. However, loss of a given gene can decrease cell growth without killing the cell, so it is more realistic to assign a numerical value to the degree of essentiality, *i.e.*, the extent to which loss of a gene, or inhibition of its product, influences fitness. In cancer, this value may depend on the genotype, transcriptome, and lineage of the cell. In principle, genes that are only essential in a few cell types might make better drug targets since inhibiting their function is less likely to cause toxicity in non-cancer tissues. For example, the epidermal growth factor receptor is strongly required in certain cancer cells, but not in normal bone marrow stem cells, making it a potentially good target (Wang et al., 2006).

The Cancer Dependency Map (DepMap) is an ongoing project to identify essential genes across hundreds of cancer cell lines using genome-wide CRISPR and shRNA screens (Tsherniak et al., 2017; Behan et al., 2019). It has already been used successfully to discover cancer cells' genetic vulnerabilities (Sandoval et al., 2018; X. Wang et al., 2019). These data represent a gold mine of useful information for biologists and drug developers, but can be challenging for non-bioinformaticians to manipulate and interpret.

The DepMap portal website (https://depmap.org/portal) provides a range of information for each gene, including 1) the cell lines and lineages dependent on the gene, 2) co-dependent genes (*i.e.*, other genes whose effects on growth are strongly positively or negatively correlated with the gene), and 3) basal transcript abundances, copy numbers, and mutations for the gene. However, the DepMap portal has no native tools to integrate CRISPR and shRNA or to examine functional relationships among essential genes beyond pairwise comparisons.

Here we describe shinyDepMap, a web tool to enable researchers to rapidly determine the essentiality and selectivity of a given gene across cell lines and to find groups of functionally related genes with similar essentiality profiles. shinyDepMap integrates data from both CRISPR and shRNA screens, yielding robust measures of the effects of gene loss on cell viability. From these combined effect scores we derive two measures for each gene: the degree to which loss of the gene reduces cell growth in sensitive lines ("efficacy"), and the degree to which its essentiality varies across lines ("selectivity"). To help researchers identify potential therapeutic targets we clustered genes with strong efficacy scores into functional units, many of which represent complexes or biological pathways, as previously reported (Pan et al., 2018). The results of this analysis are accessible via a simple interactive web-tool at https://labsyspharm.shinyapps.io/depmap.

Results

Assessment of consistency between CRISPR and shRNA dependency scores

The DepMap project (https://depmap.org/) provides two separate pre-processed genome-wide genetic perturbation datasets for hundreds of cell lines using either shRNA or CRISPR (Meyers et al., 2017; McFarland et al., 2018). In both datasets, the preprocessed scores represent the growth effects of knocking the gene down or out, with a strongly negative value in a particular cell line indicating essentiality. Though the pre-processing algorithms for the shRNA data take “off-target” genes into account when generating essentiality scores, we nevertheless expected that the essentiality profiles would differ somewhat between shRNA and CRISPR due to their distinct mechanisms of reducing gene expression.

To assess the consistency between CRISPR and shRNA dependency scores, we first compared the gene/cell line combinations tested with both methods (15,847 genes in 423 cell lines, Figure 1 – source data 1) and computed essentiality thresholds for each distribution such that a dependency score more negative than the threshold is considered essential (Figure 1A; Methods). These thresholds define the subsets of cell line/gene combinations that are determined to be essential by either CRISPR or shRNA but not both (areas A and B in Figure 1A). The two methods were somewhat consistent on average, with both methods yielding approximately normal dependency score distributions with mean zero and a left-skewing tail corresponding to the subset of essential genes (Pearson correlation: 0.456, Spearman correlation: 0.201).

Despite this concordance, the comparison highlighted differences between the two methods at the individual gene level. First, CRISPR tends to detect weak to moderate gene

deletion effects more sensitively, as evidenced by the greater density of CRISPR-essential genes above the diagonal in the joint distribution plot (Figure 1A). For example, while both methods identify RAN, CRISPR identifies CCND1 as more essential (Figure 1B). Second, some genes were shown essential only by CRISPR or shRNA, but not by the other method (e.g., FOXD4 and EIF5B; Figure 1B).

To better understand these inconsistent dependencies, we used Fisher's exact test to determine which genes, across their perturbations in all 423 cell lines, were enriched for inconsistent dependencies. We found that 958 and 20 genes were claimed commonly essential only by CRISPR or shRNA, respectively (Figure 1C). Notably, these two sets of inconsistently essential genes are enriched for involvement in distinct pathways: for example, tRNA metabolic process and mitochondrial translation are overrepresented in the CRISPR-only set, whereas cytosolic translation initiation is overrepresented in the shRNA-only set (Figure 1D). This suggests that CRISPR and shRNA have distinct biases in assessing some genes' essentiality, affecting different classes of genes. While CRISPR is considered to be less susceptible to off-target effects (Smith et al., 2017) and thus now generally preferred over shRNA, the results and their therapeutic relevance may also depend on the genes of interest. For example, EIF5B, a gene involved in translation initiation, is highly conserved throughout Bacteria, Archaea, and Eukarya, suggesting that it may be essential for most human cells (Sørensen et al., 2001). However, only shRNA, but not CRISPR, highlighted it as an essential gene. A method that can combine the two dependency scores would compensate for each other's artifacts and give more robust scores.

A new dependency score combining CRISPR and shRNA

To summarize both CRISPR and shRNA dependency scores, S^C and S^R, we developed a new dependency score by combining them. Recent studies have shown that similar approaches are practical (Gilvary et al., 2019; W. Wang et al., 2019). Our new dependency score, S^θ, was computed as the weighted average of the two such that $S^\theta = \theta S^C + (1-\theta)S^R$, where θ is the mixing ratio of the two scores. It is appropriate to have any $\theta \in [0,1]$, but we selected six values of $\theta : \theta \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$ in this study, which are equivalent to mixing CRISPR and shRNA scores at 0:100 (=S^R), 20:80, 40:60, 60:40, 80:20, and 100:0 (=S^C). Using the equation above, we computed S^θ for 15,847 genes in 423 cell lines for each θ (Figure 2 – source data 2). The distribution of S^θ was located between $S^C (= S^1)$ and $S^R (= S^0)$ (Figure 2B).

Efficacy: gene essentiality in a sensitive cell line

We used the combined CRISPR-shRNA gene dependency scores to identify genes that are either commonly or selectively essential in the cell line panel. This distinction is important for identifying therapeutic targets because inhibition of commonly essential genes may be toxic to both cancer cells and normal cells, whereas genes that are selectively essential to particular cancers may allow for a greater therapeutic window. To capture the therapeutic potential of selectively essential genes we characterized gene dependency effects with two parameters: the *efficacy*, which defines the strength of the effect in a sensitive cell line, and the *selectivity*, which describes the variation of the effect across cell lines.

We defined the efficacy $\mathcal{E}^\theta_{G,X}$ as the *X*-th percentile of the distribution of combined dependency scores S^θ for gene G across all cell lines, denoted S^θ_G. For a given *X*, we defined a gene as essential when $\mathcal{E}^\theta_{G,X}$ is lower than the essentiality threshold T_θ, which is determined from the distribution of S^θ for all genes in all cell lines (Figure 2C, top panel). Smaller values of

X (lower percentiles) lead to more extreme efficacy values $\mathcal{E}_{G,X}^{\theta}$ and identify more essential genes in smaller subsets of cell lines (Figure 2C, bottom panel). We selected X=1 for most of our analysis. This is equivalent to claiming a gene essential when roughly 5 out of 423 cell lines show dependence on the gene (p=7.8e-5, binomial test). We discovered from 3,621 to 5,094 commonly and selectively essential genes from S^{θ} with different θ (Figure 2D). Reflecting the inconsistencies between CRISPR and shRNA, only 56% (2,037 genes) of the essential genes overlapped between S^C and S^R (Figure 2D). As for the 958 and 20 genes claimed commonly essential only by CRISPR or shRNA (Figure 1C), the essential genes discovered with the same method naturally included all of them; however, the essential genes discovered with the other method included only 40 to 60% of them, highlighting the inconsistencies between them (Figure 2E). On the other hand, the combined dependency scores S^{θ}, particularly when $\theta = 0.2, 0.4, \text{and } 0.6$, provide a more sensitive measure, discovering most of the essential genes claimed by either method. The first principal component line between S^C and S^R was parallel to the line with $\theta = 0.66$ in Figure 2A, which maximizes the variance of (S^C, S^R). Among the six lines, $\theta = 0.6$ (CRISPR: shRNA=60:40), is most similar to this principal component line. Therefore, we chose $\theta = 0.6$ or the corresponding S^{θ} primarily for the rest of the analysis and compare the performance of different θ later.

Selectivity: the difference in gene essentiality among cell lines

We next defined selectivity, a measure of the cell line dependence of the response to the loss of a gene. Selectivity implies that gene loss has a widely varying effect across the population of cell lines, such that the dispersion of the score distribution for a selectively essential gene would be greater than that for a commonly essential gene. We defined the

dispersion of gene *G*, $\mathcal{D}_{G,X}^{\theta}$, as the difference between the X-th and (100-X)-th percentiles of S_G^{θ}, or $\mathcal{D}_{G,X}^{\theta} = \mathcal{E}_{G,100-X}^{\theta} - \mathcal{E}_{G,X}^{\theta}$. We found that $\mathcal{E}_{G,X}^{\theta}$ and $\mathcal{E}_{G,100-X}^{\theta}$ were related linearly for the majority of genes, corresponding to non-essential and commonly essential genes, while some genes had large positive residuals, corresponding to selectively essential genes (e.g., green vs orange in Figure 3A-B). We therefore defined the selectivity $\mathcal{S}_{G,X}^{\theta}$ using the residuals of the (100-X)-th percentile values for $\mathcal{E}_{G,100-X}^{\theta}$ relative to the red regression line Figure 3A, which we denote $\mathcal{R}_{G,X}^{\theta}$:

$$\mathcal{S}_{G,X}^{\theta} = \mathcal{R}_{G,X}^{\theta} / \widehat{\mathcal{D}_{G,X}^{\theta}} = (\mathcal{E}_{G,100-X}^{\theta} - \widehat{\mathcal{E}_{G,100-X}^{\theta}}) / \widehat{\mathcal{D}_{G,X}^{\theta}} = (\mathcal{D}_{G,X}^{\theta} - \widehat{\mathcal{D}_{G,X}^{\theta}}) / \widehat{\mathcal{D}_{G,X}^{\theta}}$$

where $\widehat{\mathcal{D}_{G,X}^{\theta}}$ is the expected dispersion of dependency scores based on the robust linear regression of $\mathcal{E}_{G,100-X}^{\theta}$ given $\mathcal{E}_{G,X}^{\theta}$, or

$$\widehat{\mathcal{D}_{G,X}^{\theta}} = \widehat{\mathcal{E}_{G,100-X}^{\theta}} - \mathcal{E}_{G,X}^{\theta}.$$

The expected dispersion $\widehat{\mathcal{D}_{G,X}^{\theta}}$ increases for more strongly negative efficacy scores $\mathcal{E}_{G,X}^{\theta}$, indicating greater variances in dependency effects for commonly or selectively essential genes (e.g., greater variance for RAN vs. ZCWPW1 in Figure 3B). This could be a result of experimental noise (e.g., fewer sequencing reads for negatively selected genes) or greater biological variability in dependency effects for these genes. To better distinguish whether genes have dispersion greater than would be expected simply based on their efficacy scores, we normalize the residual values $\mathcal{R}_{G,X}^{\theta}$ by dividing by the expected dispersion $\widehat{\mathcal{D}_{G,X}^{\theta}}$ to obtain a measure of selectivity that accounts for the variation of $\widehat{\mathcal{D}_{G,X}^{\theta}}$ with the X-th percentile efficacy, $\mathcal{E}_{G,X}^{\theta}$. Both the efficacy and the selectivity of all the genes across different θ is available to download (Figure 3C – source data 3).

Characteristics of essential genes

By comparing the efficacy and the selectivity, we found that genes with strongly negative efficacy tend to be less selective, whereas selectively essential genes tend to have only moderate efficacy (Figure 3C). To characterize the genes that have either negative efficacy or positive selectivity, we performed gene set enrichment analysis of 6,551 pathways. This revealed the pathways overrepresented among essential, selective, and both selective and essential genes. For example, chromatin regulation is overrepresented among genes that are both selective and essential; nuclear metabolism and translation are overrepresented among essential (but not selective) genes; and regulation of kinase activity and tissue development are overrepresented among selective (but not essential) genes (Figure 3D, Figure 3D – source data 4).

Relationship of the selectivity and the lineage specificity

Cells in different lineages tend to depend on distinct essential genes compared to cells in the same lineage. We examined the extent to which lineage-specific dependence contributes to selectivity. For each gene, we assessed the relationship between the number of cell lines dependent on the gene and the gene's efficacy and selectivity, and confirmed that lower selectivity and more negative efficacy and are associated with a greater number of dependent cell lines (Figure 4A).

We computed the number of distinct lineages dependent on each gene using the Adaptive Daisy Model (ADaM), a permutation-based statistical model reported previously (Behan et al., 2019). As with dependent cell lines, a greater number of dependent lineages was associated with more negative efficacy and lower selectivity (Figure 4B). Overall, we

found that 1,050 genes are commonly essential across all the lineages, 670 genes are essential in at least one lineage, and 2,581 essential genes were not lineage-dependent (Figure 4B-C, Figure 4 – source data 5). Unsurprisingly, the number of dependent cell lines is strongly associated with the number of dependent lineages (Figure 4C).

Using an independent CRISPR screening dataset, Behan et al. also proposed candidate drug targets based on selectively essential genes they identified for each lineage. We found that most genes with high selectivity scores were also identified as targets for one or more lineages in their analysis (Figure 4D). On the other hand, targets proposed for multiple lineages in Behan et al. (orange and red points in Figure 4D) tended to show moderate efficacy scores, but not necessarily high selectivity scores.

Though we saw a strong relationship between lower selectivity and a greater number of dependent cell lines and lineages (Figs. 4A-B), we note that lineage specificity is not the only cause of high selectivity. Conventionally, high selectivity is interpreted as being commonly essential within a few lineages but non-essential in others. Some genes do manifest this type of selectivity (e.g., broad dependence on CTNNB1 in colorectal cancer, MYB in leukemia, IRF4 in multiple myeloma, KRAS in pancreatic cancer, and SOX10 in skin cancer, Figure 4E), but such common essentiality within a lineage is relatively unusual. In many more cases, a gene that shows high selectivity is selectively essential within each lineage (e.g., only partial dependence on CTNNB1 in liver, lung, and pancreatic cancers, or on IRF4 in skin cancer). CDK4 is particularly strong example of this pattern as it is not commonly essential in any lineages, but is selectively essential in many. While more negative efficacy $\mathcal{E}_{G,X}^{\theta}$ is strongly correlated with larger dispersion $\mathcal{D}_{G,X}^{\theta}$ (e.g., PSMA1 vs ISL1, Figure 4E; see also Figs. 3A-B), some commonly essential genes that show similar efficacy have substantially higher selectivity

than others (e.g., the selective PSMB5 vs. the non-selective PSMA1) (Figure 4E; see also Figs 4A and 4C). These genes that are essential in many lineages but nevertheless have high selectivity could be promising drug targets given appropriate biomarkers to characterize sensitive cell lines.

Clustering essential genes to find related targets

The dependency profile of a gene carries information about the functions of the gene that make it essential in certain cellular contexts. When a set of genes comprise a functional unit (e.g., a pathway or a complex) that regulates cell viability, these genes would be expected to have similar dependency profiles. Gene-wise cluster analysis of the dependency data should therefore reveal functional units that connect essential genes into pathways or protein complexes. It may also be easier to relate the essentiality of pathways to cancer genotypes than to interpret the essentiality of individual genes. Pathway analysis can help identify druggable vulnerabilities at the pathway level that might be missed by single-gene analysis.

We clustered essential genes based on the similarity of the combined CRISPR-shRNA dependency scores across 423 cell lines. Our approach is based on a related pair of popular algorithms, t-distributed stochastic neighbor embedding (t-SNE) and density based spatial clustering and noise (DBSCAN) (Maaten and Hinton, 2008; Maaten, 2014; Ester et al., 1996). t-SNE is a technique that reduces the dimensionality of multi-dimensional data while preserving the pairwise distances between data points at high dimensions as much as possible. It has been widely used for visualizing high dimensional data, such as single-cell RNA-seq data (Mass et al., 2016). DBSCAN is a clustering algorithm that detects regions where the data points are gathered at high density and clusters them; it is often used to cluster

data points based on their coordinates in the t-SNE plot. The combination of t-SNE and DBSCAN (expressed as 't-SNE + DBSCAN', hereafter) is a powerful clustering algorithm for high-dimensional data, such as single-cell transcriptomes (Haber et al., 2017).

One limitation of this approach is that the t-SNE algorithm is stochastic, producing different results and clusters with different initial seeds. However, when we compared clusters yielded by t-SNE + DBSCAN from multiple runs, we found that strongly positively correlated points are always clustered together while weakly positively correlated points are less consistently so. To obtain robust cluster assignments from t-SNE + DBSCAN we therefore used a workflow we call ensemble clustering with hierarchy over DBSCAN on t-SNE with Spearman distance matrix (ECHODOTS).

Briefly, ECHODOTS consists of four steps (Figure 5A, Figure 5 – figure supplement 1): it 1) computes the pairwise Spearman distance matrix among essential genes, 2) feeds the distance matrix as input to run t-SNE with different initial seeds 200 times, 3) clusters data points based on their coordinates in the t-SNE plot with DBSCAN, and 4) identifies the sets of genes assigned to the same cluster consistently across the 200 sets of clusters using a technique called ensemble clustering (Hornik, 2005). ECHODOTS produces more reliable clusters than a single run of t-SNE + DBSCAN by seeking data points that are consistently clustered together.

Cluster reveals known and new connections among essential genes

We ran ECHODOTS against the combined dependency score S^{θ} of the 4,301 essential genes (θ=0.6, X=1), and assigned them into 879 small, 608 medium, and 338 large clusters (Figure 5 – source data 6). Genes in the same cluster tended to be close to each other on t-

SNE maps from individual runs, and some clusters were enriched for genes known to be members of specific biological pathways or complexes (Figure 5B, Figure 5B – source data 7). The median efficacy and selectivity of the clusters varied widely, suggesting that some represent more promising sets of drug targets than others (Figure 5C-D).

In examining the clusters, we found that they often included genes that were not all mutually correlated with one another. While strongly positively correlated genes tend to be located in the same neighborhood on the t-SNE map and subsequently clustered together in ECHODOTS (Figure 5B), for a gene to be added to a cluster it only needs to be correlated with at least one other gene in the cluster. The structure of the correlations among the genes within a cluster can therefore highlight subtle functional relationships. For example, we plotted the correlations among six highly essential genes in cluster S152, with an edge between genes indicating a Spearman correlation greater than 0.1 (Figure 5E). In this cluster, KEAP1 and KCTD10 are both strongly correlated with the E3 ubiquitin ligase CUL3 (with correlations of 0.387 and 0.29, respectively), but have no correlation with each other (correlation -0.002). This is likely due to the fact that KEAP1 and KCTD10 interact with CUL3 in a mutually exclusive manner: both serve as adaptor proteins binding the same site on CUL3 but the resulting complexes degrade distinct target proteins (NFE2L2 and RHOB, respectively) (Cullinan et al., 2004; Kovačević et al., 2018).

Perhaps more intriguing are clusters that appear to show a connection between specific cellular processes and genes not otherwise known to be involved in that process. We offer two examples. One is Cluster L119 (Figure 5F), which is comprised of three small clusters. Clusters S369 and S745 contain the core MAP kinase (MAPK) pathway proteins, including KRAS, RAF1, BRAF, and MAPK1, while cluster S641 consists of CTNNB1 (β-catenin) and

TCF7L2, which form a bipartite transcription factor complex that is a key effector of the Wnt signaling pathway (Jin and Liu, 2008). These small clusters are within the same large cluster, suggesting that MAPK and Wnt signaling are functionally related in dependent cancer cell lines (Jeong et al., 2018). Cluster L119 also contains SHOC2, which was positively correlated to KRAS, RAF1, BRAF, and MAPK1. Multiple KRAS-mutant cancers were recently shown to be vulnerable to the loss of SHOC2 in the context of MEK inhibition, confirming the link between SHOC2 and MAP kinase pathway driven cancers (Sulahian et al., 2019).

A final example is the cluster L91 (Figure 5G), also consisting of multiple smaller clusters. Cluster S154 contains the selenoprotein GPX4. GPX4 encodes glutathione peroxidase 4, an antioxidant enzyme, that reduces cytotoxic lipid peroxides and protect cells from a non-apoptotic cell death, called ferroptosis. Intriguingly, we found the other genes in Cluster S154 were involved in selenoprotein synthesis (SEPHS2, SEPSECS, PSTK, EFFSEC) (Squires and Berry, 2008), suggesting that the primary role of these genes in dependent cell lines is to synthesize GPX4. S572 contains another selenoprotein TXNRD1 and its substrate, TXNDC17 (Espinosa and Arnér, 2018), both of which are also strongly correlated with four selenoprotein synthesis genes. Overall, L91 seems to represent a gene set related to the sensitivity to ferroptosis (Abdalkader et al., 2018; Ingold et al., 2018).

Comparison between different dependency scores

We have so far computed the efficacy, the selectivity, and the clusters of essential genes using S^{θ} with the fixed mixing ratio, θ =0.6, since this S^{θ} retains the largest variance of the original CRISPR and shRNA scores (Figure 2A). Here we compare the performance of different θ.

For efficacy, larger θ (i.e., with a larger contribution of S^C to S^θ) gave more genes with strongly negative efficacy (Figure 5 – figure supplement 2A). Consistently, the number of dependent cell lines and the number of dependent lineages per gene increased with larger θ (Figure 5 – figure supplement 2B-F). For selectivity, larger θ gave more genes with high selectivity (Figure 5 – figure supplement 2G). We showed earlier that essential (i.e., negative efficacy), selective (i.e., high selectivity), and both selective and essential genes overrepresent different pathways when θ = 0.6 (Figure 3D). Similar pathways were overrepresented by selective and both selective and essential genes when θ = 0.8 and 1. However, no pathways were associated with selective or both selective and essential genes when θ < 0.5, suggesting that a high selectivity was given to genes more randomly (Figure 5 – figure supplement 2H). Since it is more likely that selective genes represent certain pathways such as the ones shown in Figure 2D, S^θ with θ > 0.5 are more reasonable ones to choose.

Next, we compared the clusters of the essential genes using ECHODOTS. Since the number of discovered essential genes varies with different θ (Figure 2D), we expect the number of clusters to be different. Therefore, we did not fix the number of clusters across θ. Instead, we sought the upper bound of the neighborhood threshold ε (termed ε_0) in DBSCAN for each θ because as ε gets larger than a certain value, ε_0, most points on the t-SNE plot would start to merge to form a single large cluster (See Methods). We can detect incorrect merging by measuring the ratio between the 1st and 2nd largest cluster sizes (Figure 5 – figure supplement 3A-B). We found that ε_0 is particularly small for θ = 0.8 and 1 compared to the rest of θ, and more clusters were discovered for these θ consequently (Figure 5 – figure supplement 3C-D). We compared cluster memberships of the 2,008 genes identified as essential among all θ, and found that θ > 0.5 and θ < 0.5 gave substantially different clusters

(Figure 5 – figure supplement 3E). Through the comparison, we concluded that our initial choice of θ = 0.6 was a reasonable one since the combined dependency score is more informative with more weight on CRISPR than shRNA (θ > 0.5, Figure 5 – figure supplement 3) while having some contribution from shRNA is more beneficial than CRISPR alone (Figure 2E).

shinyDepMap: an interactive web tool to explore the essentiality of genes

Both the clusters of essential genes and the gene efficacy and selectivity scores provide valuable information for finding potential chemotherapeutic drug targets. To make this information accessible to the broader community of experimental drug discovery researchers, we developed a web-based tool to explore these analyses, called shinyDepMap. shinyDepMap is written in R (Chang et al., 2019) using the shiny package for building interactive visualization tools. It consists of two apps: “Gene essentiality” and “Gene cluster.” Each app is a dashboard-style website (Figure 6A). shinyDepMap can be used in three ways: 1) via the website https://labsyspharm.shinyapps.io/depmap, 2) by downloading the code and pre-processed data from the GitHub repository (https://github.com/kenichi-shimada/shinyDepMap) and running it on a local computer, and 3) running the app from a Docker container using the image at https://hub.docker.com/r/labsyspharm/shinydepmap. The analysis workflow in the application is explained below (Figure 6B).

Gene essentiality (all protein-encoding genes) This app allows a user to explore the essentiality of all the genes tested in the DepMap genetic perturbation experiments. Its output has two panels. A scatterplot in the middle displays efficacy and selectivity scores for all genes (**3**, bold numbers correspond to the panels in Figure 6B-C). By hovering over the plot points

with the cursor, one can find the genes corresponding to each point. When a gene name to search is provided in the input text box (**1**), corresponding genes will be highlighted in orange/red in the Efficacy-Selectivity plot (**3**). Genes matched with the query will be listed on the "Matched genes" tab in the right (**4**),in which by clicking a gene's name, the description of a gene in GeneCards (https://www.genecards.org/) will be open on a new page. By further selecting a matched gene from the dropdown menu (**2**), one can see the combined dependency scores of the gene in 423 cell lines in the "Dependency scores" tab on the right (**5**). The definition of combined dependency scores, the efficacy, and the selectivity can be changed by tuning the mix ratio (equivalent to θ; **6**) and the efficacy threshold (X-th percentile, **7**) from the input panel. We set them θ=0.6 and X=1 by default.

Gene cluster (essential genes) This app allows a user to explore gene clusters among the essential genes. When a user first selects an essential gene from the top-left dropdown menu (**8**), genes clustered with the query gene will be shown on the output panels. There are three output panels in the app. The top-center is the Efficacy-Selectivity plot for the essential genes (**9**). The bottom-center shows the t-SNE plot, indicating the similarity of the dependency scores among essential genes (**10**). The list of "Clustered genes" will be shown in the right panel (**11**). "Connectivity" tab will show how the clustered genes are connected (i.e., strongly correlated) (**12**). The graphs can be downloaded in the GraphML format. "Correlation" tab shows the Spearman correlation coefficients between the selected gene and the other essential genes grouped by the clusters (**13**). While all the genes in the Small cluster are shown by default, one can change it by tuning the "cluster size" parameter (**14**) and the probability threshold (**15**) in the left input panel. In ECHODOTS, we computed a probability at which each gene belongs to the assigned cluster. By setting the probability threshold close to

one, one can show only genes that are assigned to the same clusters consistently across many runs of t-SNE + DBSCAN. This app allows the users to tune the mix ratio (**16**) and the efficacy threshold (**17**) like the Gene essentiality app. Essential genes were defined based on these two parameters. Consequently, modifying these parameters affects the clusters.

Discussion

In this paper we described an interactive software tool, shinyDepMap, that allows users to rapidly determine the efficacy and selectivity of a gene of interest and thereby find highly selective genes that may offer promising therapeutic targets. shinyDepMap is based on both the CRISPR and shRNA genome-wide screening DepMap datasets, which we combined to generate a unified dependency score that is more informative than data from either dataset alone. Using this combined dependency score we computed “efficacy” and “selectivity” scores for each gene and highlighted how these scores can be used to characterize the therapeutic potential of targeting different genes across cell lines and lineages. Finally, we performed robust clustering of commonly and selectively essential genes to highlight functional relationships. shinyDepMap allows users to interactively explore both the essentiality and clustering results and is available as a deployed web-based application at https://labsyspharm.shinyapps.io/depmap and via source code or Docker image.

Our cluster analysis of the dependency scores highlighted genes comprising complexes and pathways, as previously reported (Pan et al., 2018). Our research complements published work in the area of complex and pathway annotation, in part because we were able to combine both DepMap datasets. We provide cluster information in a browsable form in shinyDepMap, including the ability to tune the size of clusters. One application of this tool is “target hopping,”

i.e., moving from one drug target to another while keeping the selectivity profile similar (Schenone et al., 2013). One goal of target hopping is to identify druggable targets with similar dependency score profiles to genes of interest that are not conventionally druggable. A classic example is KRAS, which is essential to many cancers but until recently has been considered "undruggable." KRAS appears in cluster L119 with the druggable kinases RAF1 and MAPK1, highlighting these proteins as relevant alternative targets. One goal of the shinyDepMap is to help researchers identify similar therapeutic opportunities among less well-studied genes.

Despite its potential value for therapeutic discovery, the DepMap dataset and our corresponding analysis in shinyDepMap has important practical limitations. The DepMap data characterizes the genetic requirements for cells grown in culture, which differs from the in vivo tumor environment in critical ways: the presence of nutrient-rich media, a two-dimensional rather than a three-dimensional substrate, and the lack of a functional immune system or physiological microenvironment. In addition, the effects of genetic perturbations (knockdown and knockout) do not necessarily correspond to those of chemical inhibition, a discrepancy that makes target identification from datasets like the DepMap less straightforward (Weiss et al., 2007). Finally, because the DepMap includes only cultured cancer lines and no wild-type cell lines or tissues, a highly selective gene in our analysis is not guaranteed to be less toxic to normal tissues when inhibited. Provided these limitations are kept in mind, the DepMap is a powerful resource and we hope the shinyDepMap tool makes it accessible to a broad community of researchers.

Methods

Data and Code Availability Statement

Following data of the 2019 Q3 release were downloaded from the DepMap project website: CRISPR (avana) (“Achilles_gene_effect.csv”), combined RNAi (“D2_combined_gene_dep_scores.csv”), and the cell line metadata (“sample_info.csv”). The CRISPR and shRNA efficacy data provided by the DepMap project were normalized with CERES and DEMETER2 algorithms by the Broad Institute, respectively. To compute the combined dependency score, we use the data of 15,847 genes in 423 cell lines, which were examined with both CRISPR and shRNA. The codes generated during this study are available at https://github.com/kenichi-shimada/depmap-analysis (data processing and analysis) and https://github.com/kenichi-shimada/shinyDepMap (standalone shinyDepMap). This study did not generate unique datasets.

Imputing missing values in shRNA and CRISPR dependency scores

We first took conditions that were tested in both CRISPR and shRNA were plotted. 15,847 genes were perturbed using both methods in 423 cell lines, and we compared 4,846,055 conditions that were non-missing values. Here, one condition is defined as a perturbation of one gene, either using CRISPR or shRNA in one cell line. We next imputed missing values in CRISPR and shRNA datasets using non-missing data from the other method. 1,345,642 conditions (20%) were tested using CRISPR, but not shRNA. 19,001 (0.28%) were tested using shRNA, but not CRISPR. In these cases, the missing values were imputed from the other data using local polynomial regression (loess) function in R. 4,457 (0.066%) conditions were not tested by either CRISPR or shRNA, which were left as missing. After missing values were imputed, the combined dependency score from two values was computed.

Computing the dependency score combining CRISPR and shRNA data

In this method section, we use $S^{\theta}_{G,L}$ instead of S^{θ} to represent the dependency score of gene *G* in cell line *L*, to make the argument clearer. To summarize CRISPR and shRNA dependency scores, we developed a new dependency score of a gene *G* in a cell line *L*, $S^{\theta}_{G,L}$, using the following equation:

$$S^{\theta}_{G,L} = \theta S^{C}_{G,L} + (1-\theta) S^{R}_{G,L} \quad (0 \le \theta \le 1)$$

where $S^{C}_{G,L}$ and $S^{R}_{G,L}$ are the dependency scores of the gene *g* in a cell line *l*, given by CRISPR and shRNA alone, respectively. The combined dependency score is a function of the mixing value θ. We particularly chose and compared six values of θ: $\theta \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. Resulting combined dependency scores, $S^{\theta}_{G,L}$, were computed for 15,847 genes in 423 cell lines for each θ.

Defining essential conditions

We defined the loss of a gene *G* is essential in a cell line *L* when the score $S^{\theta}_{G,L}$ is lower than the essentiality threshold T_{θ}: $S^{\theta}_{G,L} < T_{\theta}$. T_{θ} is defined as follows: we first fit the kernel density estimate function to the entire distribution of the dependency scores of across all genes and cell lines, or $S^{\theta}_{:,:}$. This distribution is well fitted with a normal distribution with a heavy left tail. We computed the mean μ and standard deviation σ from the right-half of the data and defined T_{θ} such that $P(S^{\theta}_{G,L} < T_{\theta}) = 0.001$ where $S^{\theta}_{G,L} \sim N(\mu, \sigma)$.

Identifying inconsistent essential genes between CRISPR and shRNA

To find out genes that are claimed essential only by CRISPR or shRNA, we first computed T_θ for CRISPR and shRNA, or T_C and T_R. Next, we sought the set of scores targeting the same gene in all the cells ($S^C_{G,\cdot}$, $S^R_{G,\cdot}$). The essentiality claimed only by CRISPR or shRNA were expressed as $S^C_{G,\cdot} < T_C \cap S^R_{G,\cdot} \geq T_R$ and $S^C_{G,\cdot} \geq T_C \cap S^R_{G,\cdot} < T_R$, respectively. They were illustrated as areas A and B in Fib. 1B. Using one-tailed Fisher's exact test, we computed the statistical significance of the enrichment of the data points ($S^C_{G,\cdot}$, $S^R_{G,\cdot}$) in the areas A and B and found 958 and 20 genes were claimed essential only by CRISPR and shRNA, respectively. We also computed the statistical significance of the overlap between these gene sets with publicly available gene annotations, Molecular Signature Database (MSigDB v7.0) (Subramanian et al., 2005), using Fisher's exact test.

Computing the efficacy and selectivity for each gene

The efficacy $\mathcal{E}^\theta_{G,X}$ measures how essential a gene is in a sensitive cell line:

$$\mathcal{E}^\theta_{G,X} = \text{ the X}^{\text{th}} \text{ percentile of } S^\theta_{G,\cdot} \qquad (X \in \{1, 2.5, 5, 10, 25\})$$

We defined a gene as essential for a given X when $\mathcal{E}^\theta_{G,X} < T_\theta$ (Figure 3A). Since T_θ is selected such that the probability $P(S^\theta_{G,L} < T_\theta) = 0.001$ when *G* is non-essential, we computed the probability of *N* cell lines of

To define the selectivity, we first determined the dispersion of the distribution of $S^C_{G,\cdot}$, $\mathcal{D}^\theta_{G,X}$:

$$\mathcal{D}^\theta_{G,X} = \mathcal{E}^\theta_{G,100-X} - \mathcal{E}^\theta_{G,X}$$

$\mathcal{E}^{\theta}_{G,X}$ and $\mathcal{E}^{\theta}_{G,100-X}$ were in a strong linear relationship for most of the genes, which correspond to commonly essential genes, while some genes have large positive residuals. We defined the residual from the regression line, $\mathcal{R}^{\theta}_{G,X}$, as follows:

$$\mathcal{R}^{\theta}_{G,X} = \mathcal{D}^{\theta}_{G,X} - \widehat{\mathcal{D}^{\theta}_{G,X}}$$

$$\widehat{\mathcal{D}^{\theta}_{G,X}} = \widehat{\mathcal{E}^{\theta}_{G,100-X}} - \mathcal{E}^{\theta}_{G,X} = f\left(\mathcal{E}^{\theta}_{G,X}\right) - \mathcal{E}^{\theta}_{G,X}$$

and the selectivity of the gene *G*, $\mathcal{S}^{\theta}_{G,X}$, was defined as:

$$\mathcal{S}^{\theta}_{G,X} = \mathcal{R}^{\theta}_{G,X} / \widehat{\mathcal{D}^{\theta}_{G,X}}.$$

Overlap of essential genes among different θ

The number of essential genes, defined as $\mathcal{E}^{\theta}_{G,1} < T_{\theta}$, depends on the mixing ratio θ. To assess the overlap between the essential gene sets across different θ, we computed an overlap index for any pairs of θ that is denoted as $O_X(\theta_1, \theta_2)$:

$$O_X(\theta_1, \theta_2) = N_X^{\theta_1 \cap \theta_2} / \min(N_X^{\theta_1}, N_X^{\theta_2})$$

where θ_1 and θ_2 are specific values of θ, $N_X^{\theta_1}$ is the number of essential genes when $\theta = \theta_1$, and $N_X^{\theta_1 \cap \theta_2}$ is the number of shared genes between the two essential gene sets. By definition, $O_X(\theta_1, \theta_2)$ can take any values between 0 and 1: the index is zero when the two essential gene sets do not share any genes; the index is one when the two essential genes are identical (Figure 2D).

Identifying pathways overrepresented by essential and/or selective genes

For each of essential (i.e., negative efficacy), selective (i.e., high selectivity), or both selective and essential genes, we sought pathways that were overrepresented by each gene.

We sorted all the genes by the efficacy and the selectivity in descending order and ran gene set enrichment analysis (GSEA) with the sorted genes against the pathways from MSigDB. GSEA was performed utilizing fgsea package with 10^7 permutations (Sergushichev, 2016).

Identifying lineage-specific and universally essential genes

To compute the lineage specificity, we utilized the Adaptive Daisy Model (ADaM). ADaM calculates the minimum number of dependent cell lines that are required for a gene to be considered as commonly essential among the cell lines in question (Behan et al., 2019). It is implemented in the ADAM2 R package (https://github.com/DepMap-Analytics/ADAM2). The dependency score matrix, $S^{\theta}_{:,:}$, contains the information of 423 cell lines representing 28 lineages. We focused on a subset of 387 cell lines in 17 lineages that includes ten or more cell lines. We computed the binary essentiality matrix for each G and L, where 1 if $S^{\theta}_{G,L} < T_{\theta}$ and 0 otherwise. We then provided a subset of the matrix attributed to each lineage as input and calculated the minimum number of dependent cell lines for the lineage. Each lineage is considered dependent on *G* when the number of dependent cell lines in the lineage is equal to or greater than the minimum number of the dependent cell lines. To compute universally essential genes utilizing ADaM, we calculated the binary essential matrix for the 17 lineages instead of cell lines and the minimum number of dependent lineages providing the matrix as input.

Behan et al. also provides a list of genes that are good targets for chemotherapies for each lineage. We counted the number of lineages they suggested was a good target for each gene and mapped them onto the Efficacy/Selectivity plot. (Figure 4D).

Robust cluster analysis utilizing t-SNE and DBSCAN: ECHODOTS

We implemented a new cluster algorithm, ensemble clustering with hierarchy over DBSCAN on t-SNE with Spearman distance matrix (ECHODOTS), extending the combination of t-SNE and DBSCAN. It is graphically summarized in Figure 5A and its pseudocode is provided in Figure 5 – figure supplement 1.

First, we computed the Spearman distance matrix or 1-Spearman correlation coefficient across all pairs of essential genes (line (1) in Figure 5 – figure supplement 1). This matrix was provided as input, and the coordinates of each data point in a 2D plane was computed with t-SNE (line (2)). Next, we clustered data points based on their coordinates with DBSCAN such that any two points whose distance is smaller than the neighborhood threshold ε are assigned into the same cluster (line (4)). We note that the range of the coordinates, *L*, varies among different runs of t-SNE. It is more reasonable to make the denominator *d* constant rather than to make ε constant, therefore we determined ε relative to *L*: $\varepsilon = L/d$ (line (3)). The resulting cluster set *C* was computed by DBSCAN using ε derived from the preset parameter *d* (line (4)), which we set an integer ranging from 30 to 200. We ran t-SNE and DBSCAN to compute *C* 200 times using different initial seeds.

Next, we computed consistent clusters *CC* across 200 cluster sets *C* for each *d* (or equivalently ε) using ensemble clustering available in clue R package (line (5)). When *d* is too small, (ε is too large), most of the points are erroneously connected. The mean size and the total number of consistent clusters depend on the neighborhood threshold ε. The smallest ε yields the largest number of small clusters, in which data points within each cluster are most tightly connected. The largest ε yields the smallest number of large clusters, in which data points within each cluster are most loosely connected. When ε is too large (correspondingly, *d*

was too small), most genes form one massive cluster in an extreme case. To avoid this, we set a lower bound for *d* or equivalently upper bound for ε such that $\varepsilon = L/d \leq L/d_0 = \varepsilon_0$ (line (6)). We determined d_0 by looking at the ratio between the sizes of the first and second largest clusters (Figure 5 – figure supplement 3A). When ε gets smaller (*d* gets larger), clusters get smaller and tighter. With smaller ε, some genes are not clustered with any other genes. We called these genes forming clusters only by themselves 'noise' and distinguished them from other clusters containing more than one gene. Eventually, all the genes become isolated or noise, but we stopped our analysis far ahead (*d* = 200). When θ = 0.6, we chose three different values of *d* (*d* = 65, 100, 141), which discovered 338, 608, and 879 clusters (line (7)). Most of the smaller clusters derived with larger *d* were contained in the larger clusters derived with smaller *d*. Therefore, we constructed a hierarchical relationship between the three sets of clusters. There were a few cases where a gene belongs to distinct clusters with different *d,* but the smaller cluster is not a part of the larger cluster and thus the hierarchy is not constructed.

Note that DBSCAN is a hard clustering algorithm that assigns each gene into only one cluster. Once clustered, both strongly and weakly correlated genes become indistinguishable in hard clusters. On the other hand, ECHODOTS performs a soft clustering that can assign each gene to more than one cluster. It conducts the 'majority vote' among 200 runs of the clustering and computes the probability of a point being assigned to each cluster, which provides us with the strength of evidence for the cluster assignment of each gene. Thus, ECHODOTS produces more reliable clusters than a single run of t-SNE + DBSCAN by seeking data points that are consistently clustered together.

Implementation of shinyDepMap website

shinyDepMap was built using shiny package. Following packages are also used to implement the tool: ggplot2 (v3.2.1), RColorBrewer (v1.1-2), shinyWidgets (v0.4.8), plotly (v4.9.0), DT (v0.8), visNetwork (v2.0.8), tibble (v2.1.3), dplyr (v0.8.3), tidyr (v0.8.3). shinyDepMap can be run locally without an internet connection. One can download the code and data at https://github.com/kenichi-shimada/shinyDepMap, and run locally following the link's instruction.

Acknowledgments

The authors thank Laura Maliszewski, Peter Sorger, and Rebecca Ward of Harvard Medical School for helpful discussions. This work is financially supported by the Japan Society for the Promotion of Science Overseas Research Fellowship (to K.S.), NIH R35GM131753 (to T.J.M.), and P50GM107618.

Competing Interests

J.A.B. has received consulting fees from Two Six Labs, LLC. The other authors do not have competing interests to declare.

Figure Legends

Figure 1. Systematic biases in CRISPR and shRNA dependency scores

A Comparison of normalized CRISPR and shRNA dependency scores of 15,847 protein-encoding genes in 423 cell lines. The density and contour plot corresponds to the distribution of the scores from all gene perturbations. Vertical and horizontal solid lines indicate the essentiality thresholds for CRISPR and shRNA dependency scores, respectively. Areas A, B, C correspond to the regions where only CRISPR, only shRNA, or both CRISPR and shRNA claimed essential. **B** Comparison of CRISPR and shRNA targeting four genes. In each panel, data points correspond to each gene's perturbation in 423 cell lines. Each point corresponds to one cell line. **C** 958 and 20 genes were claimed essential only by CRISPR or shRNA but not by the other method, respectively (Fisher's exact test, p-value < 1e-3). **D** Assessment of the pathways overrepresented by the essential genes claimed only by CRISPR or shRNA, highlighted in **C** (Fisher's exact test).

Figure 2. Identification of essential genes based on combined dependency score

A Dependency scores defined with different mixing ratios are computed by projecting each point onto the corresponding lines. θ denotes the fraction of CRISPR dependency scores. PC1 is the direction of the primary principal component line. **B** The distributions of combined dependency scores for four genes showed in Figure 1B. **C** Top panel: the distribution of the combined dependency score $S^{0.6}$. The essentiality threshold $T^{0.6}$ is determined based on this distribution. Bottom panel: the distribution of efficacy scores $\mathcal{E}_{G,X}^{0.6}$ with various X (-th percentile). Genes that satisfy $\mathcal{E}_{G,X}^{0.6} < T^{0.6}$ are defined as commonly or selectively essential, and the number of essential genes depends on X. **D** On the diagonal line, the numbers of commonly or

selectively essential genes identified with various θ are shown. In the off-diagonal area, the numbers of essential genes identified with two distinct θ are shown. Color code indicates an overlap index, a measure of overlap between two essential gene sets. The overlap index ranges from 0 (no shared genes) to 1 (the smaller set is included in the larger set). **E** The extent to which the genes claimed essential by only CRISPR or shRNA are covered by the essential genes discovered by each mixing ratio.

Figure 3. The efficacy and the selectivity

A The 1st and 99th percentiles of the combined dependency score of each gene where θ = 0.6. Each point corresponds to one gene. X- and Y-axis are equivalent to the efficacy with X=1 and X=99, respectively. Solid red line and Dashed black line are robust linear regression and identity lines, respectively. **B** Distribution of the combined dependency scores of four selective and four non-selective genes. The 1st and 99th percentile values within each distribution are also highlighted. **C** The efficacy and the selectivity of all genes are plotted. **D** Summary of the pathways overrepresented by genes with strongly negative efficacy and high selectivity, strongly negative efficacy, and high selectivity, respectively.

Figure 4. The lineage dependency

In the following panels **A**, **B**, and **D**, the efficacy and the selectivity scatterplots (Figure **3C**) are color-coded differently, highlighting the following properties of each gene. **A** The numbers of dependent cell lines. **B** The number of dependent lineages, computed with ADaM. **C** The relationship between the dependent cell lines and the dependent lineages. **D** The number of lineages in which Behan et al. suggested suitable for chemotherapy targets. **E** Nine genes'

649 dependency scores grouped by lineages, together with the number of dependent lineages and
650 cell lines, the efficacy, and the selectivity. All the panels in this figure were computed using θ =
651 0.6 and X=1.
652
653 **Figure 5. Essential gene clustering**
654 **A** The framework of ECHODOTS algorithm. **B** Nine gene clusters and their associated
655 pathways. **C** Median efficacy and selectivity of Large Clusters. **D** Genes consisting Large
656 Clusters with high selectivity highlighted in **C**. **E-G** The intra-cluster connectivity of three gene
657 clusters as exemplars. The colors of nodes indicate their membership of small clusters, and
658 the edges indicate that the two connected genes have Spearman correlation coefficient greater
659 than 0.1. Numbers in **E** indicate Spearman correlation coefficients.
660
661 **Figure 6. shinyDepMap: a web-tool to explore DepMap dataset**
662 **A** Top pag. **B** Gene essentiality app. **1** Textbox to type in a (partial) gene symbol to query. **2**
663 Dropdown menu to select a gene symbol that matches the query. **3** Efficacy-selectivity scatter
664 plot. **4** List of matched genes. **5** Combined dependency score profile of the gene selected in **2**.
665 **6** Mix ratio **7** Efficacy threshold
666 **C** Gene clusters app. **8** Dropdown menu to select an essential gene to explore. **9** Efficacy-
667 selectivity plot for essential genes. **10** t-SNE plot. **11** List of clustered genes. **12** Connectivity
668 plot. **13** Spearman correlation between the selected gene and the other essential genes. **14**
669 Cluster size input. **15** Probability threshold input. **16** Mix ratio. **17** Efficacy threshold.
670
671 **Figure 5 – figure supplement 1. ECHODOTS algorithm**

672 ECHODOTS algorithm is written using pseudocodes. The line numbers correspond to the line
673 numbers in the main text.
674
675 **Figure 5 – figure supplement 2. Efficacy, selectivity and the dependent lineages with**
676 **various θ**
677 **A** Empirical cumulative density functions (CDF) of the efficacy $\mathcal{E}_{G,X}^{\theta}$ across all the genes with
678 various θ. **B** Empirical CDFs of the number of dependent cell lines across all the genes with
679 various θ. **C** The distribution of the number of dependent lineages among essential genes with
680 various θ. The genes in the left- and right-hand side of the vertical lines are considered
681 selectively and commonly essential according to ADaM. **D-E** The efficacy-selectivity plot of all
682 the genes with various θ (X=1). The genes are color-coded based on the number of dependent
683 cell lines (**D**) and lineages (**E**). **F** Relationship between the number of dependent lineages and
684 the number of dependent cell lines with various θ (X=1). **G** Empirical CDFs of the selectivity
685 $\mathcal{S}_{G,X}^{\theta}$ across all the genes with various θ. **H** The number of overrepresented pathways
686 associated with genes with strongly negative efficacy and high selectivity, strongly negative
687 efficacy, and high selectivity.
688
689 **Figure 5 – figure supplement 3. Dependent cell lines and lineages using six dependency**
690 **scores**
691 In the panels **A-D**, four parameters were plotted on Y-axis against $d = L/\varepsilon$ on X-axis, where ε
692 is a neighborhood threshold in DBSCAN, for various for various θ. **A** The ratio between the
693 sizes of the first and second largest clusters (N_1/N_2). **B** The number of genes assigned into the
694 first and second largest clusters (N_1 and N_2) and the number of noise genes (N_n), *i.e.,* the

695 genes which are not clustered with other genes. **C** The number of clusters. **D** The mean cluster
696 size. **E** The similarity of the clusters with various θ. Cluster membership of the 2,008 genes
697 which were found essential with all θ was compared using cl_dissimilarity in clue R package.
698
699
700

Source Data

Source Data were made publicly available from the following link:

https://figshare.com/projects/shinyDepMap_Source_Data/97382

Figure 1 – source data 1. Information of the 423 cell lines in which both CRIPSR and shRNA screening were tested.

Figure 2 – source data 2. The combined dependency scores for 15,847 protein-coding genes in 423 cell lines for six θ.

θ = 0 and 1 corresponds to shRNA and CRISPR scores compared in Figure 1.

Figure 3C – source data 3. Efficacy and Selectivity for 15,847 genes for the six θ and five X: $X = \{1, 2.5, 5, 10, 25\}$

Figure 3D – source data 4. GO/KEGG Pathways overrepresented by genes with strongly negative efficacy, high selectivity, or strongly negative efficacy and high selectivity for six θ and X=1

Figure 4 – source data 5. Lineage-dependent essentiality of 17 lineages and common essentiality computed using ADaM for six θ

Figure 5 – Source Data 6. Cluster membership of essential genes and probability of their assignment to clusters for six θ

Figure 5B – Source Data 7. Pathways overrepresented in Large Clusters for six θ.

Clusters that contains 15 genes or more are only considered in this analysis.

References

Abdalkader M, Lampinen R, Kanninen KM, Malm TM, Liddell JR. 2018. Targeting Nrf2 to Suppress Ferroptosis and Mitochondrial Dysfunction in Neurodegeneration. *Front Neurosci* **12**. doi:10.3389/fnins.2018.00466

Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, Colaprico A, Wendl MC, Kim J, Reardon B, Ng PK-S, Jeong KJ, Cao S, Wang Zixing, Gao J, Gao Q, Wang F, Liu EM, Mularoni L, Rubio-Perez C, Nagarajan N, Cortés-Ciriano I, Zhou DC, Liang W-W, Hess JM, Yellapantula VD, Tamborero D, Gonzalez-Perez A, Suphavilai C, Ko JY, Khurana E, Park PJ, Van Allen EM, Liang H, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Wang Zhining, Yang L, Zenklusen JC, Zhang J (Julia), Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DI, Kim J, Lawrence MS, Lin P, Meier S, Noble MS, Saksena G, Voet D, Zhang Hailei, Bernard B, Chambwe N, Dhankani V, Knijnenburg T, Kramer R, Leinonen K, Liu Y, Miller M, Reynolds S, Shmulevich I, Thorsson V, Zhang W, Akbani R, Broom BM, Hegde AM, Ju Z, Kanchi RS, Korkut A, Li J, Liang H, Ling S, Liu W, Lu Y, Mills GB, Ng K-S, Rao A, Ryan M, Wang Jing, Weinstein JN, Zhang J, Abeshouse A, Armenia J, Chakravarty D, Chatila WK, de Bruijn I, Gao J, Gross BE, Heins ZJ, Kundra R, La K, Ladanyi M, Luna A, Nissan MG, Ochoa A, Phillips SM, Reznik E, Sanchez-Vega F, Sander C, Schultz N, Sheridan R, Sumer SO, Sun Y, Taylor BS, Wang Jioajiao, Zhang Hongxin, Anur P, Peto M, Spellman P, Benz C, Stuart JM, Wong CK, Yau C, Hayes DN, Parker JS, Wilkerson MD, Ally A, Balasundaram M, Bowlby R, Brooks D, Carlsen R, Chuah E, Dhalla N, Holt R, Jones SJM, Kasaian K, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Robertson AG, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Berger AC, Beroukhim R, Cherniack AD, Cibulskis C, Gabriel SB, Gao GF, Ha G, Meyerson M, Schumacher SE, Shih J, Kucherlapati MH, Kucherlapati RS, Baylin S, Cope L, Danilova L, Bootwalla MS, Lai PH, Maglinte DT, Van Den Berg DJ, Weisenberger DJ, Auman JT, Balu S, Bodenheimer T, Fan C, Hoadley KA, Hoyle AP, Jefferys SR, Jones CD, Meng S, Mieczkowski PA, Mose LE, Perou AH, Perou CM, Roach J, Shi Y, Simons JV, Skelly T, Soloway MG, Tan D, Veluvolu U, Fan H, Hinoue T, Laird PW, Shen H, Zhou W, Bellair M, Chang K, Covington K, Creighton CJ, Dinh H, Doddapaneni H, Donehower LA, Drummond J, Gibbs RA, Glenn R, Hale W, Han Y, Hu J, Korchina V, Lee S, Lewis L, Li W, Liu X, Morgan M, Morton D, Muzny D, Santibanez J, Sheth M, Shinbrot E, Wang L, Wang M, Wheeler DA, Xi L, Zhao F, Hess J, Appelbaum EL, Bailey M, Cordes MG, Ding L, Fronick CC, Fulton LA, Fulton RS, Kandoth C, Mardis ER, McLellan MD, Miller CA, Schmidt HK, Wilson RK, Crain D, Curley E, Gardner J, Lau K, Mallery D, Morris S, Paulauskis J, Penny R, Shelton C, Shelton T, Sherman M, Thompson E, Yena P, Bowen J, Gastier-Foster JM, Gerken M, Leraas KM, Lichtenberg TM, Ramirez NC, Wise L, Zmuda E, Corcoran N, Costello T, Hovens C, Carvalho AL, de Carvalho AC, Fregnani JH, Longatto-Filho A, Reis RM, Scapulatempo-Neto C, Silveira HCS, Vidal DO, Burnette A, Eschbacher J, Hermes B, Noss A, Singh R, Anderson ML, Castro PD, Ittmann M, Huntsman D, Kohl B, Le X, Thorp R, Andry C, Duffy ER, Lyadov V, Paklina O, Setdikova G, Shabunin A, Tavobilov M, McPherson C, Warnick R, Berkowitz R, Cramer D, Feltmate C, Horowitz N, Kibel A, Muto M, Raut CP, Malykh A, Barnholtz-Sloan JS, Barrett W, Devine K, Fulop J, Ostrom QT, Shimmel K, Wolinsky Y, Sloan AE, De Rose A, Giuliante F, Goodman M, Karlan BY, Hagedorn CH, Eckman J, Harr J, Myers J, Tucker K, Zach LA, Deyarmin B, Hu H, Kvecher L, Larson C, Mural RJ, Somiari S, Vicha A, Zelinka T, Bennett J, Iacocca M, Rabeno B,

Swanson P, Latour M, Lacombe L, Têtu B, Bergeron A, McGraw M, Staugaitis SM, Chabot J, Hibshoosh H, Sepulveda A, Su T, Wang T, Potapova O, Voronina O, Desjardins L, Mariani O, Roman-Roman S, Sastre X, Stern M-H, Cheng F, Signoretti S, Berchuck A, Bigner D, Lipp E, Marks J, McCall S, McLendon R, Secord A, Sharp A, Behera M, Brat DJ, Chen A, Delman K, Force S, Khuri F, Magliocca K, Maithel S, Olson JJ, Owonikoko T, Pickens A, Ramalingam S, Shin DM, Sica G, Van Meir EG, Zhang Hongzheng, Eijckenboom W, Gillis A, Korpershoek E, Looijenga L, Oosterhuis W, Stoop H, van Kessel KE, Zwarthoff EC, Calatozzolo C, Cuppini L, Cuzzubbo S, DiMeco F, Finocchiaro G, Mattei L, Perin A, Pollo B, Chen C, Houck J, Lohavanichbutr P, Hartmann A, Stoehr C, Stoehr R, Taubert H, Wach S, Wullich B, Kycler W, Murawa D, Wiznerowicz M, Chung K, Edenfield WJ, Martin J, Baudin E, Bubley G, Bueno R, De Rienzo A, Richards WG, Kalkanis S, Mikkelsen T, Noushmehr H, Scarpace L, Girard N, Aymerich M, Campo E, Giné E, Guillermo AL, Van Bang N, Hanh PT, Phu BD, Tang Y, Colman H, Evason K, Dottino PR, Martignetti JA, Gabra H, Juhl H, Akeredolu T, Stepa S, Hoon D, Ahn K, Kang KJ, Beuschlein F, Breggia A, Birrer M, Bell D, Borad M, Bryce AH, Castle E, Chandan V, Cheville J, Copland JA, Farnell M, Flotte T, Giama N, Ho T, Kendrick M, Kocher J-P, Kopp K, Moser C, Nagorney D, O'Brien D, O'Neill BP, Patel T, Petersen G, Que F, Rivera M, Roberts L, Smallridge R, Smyrk T, Stanton M, Thompson RH, Torbenson M, Yang JD, Zhang L, Brimo F, Ajani JA, Gonzalez AMA, Behrens C, Bondaruk J, Broaddus R, Czerniak B, Esmaeli B, Fujimoto J, Gershenwald J, Guo C, Lazar AJ, Logothetis C, Meric-Bernstam F, Moran C, Ramondetta L, Rice D, Sood A, Tamboli P, Thompson T, Troncoso P, Tsao A, Wistuba I, Carter C, Haydu L, Hersey P, Jakrot V, Kakavand H, Kefford R, Lee K, Long G, Mann G, Quinn M, Saw R, Scolyer R, Shannon K, Spillane A, Stretch J, Synott M, Thompson J, Wilmott J, Al-Ahmadie H, Chan TA, Ghossein R, Gopalan A, Levine DA, Reuter V, Singer S, Singh B, Tien NV, Broudy T, Mirsaidi C, Nair P, Drwiega P, Miller J, Smith J, Zaren H, Park J-W, Hung NP, Kebebew E, Linehan WM, Metwalli AR, Pacak K, Pinto PA, Schiffman M, Schmidt LS, Vocke CD, Wentzensen N, Worrell R, Yang H, Moncrieff M, Goparaju C, Melamed J, Pass H, Botnariuc N, Caraman I, Cernat M, Chemencedji I, Clipca A, Doruc S, Gorincioi G, Mura S, Pirtac M, Stancul I, Tcaciuc D, Albert M, Alexopoulou I, Arnaout A, Bartlett J, Engel J, Gilbert S, Parfitt J, Sekhon H, Thomas G, Rassl DM, Rintoul RC, Bifulco C, Tamakawa R, Urba W, Hayward N, Timmers H, Antenucci A, Facciolo F, Grazi G, Marino M, Merola R, de Krijger R, Gimenez-Roqueplo A-P, Piché A, Chevalier S, McKercher G, Birsoy K, Barnett G, Brewer C, Farver C, Naska T, Pennell NA, Raymond D, Schilero C, Smolenski K, Williams F, Morrison C, Borgia JA, Liptay MJ, Pool M, Seder CW, Junker K, Omberg L, Dinkin M, Manikhas G, Alvaro D, Bragazzi MC, Cardinale V, Carpino G, Gaudio E, Chesla D, Cottingham S, Dubina M, Moiseenko F, Dhanasekaran R, Becker K-F, Janssen K-P, Slotta-Huspenina J, Abdel-Rahman MH, Aziz D, Bell S, Cebulla CM, Davis A, Duell R, Elder JB, Hilty J, Kumar B, Lang J, Lehman NL, Mandt R, Nguyen P, Pilarski R, Rai K, Schoenfield L, Senecal K, Wakely P, Hansen P, Lechan R, Powers J, Tischler A, Grizzle WE, Sexton KC, Kastl A, Henderson J, Porten S, Waldmann J, Fassnacht M, Asa SL, Schadendorf D, Couce M, Graefen M, Huland H, Sauter G, Schlomm T, Simon R, Tennstedt P, Olabode O, Nelson M, Bathe O, Carroll PR, Chan JM, Disaia P, Glenn P, Kelley RK, Landen CN, Phillips J, Prados M, Simko J, Smith-McCune K, VandenBerg S, Roggin K, Fehrenbach A, Kendler A, Sifri S, Steele R, Jimeno A, Carey F, Forgie I, Mannelli M, Carney M, Hernandez B, Campos B, Herold-Mende C, Jungk C, Unterberg A, von Deimling A, Bossler A, Galbraith J, Jacobus L, Knudson M, Knutson T, Ma D, Milhem M, Sigmund R, Godwin AK, Madan R, Rosenthal HG, Adebamowo C, Adebamowo SN, Boussioutas A, Beer D, Giordano T, Mes-Masson A-M, Saad F, Bocklage T, Landrum L,

Mannel R, Moore K, Moxley K, Postier R, Walker J, Zuna R, Feldman M, Valdivieso F, Dhir R, Luketich J, Pinero EMM, Quintero-Aguilo M, Carlotti CG, Dos Santos JS, Kemp R, Sankarankuty A, Tirapelli D, Catto J, Agnew K, Swisher E, Creaney J, Robinson B, Shelley CS, Godwin EM, Kendall S, Shipman C, Bradford C, Carey T, Haddad A, Moyer J, Peterson L, Prince M, Rozek L, Wolf G, Bowman R, Fong KM, Yang I, Korst R, Rathmell WK, Fantacone-Campbell JL, Hooke JA, Kovatich AJ, Shriver CD, DiPersio J, Drake B, Govindan R, Heath S, Ley T, Van Tine B, Westervelt P, Rubin MA, Lee JI, Aredes ND, Mariamidze A, Lawrence MS, Godzik A, Lopez-Bigas N, Stuart J, Wheeler D, Getz G, Chen K, Lazar AJ, Mills GB, Karchin R, Ding L. 2018. Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**:371-385.e18. doi:10.1016/j.cell.2018.02.060

Behan FM, Iorio F, Picco G, Gonçalves E, Beaver CM, Migliardi G, Santos R, Rao Y, Sassi F, Pinnelli M, Ansari R, Harper S, Jackson DA, McRae R, Pooley R, Wilkinson P, Meer D van der, Dow D, Buser-Doepner C, Bertotti A, Trusolino L, Stronach EA, Saez-Rodriguez J, Yusa K, Garnett MJ. 2019. Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens. *Nature* **568**:511–516. doi:10.1038/s41586-019-1103-9

Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J. 2019. shiny: Web Application Framework for R.

Cullinan SB, Gordan JD, Jin J, Harper JW, Diehl JA. 2004. The Keap1-BTB Protein Is an Adaptor That Bridges Nrf2 to a Cul3-Based E3 Ligase: Oxidative Stress Sensing by a Cul3-Keap1 Ligase. *Molecular and Cellular Biology* **24**:8477–8486. doi:10.1128/MCB.24.19.8477-8486.2004

Espinosa B, Arnér ESJ. 2018. Thioredoxin-related protein of 14 kDa as a modulator of redox signalling pathways. *British Journal of Pharmacology* 544–553. doi:10.1111/bph.14479@10.1111/(ISSN)1476-5381.nitric_oxide_the_1998_prize_virtual_issue.html

Ester M, Kriegel H-P, Sander J, Xu X. 1996. A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with NoiseProceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96. Portland, Oregon: AAAI Press. pp. 226–231.

Friedman AA, Letai A, Fisher DE, Flaherty KT. 2015. Precision medicine for cancer with next-generation functional diagnostics. *Nature Reviews Cancer* **15**:747–756. doi:10.1038/nrc4015

Gilvary C, Madhukar NS, Gayvert K, Foronda M, Perez A, Leslie CS, Dow L, Pandey G, Elemento O. 2019. A machine learning approach predicts essential genes and pharmacological targets in cancer. *bioRxiv* 692277. doi:10.1101/692277

Haber AL, Biton M, Rogel N, Herbst RH, Shekhar K, Smillie C, Burgin G, Delorey TM, Howitt MR, Katz Y, Tirosh I, Beyaz S, Dionne D, Zhang M, Raychowdhury R, Garrett WS, Rozenblatt-Rosen O, Shi HN, Yilmaz O, Xavier RJ, Regev A. 2017. A single-cell survey of the small intestinal epithelium. *Nature* **551**:333–339. doi:10.1038/nature24489

Hornik K. 2005. A CLUE for CLUster ensembles. *Journal of Statistical Software* **14**:1–25.

Ingold I, Berndt C, Schmitt S, Doll S, Poschmann G, Buday K, Roveri A, Peng X, Porto Freitas F, Seibt T, Mehr L, Aichler M, Walch A, Lamp D, Jastroch M, Miyamoto S, Wurst W, Ursini F, Arnér ESJ, Fradejas-Villar N, Schweizer U, Zischka H, Friedmann Angeli JP, Conrad M. 2018. Selenium Utilization by GPX4 Is Required to Prevent Hydroperoxide-Induced Ferroptosis. *Cell* **172**:409-422.e21. doi:10.1016/j.cell.2017.11.048

Jeong W-J, Ro EJ, Choi K-Y. 2018. Interaction between Wnt/β-catenin and RAS-ERK pathways and an anti-cancer strategy via degradations of β-catenin and RAS by targeting the Wnt/β-catenin pathway. *npj Precision Oncology* **2**:1–10. doi:10.1038/s41698-018-0049-y

Jin T, Liu L. 2008. Minireview: The Wnt Signaling Pathway Effector TCF7L2 and Type 2 Diabetes Mellitus. *Molecular Endocrinology* **22**:2383–2392. doi:10.1210/me.2008-0135

Kovačević I, Sakaue T, Majoleé J, Pronk MC, Maekawa M, Geerts D, Fernandez-Borja M, Higashiyama S, Hordijk PL. 2018. The Cullin-3–Rbx1–KCTD10 complex controls endothelial barrier function via K63 ubiquitination of RhoB. *J Cell Biol* **217**:1015–1032. doi:10.1083/jcb.201606055

Maaten L van der. 2014. Accelerating t-SNE using Tree-Based Algorithms. *Journal of Machine Learning Research* **15**:3221–3245.

Maaten L van der, Hinton G. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**:2579–2605.

Mass E, Ballesteros I, Farlik M, Halbritter F, Günther P, Crozet L, Jacome-Galarza CE, Händler K, Klughammer J, Kobayashi Y, Gomez-Perdiguero E, Schultze JL, Beyer M, Bock C, Geissmann F. 2016. Specification of tissue-resident macrophages during organogenesis. *Science* **353**:aaf4238. doi:10.1126/science.aaf4238

McFarland JM, Ho ZV, Kugener G, Dempster JM, Montgomery PG, Bryan JG, Krill-Burger JM, Green TM, Vazquez F, Boehm JS, Golub TR, Hahn WC, Root DE, Tsherniak A. 2018. Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration. *Nat Commun* **9**. doi:10.1038/s41467-018-06916-5

Meyers RM, Bryan JG, McFarland JM, Weir BA, Sizemore AE, Xu H, Dharia NV, Montgomery PG, Cowley GS, Pantel S, Goodale A, Lee Y, Ali LD, Jiang G, Lubonja R, Harrington WF, Strickland M, Wu T, Hawes DC, Zhivich VA, Wyatt MR, Kalani Z, Chang JJ, Okamoto M, Stegmaier K, Golub TR, Boehm JS, Vazquez F, Root DE, Hahn WC, Tsherniak A. 2017. Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. *Nature Genetics* **49**:1779–1784. doi:10.1038/ng.3984

Pan J, Meyers RM, Michel BC, Mashtalir N, Sizemore AE, Wells JN, Cassel SH, Vazquez F, Weir BA, Hahn WC, Marsh JA, Tsherniak A, Kadoch C. 2018. Interrogation of Mammalian Protein Complex Structure, Function, and Membership Using Genome-Scale Fitness Screens. *Cell Syst* **6**:555-568.e7. doi:10.1016/j.cels.2018.04.011

Sandoval GJ, Pulice JL, Pakula H, Schenone M, Takeda DY, Pop M, Boulay G, Williamson KE, McBride MJ, Pan J, St. Pierre R, Hartman E, Garraway LA, Carr SA, Rivera MN, Li Z, Ronco L, Hahn WC, Kadoch C. 2018. Binding of TMPRSS2-ERG to BAF Chromatin Remodeling Complexes Mediates Prostate Oncogenesis. *Molecular Cell* **71**:554-566.e7. doi:10.1016/j.molcel.2018.06.040

Schenone M, Dančík V, Wagner BK, Clemons PA. 2013. Target identification and mechanism of action in chemical biology and drug discovery. *Nat Chem Biol* **9**:232–240. doi:10.1038/nchembio.1199

Sergushichev A. 2016. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *bioRxiv* 060012. doi:10.1101/060012

Smith I, Greenside PG, Natoli T, Lahr DL, Wadden D, Tirosh I, Narayan R, Root DE, Golub TR, Subramanian A, Doench JG. 2017. Evaluation of RNAi and CRISPR technologies by large-scale gene expression profiling in the Connectivity Map. *PLOS Biology* **15**:e2003213. doi:10.1371/journal.pbio.2003213

Sørensen HP, Hedegaard J, Sperling-Petersen HU, Mortensen KK. 2001. Remarkable Conservation of Translation Initiation Factors: IF1/eIF1A and IF2/eIF5B are Universally Distributed Phylogenetic Markers. *IUBMB Life* **51**:321–327. doi:10.1080/152165401317190842

Squires JE, Berry MJ. 2008. Eukaryotic selenoprotein synthesis: Mechanistic insight incorporating new factors and new functions for old factors. *IUBMB Life* **60**:232–235. doi:10.1002/iub.38

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**:15545–15550. doi:10.1073/pnas.0506580102

Sulahian R, Kwon JJ, Walsh KH, Pailler E, Bosse TL, Thaker M, Almanza D, Dempster JM, Pan J, Piccioni F, Dumont N, Gonzalez A, Rennhack J, Nabet B, Bachman JA, Goodale A, Lee Y, Bagul M, Liao R, Navarro A, Yuan TL, Ng RWS, Raghavan S, Gray NS, Tsherniak A, Vazquez F, Root DE, Firestone AJ, Settleman J, Hahn WC, Aguirre AJ. 2019. Synthetic Lethal Interaction of SHOC2 Depletion with MEK Inhibition in RAS-Driven Cancers. *Cell Reports* **29**:118-134.e8. doi:10.1016/j.celrep.2019.08.090

The Cancer Genome Atlas Research Network. 2019. The Cancer Genome Atlas Program. *National Cancer Institute*. https://www.cancer.gov/tcga

Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, Gill S, Harrington WF, Pantel S, Krill-Burger JM, Meyers RM, Ali L, Goodale A, Lee Y, Jiang G, Hsiao J, Gerath WFJ, Howell S, Merkel E, Ghandi M, Garraway LA, Root DE, Golub TR, Boehm JS, Hahn WC. 2017. Defining a Cancer Dependency Map. *Cell* **170**:564-576.e16. doi:10.1016/j.cell.2017.06.010

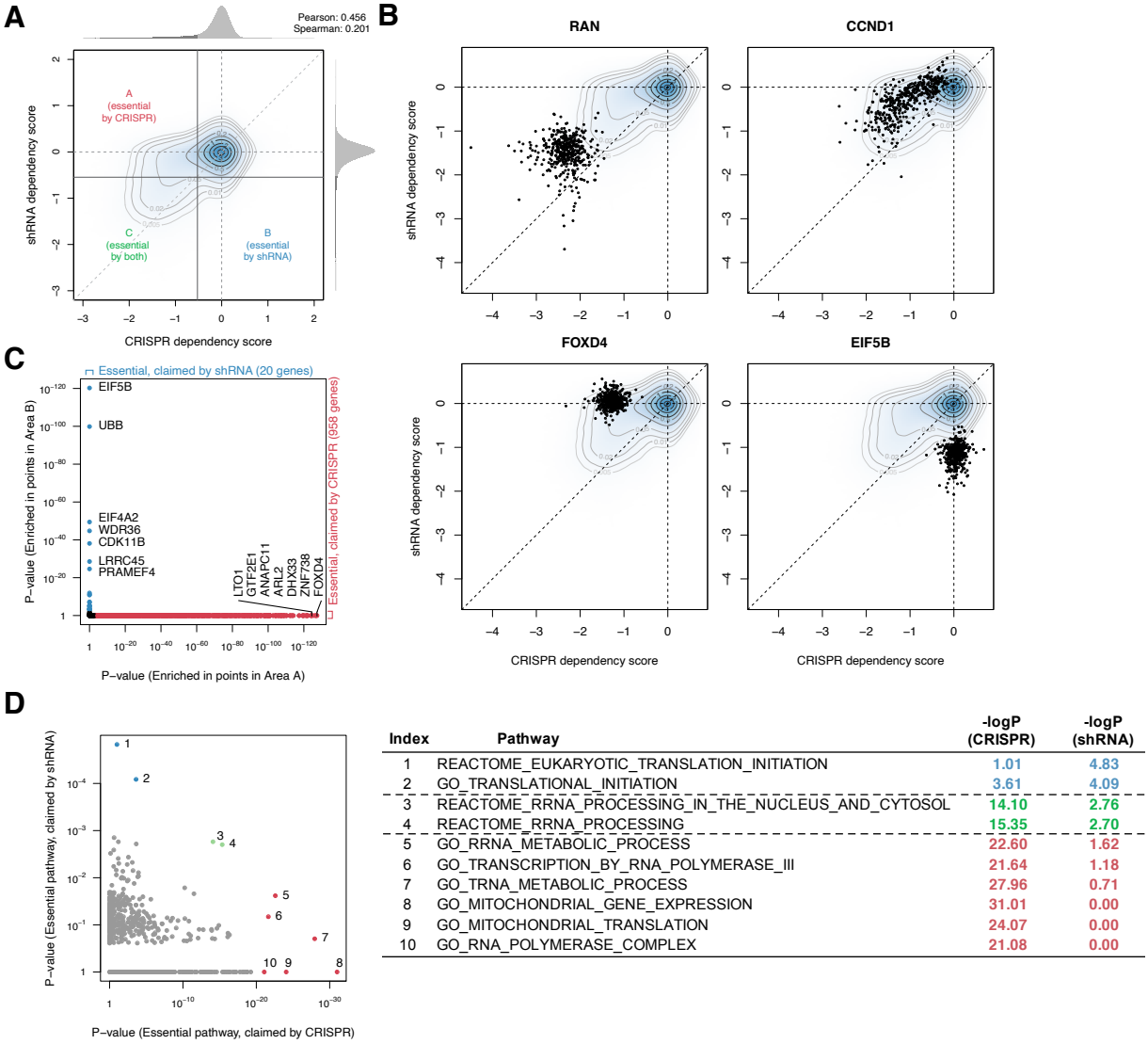
Wang W, Malyutina A, Pessia A, Saarela J, Heckman CA, Tang J. 2019. Combined gene essentiality scoring improves the prediction of cancer dependency maps. *EBioMedicine*. doi:10.1016/j.ebiom.2019.10.051

Wang X, Wang S, Troisi EC, Howard TP, Haswell JR, Wolf BK, Hawk WH, Ramos P, Oberlick EM, Tzvetkov EP, Ross A, Vazquez F, Hahn WC, Park PJ, Roberts CWM. 2019. BRD9 defines a SWI/SNF sub-complex and constitutes a specific vulnerability in malignant rhabdoid tumors. *Nat Commun* **10**:1–11. doi:10.1038/s41467-019-09891-7

Wang Y, Probin V, Zhou D. 2006. Cancer therapy-induced residual bone marrow injury-Mechanisms of induction and implication for therapy. *Curr Cancer Ther Rev* **2**:271–279.

Weiss WA, Taylor SS, Shokat KM. 2007. Recognizing and exploiting differences between RNAi and small-molecule inhibitors. *Nat Chem Biol* **3**:739–744. doi:10.1038/nchembio1207-739

Figure 1



Index	Pathway	-logP (CRISPR)	-logP (shRNA)
1	REACTOME_EUKARYOTIC_TRANSLATION_INITIATION	1.01	4.83
2	GO_TRANSLATIONAL_INITIATION	3.61	4.09
3	REACTOME_RRNA_PROCESSING_IN_THE_NUCLEUS_AND_CYTOSOL	14.10	2.76
4	REACTOME_RRNA_PROCESSING	15.35	2.70
5	GO_RRNA_METABOLIC_PROCESS	22.60	1.62
6	GO_TRANSCRIPTION_BY_RNA_POLYMERASE_III	21.64	1.18
7	GO_TRNA_METABOLIC_PROCESS	27.96	0.71
8	GO_MITOCHONDRIAL_GENE_EXPRESSION	31.01	0.00
9	GO_MITOCHONDRIAL_TRANSLATION	24.07	0.00
10	GO_RNA_POLYMERASE_COMPLEX	21.08	0.00

Figure 2

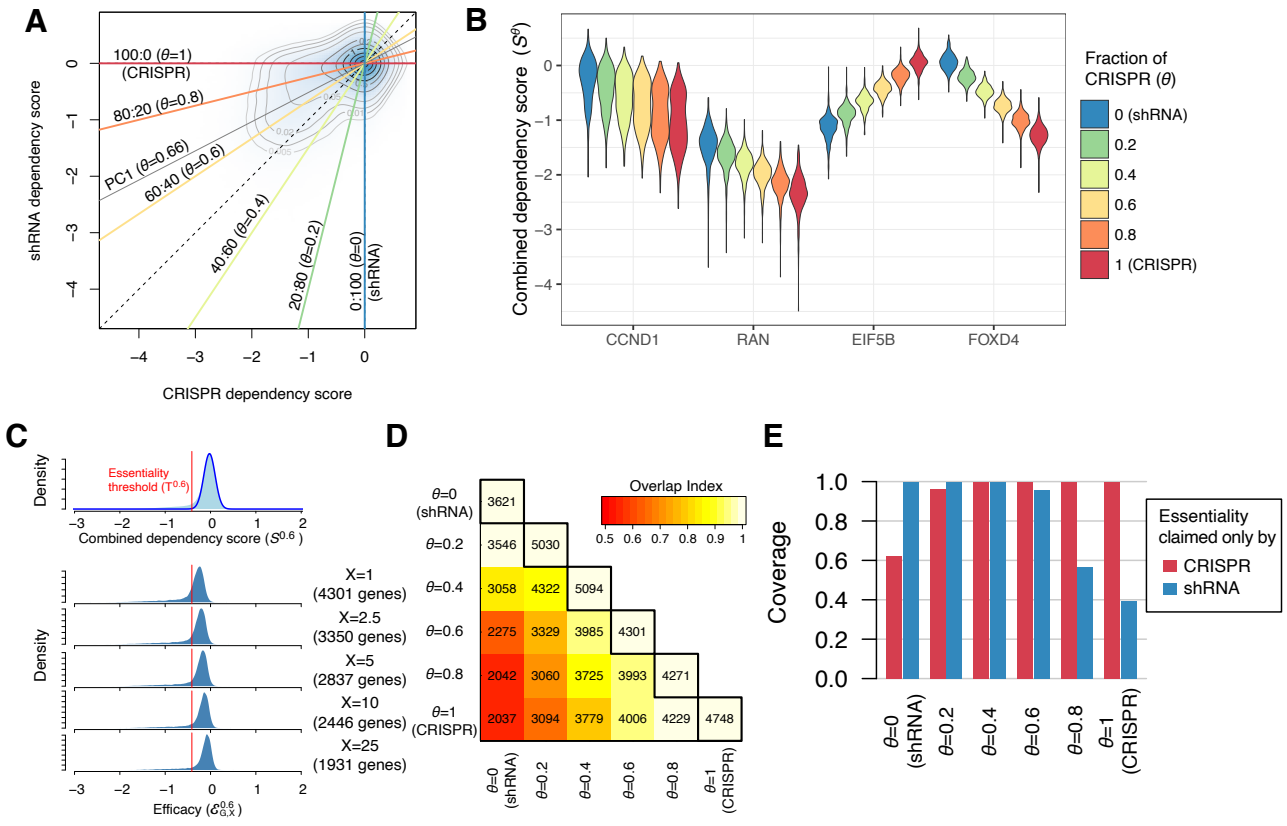
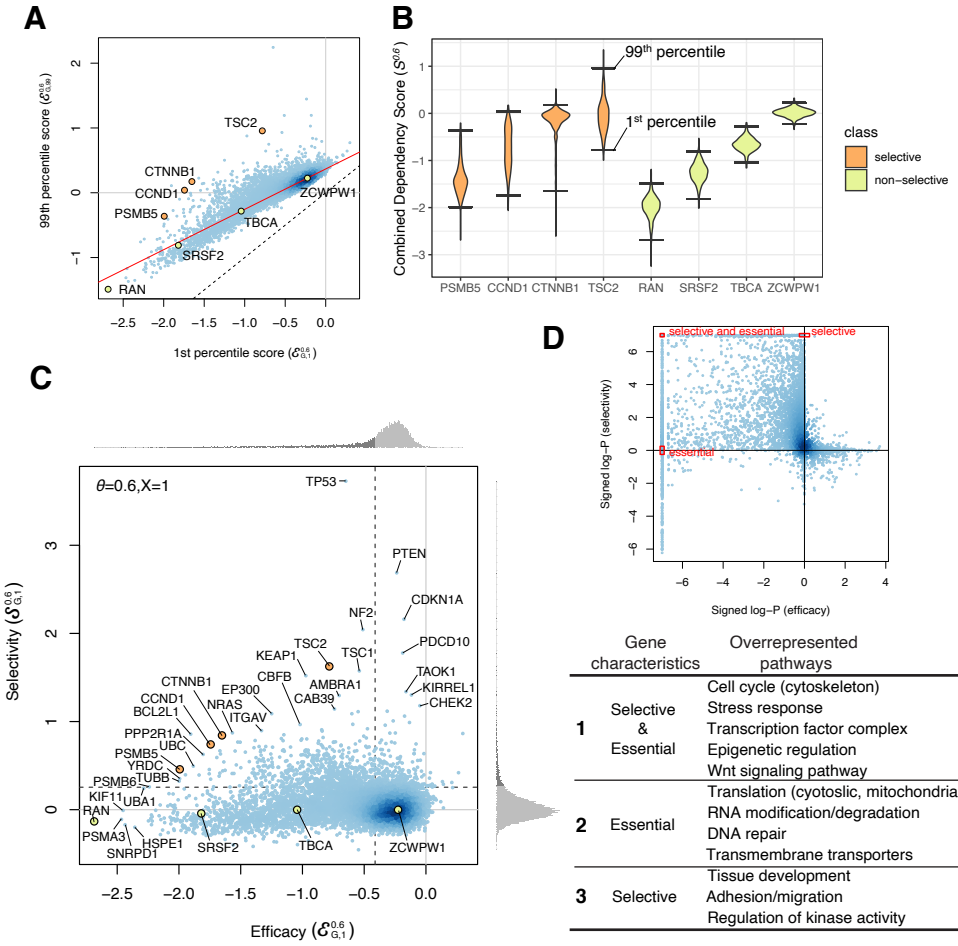
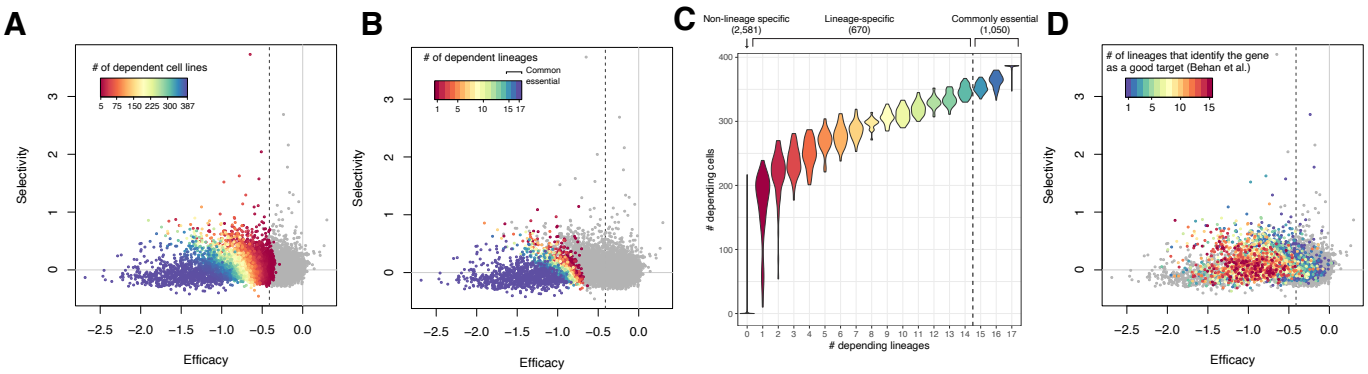

A
100:0 (θ=1)
(CRISPR)
80:20 (θ=0.8)
PC1 (θ=0.66)
60:40 (θ=0.6)
40:60 (θ=0.4)
20:80 (θ=0.2)
0:100 (θ=0)
(shRNA)
shRNA dependency score
CRISPR dependency score
B
Combined dependency score (S^θ)
CCND1
RAN
EIF5B
FOXD4
Fraction of
CRISPR (θ)
0 (shRNA)
0.2
0.4
0.6
0.8
1 (CRISPR)
C
Density
Essentiality
threshold (T^0.6)
Combined dependency score (S^0.6)
X=1
(4301 genes)
X=2.5
(3350 genes)
X=5
(2837 genes)
X=10
(2446 genes)
X=25
(1931 genes)
Efficacy (E^0.6_G,X)
D
Overlap Index
θ=0
(shRNA)
θ=0.2
θ=0.4
θ=0.6
θ=0.8
θ=1
(CRISPR)
3621
3546 5030
3058 4322 5094
2275 3329 3985 4301
2042 3060 3725 3993 4271
2037 3094 3779 4006 4229 4748
E
Coverage
Essentiality
claimed only by
CRISPR
shRNA

Figure 3

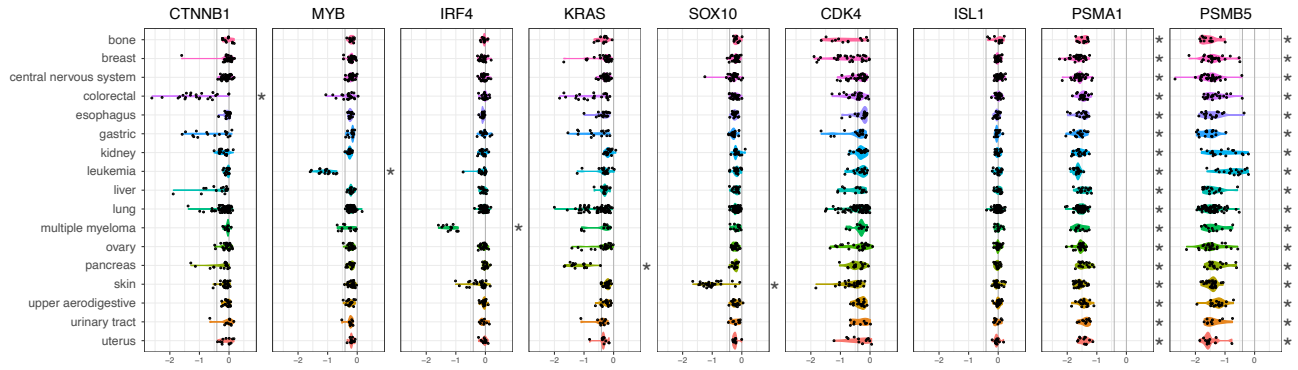


	Gene characteristics	Overrepresented pathways
1	Selective & Essential	Cell cycle (cytoskeleton) Stress response Transcription factor complex Epigenetic regulation Wnt signaling pathway
2	Essential	Translation (cytoslic, mitochondrial) RNA modification/degradation DNA repair Transmembrane transporters
3	Selective	Tissue development Adhesion/migration Regulation of kinase activity

Figure 4

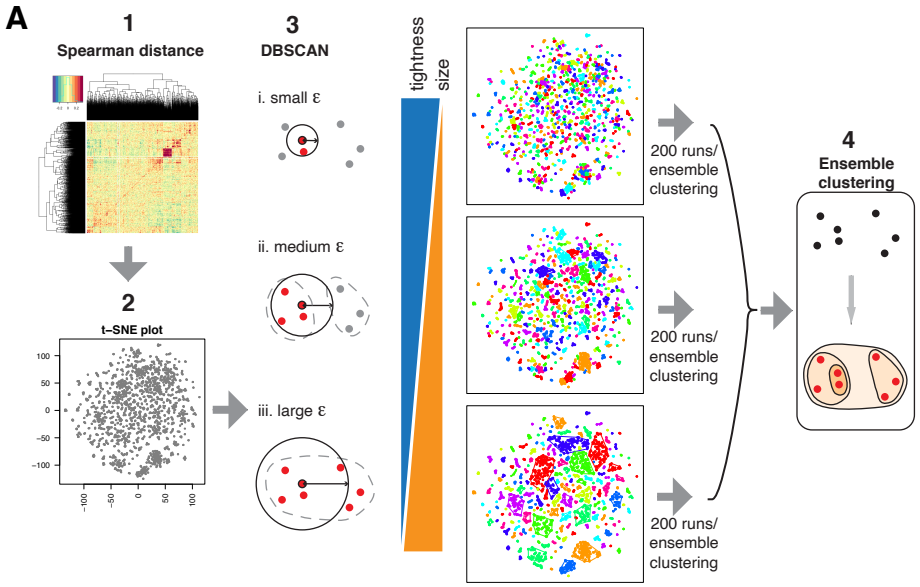


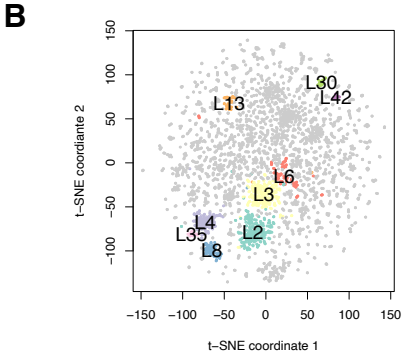
E



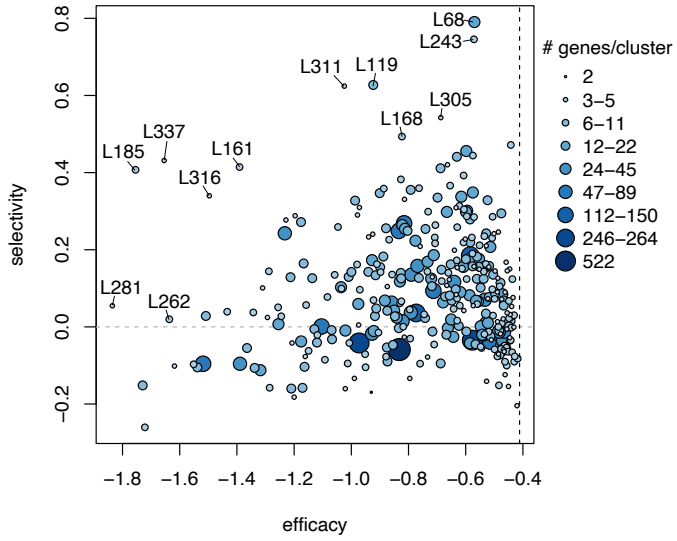
	CTNNB1	MYB	IRF4	KRAS	SOX10	CDK4	ISL1	PSMA1	PSMB5
# dependent lineages	1	1	1	1	1	0	0	17	17
# dependent cell lines	56	38	19	122	41	170	0	387	381
Efficacy	-1.652	-1.265	-1.297	-1.663	-1.31	-1.668	-0.554	-2.042	-1.993
Selectivity	0.844	0.495	0.717	0.632	0.546	0.668	0.367	-0.219	0.459

Figure 5



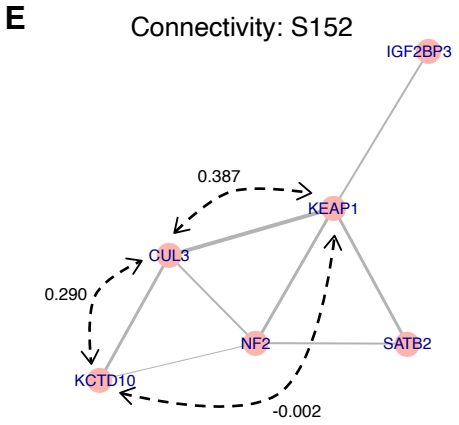


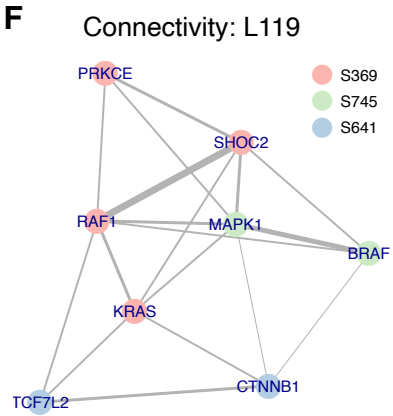
Cluster	Pathway	-logP
L4	Mitochondrial gene expression	178
L8	Respiratory electron transport	125.1
L13	Mediator complex	61
L30	DNA methylation	41.8
L2	rRNA processing in the nucleus and cytosol	37.9
L3	Processing of capped intron containing pre-mRNA	35.3
L6	mRNA splicing	33.7
L35	Formation of ATP by chemiosmotic coupling	33.4
L42	Interstrand cross link repair	33.1



D

Large	Medium	Small	Gene symbols	Efficacy	Selectivity
L68	M161	S152	CUL3,IGF2BP3,KCTD10,KEAP1,NF2,SATB2	-0.57	**0.791**
	M246	S265	FBXW11,KCTD5,PIK3R1,TSC1,TSC2		
	M648	S936	GSK3B		
L243	M431	S602	AHR,AIP,ARNT	-0.571	**0.746**
L119	M190	S369	KRAS,PRKCE,RAF1,SHOC2	-0.923	**0.627**
		S745	BRAF,MAPK1		
	M486	S641	CTNNB1,TCF7L2		
L311	M526	S688	IRF4,POU2AF1	-1.024	**0.624**
L305	M530	S748	CCDC6,FBXO42	-0.687	**0.543**
L168	M456	S599	PLAA,UBC,USP14	-0.823	**0.494**
	M519	S707	OTUD5,UBR5		
L337	M565	S741	DDX3X,RPS4X	-1.654	**0.431**
L161	M203	S500	CCND3,CDK6,EEF1A1	-1.391	**0.415**
		S829	CDAN1,DNAJC9		
L185	M222	S280	PSMB5,PSMB6,THAP1,TRIM8,ZGRF1	-1.754	**0.407**
L316	M619	S814	MAX,MYC	-1.497	**0.34**
L309	M502	S660	TMED10,TMED2	-0.972	**0.31**
L297	M513	S869	HUWE1,VCPIP1	-0.978	**0.29**
L210	M318	S415	INTS12,INTS6,INTS7,PRCC	-1.047	**0.259**
L281	M557	S892	MMS22L,TONSL	-1.835	**0.055**
L262	M434	S534	HAUS1,KIF18A,MSH2	-1.636	**0.02**





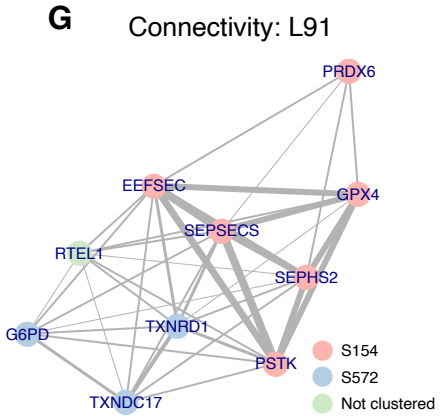
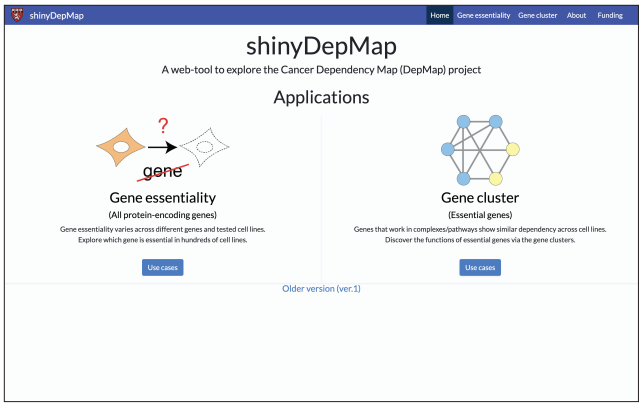
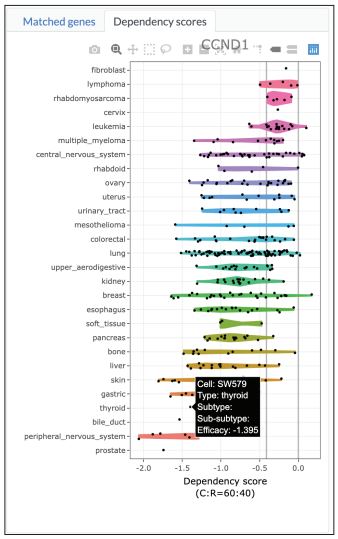


Figure 6

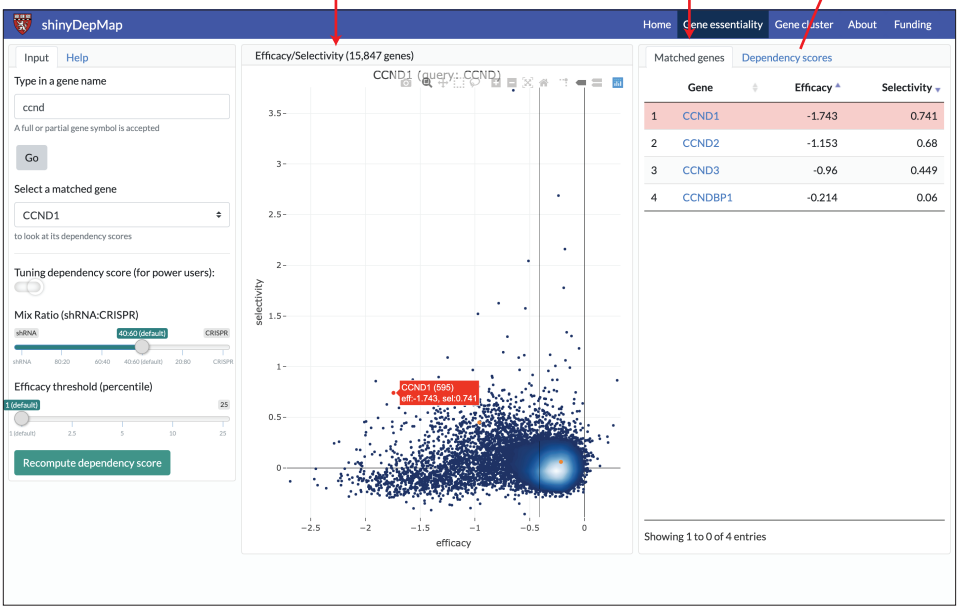
A Top page



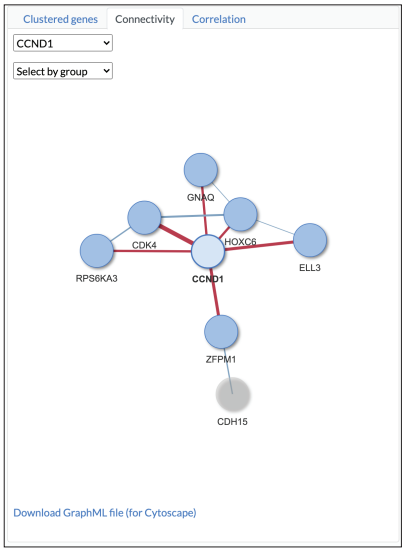
5



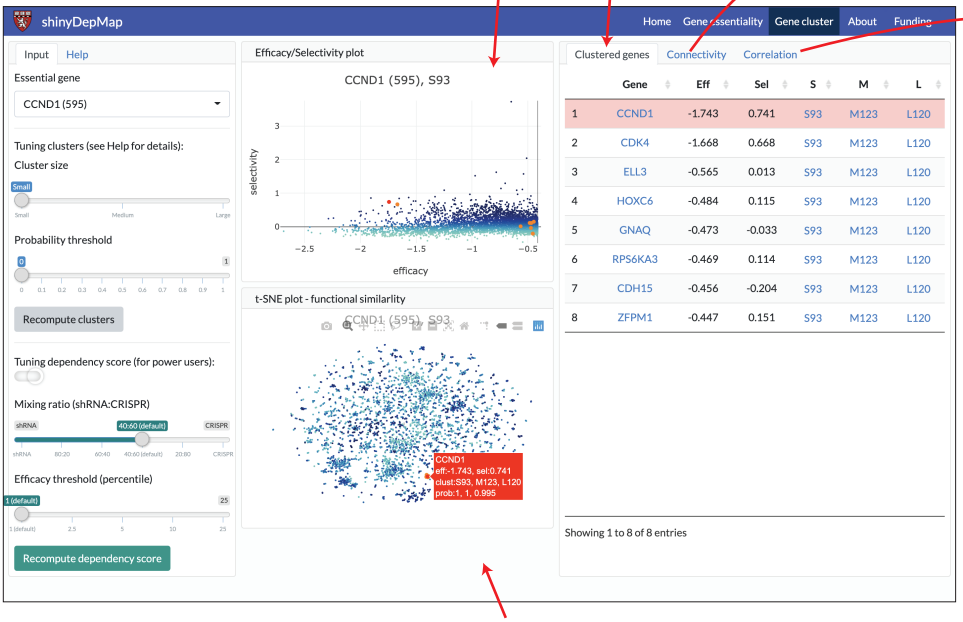
B. Gene essentiality app (all genes)



12



C. Gene clusters app (essential genes only)



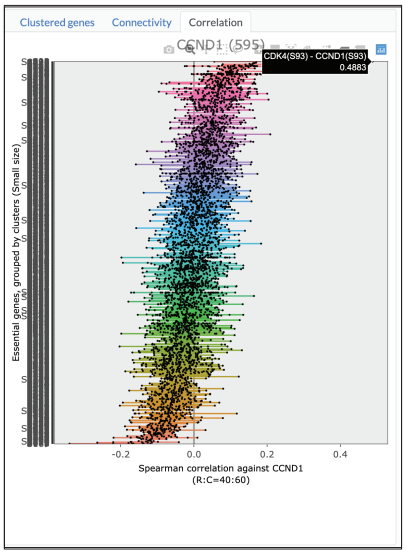


Figure 5 – figure supplement 1

ECHODOTS algorithm

Compute the Spearman distance ***D*** across all the essential genes ---------- (1)
For each ***d*** in 30:200
  For each ***i*** in 1:200
    Set ***i*** as the initial seed
    Run t-SNE with ***D*** as density matrix and get 2D coordinates ---------- (2)
    Compute the range of the data points in t-SNE map ***L(i)***
    Compute the neighborhood threshold $\varepsilon(i) = L(i)/d$ ---------- (3)
    Run DBSCAN with $\varepsilon(i)$ to find clusters for the essentail genes ***C(d,i)*** ---------- (4)
  Find consistent clusters ***CC(d)*** among ***C(d,i)***, $i = \{1,...,200\}$ ---------- (5)
Find the lower bound d_0 such that $d \geq d_0$ ---------- (6)
Pick three ***d*** and corresponding cluster sets ***CC(d)*** ---------- (7)

Figure 5 – figure supplement 2

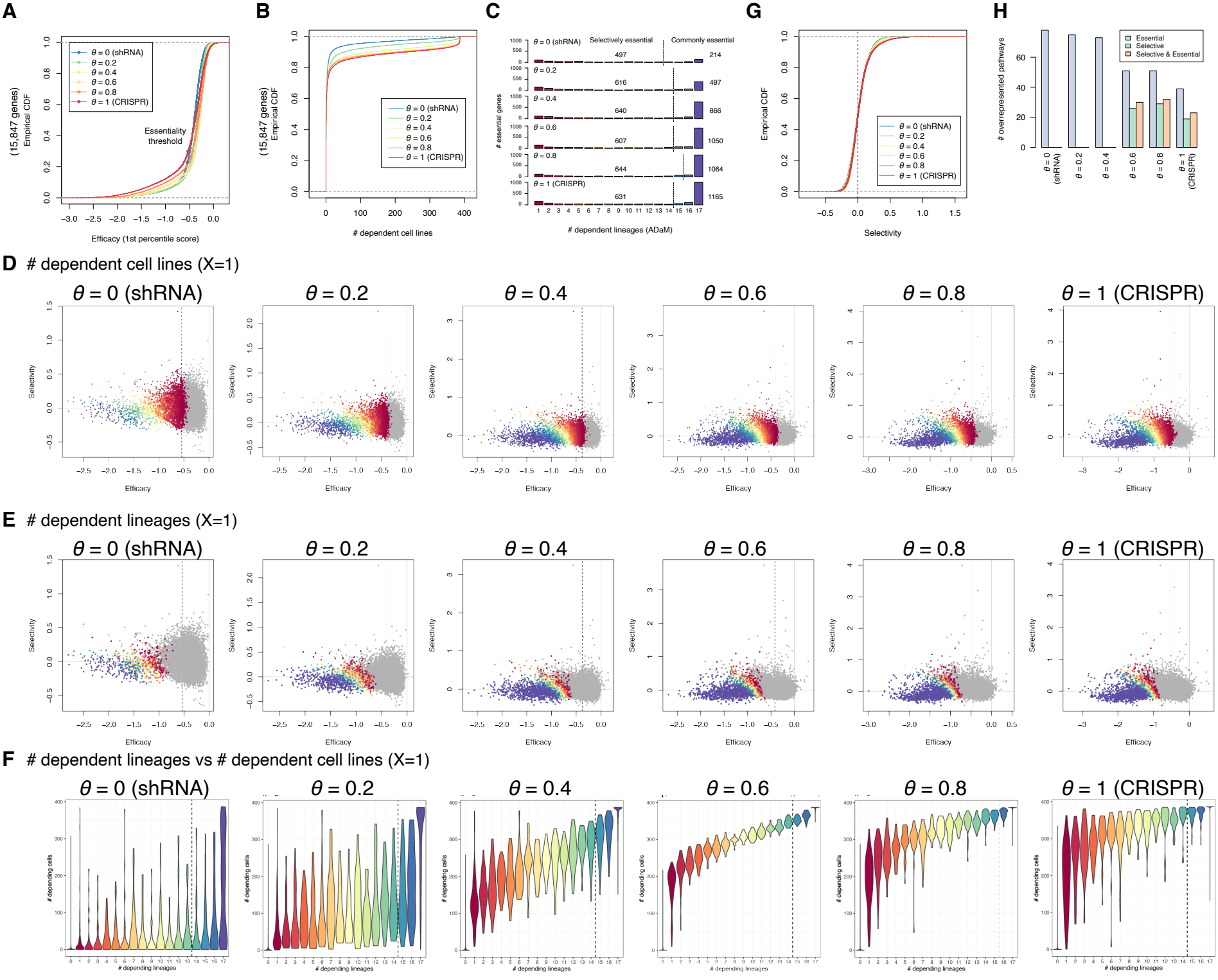


Figure 5 - figure supplement 3