Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration

James M McFarland[1], Zandra V Ho[1], Guillaume Kugener[1], Joshua M Dempster[1], Phillip G Montgomery[1], Jordan G Bryan[1], John M. Krill-Burger[1], Thomas M Green[1], Francisca Vazquez[1,2], Jesse S Boehm[1], Todd R Golub[1,2,3,4,5], William C Hahn[1,2,3,6], David E Root[1], Aviad Tsherniak[1]*

[1]Broad Institute of MIT and Harvard. [2]Dana-Farber Cancer Institute. [3]Harvard Medical School. [4]Boston Children's Hospital. [5]Howard Hughes Medical Institute. [6]Department of Medicine, Brigham and Women's Hospital.

* Corresponding author: aviad@broadinstitute.org

Abstract

The availability of multiple datasets together comprising hundreds of genome-scale RNAi viability screens across a diverse range of cancer cell lines presents new opportunities for understanding cancer vulnerabilities. Integrated analyses of these data to assess differential dependency across genes and cell lines are challenging due to confounding factors such as batch effects and variable screen quality, as well as difficulty assessing gene dependency on an absolute scale. To address these issues, we incorporated estimation of cell line screen quality parameters and hierarchical Bayesian inference into an analytical framework for analyzing RNAi screens (DEMETER2; https://depmap.org/R2-D2). We applied this model to individual large-scale datasets and show that it substantially improves estimates of gene dependency across a range of performance measures, including identification of gold-standard essential genes as well as agreement with CRISPR-Cas9-based viability screens. This model also allows us to effectively integrate information across three large RNAi screening datasets, providing a unified resource representing the most extensive compilation of cancer cell line genetic dependencies to date.

Introduction

Large-scale RNAi screens for cancer dependencies have recently been performed by multiple groups (Tsherniak et al. 2017; McDonald et al. 2017; Marcotte et al. 2016), providing systematic assessments of the effects of single-gene knock-down on cell viability, across a wide range of well-characterized cancer cell lines that are beginning to reflect the diversity of tumor types. By comparing genetic dependencies across cancer cell lines, researchers can thus identify specific cancer subtypes exhibiting a given vulnerability, as well as uncover new functional relationships between genes. In theory, integrating information across these separate RNAi datasets might greatly increase their utility -- both by providing the broadest coverage of cell lines and genes assayed, as well as by improving the accuracy and precision of individual gene dependency estimates. However, such integration requires addressing several computational challenges.

Firstly, the presence of substantial off-target effects mediated by the microRNA pathway (Jackson et al. 2006; Birmingham et al. 2006), as well as variable reagent efficacy, have long been recognized as challenges that can confound the interpretation of RNAi screening data. A number of methods have been developed to address these issues by utilizing robust statistics (König et al.

2007; Luo et al. 2008; Shao et al. 2013), mixed-effect models (Marcotte et al. 2016; Rämö et al. 2014), or explicit models of microRNA-mediated effects (Buehler et al. 2012; Schmich et al. 2015). Previously, we developed the DEMETER algorithm, a computational approach that models the 'seed-sequence' specific off-target effect of each shRNA directly, along with variable shRNA efficacy (Tsherniak et al. 2017). While DEMETER and related approaches (Shao et al. 2013) provide improved isolation of on-target gene-knockdown effects, they assess only the relative differences in gene dependency across cell lines. This limitation precludes identification of genes that are 'common essential' across cell lines, and makes direct comparisons of knockdown effects across genes difficult.

Another challenge with interpreting large-scale RNAi screens is that differences in screen quality between cell lines (as measured, for example, by the separation of positive and negative control gene dependencies) can confound comparisons of their genetic dependencies. Indeed, mRNA expression of AGO2, the catalytic component of the RNAi-inducing silencing complex (RISC), has been shown to correlate strongly with a cell line's screen quality (Hart et al. 2014), and is associated with the apparent essentiality of many genes (McDonald et al. 2017). This suggests that differences in the RNAi machinery between cell lines can bias quantification of the relative strength of their gene dependencies. Thus, it stands to reason that removal of such systematic 'screen-related' differences might provide more accurate estimation of the relevant patterns of genetic dependency.

Finally, the need to integrate RNAi screening datasets to generate a unified resource of cancer genetic dependencies raises several additional challenges. Such data integration requires analytical methods which can address batch effects at the cell line and shRNA level, and handle partial overlap of shRNAs and cell lines across datasets. Further, variable screen quality across datasets, and differences in the number and quality of reagents targeting each gene in different shRNA libraries, emphasize the need for statistically principled models which can efficiently integrate evidence across datasets to estimate gene dependency, as well as its associated uncertainty.

We thus developed a method, DEMETER2 (D2) that builds on the DEMETER model to address these challenges. We demonstrate this approach by applying it to three of the largest published RNAi datasets, showing that it improves gene dependency estimates and allows for effective integration of these data. The resulting combined dataset represents the most extensive compilation of cancer genetic dependencies to date and will facilitate discovery of therapeutic targets, as well as new cancer biology.

Results

Improved model of RNAi screening data

DEMETER2 is a model for large-scale pooled RNAi screening data, which takes as input measured changes in the relative abundance of pooled shRNA reagents across a panel of cell lines (Luo et al. 2008; Cheung et al. 2011; Cowley et al. 2014; Tsherniak et al. 2017; McDonald et al. 2017; Marcotte et al. 2016) and infers the effects of gene knockdown on the viability of each cell line. As in the original DEMETER (D1) model, DEMETER2 (D2) accounts for the depletion of each shRNA over time as a combination of the effects of suppressing the genes targeted by the shRNA, along with seed-based off-target effects determined by two 7-mer 'seed sequences' within each shRNA (Tsherniak et al. 2017). The DEMETER models also estimate the efficacy of each shRNA in eliciting these gene- and seed-effects.

D2 builds on the original D1 model by adding several additional components (**Fig. 1a**), summarized here (see Methods for details). First, D2 estimates a 'screen signal' parameter for each cell line, which accounts for overall differences in the relative strength of gene knockdown effects, such as due to variable RNAi efficacy (Vickers et al. 2007; Grimm et al. 2010; McDonald et al. 2017; Hart et al. 2014). The model also incorporates scaling and offset terms for each screen to account for global

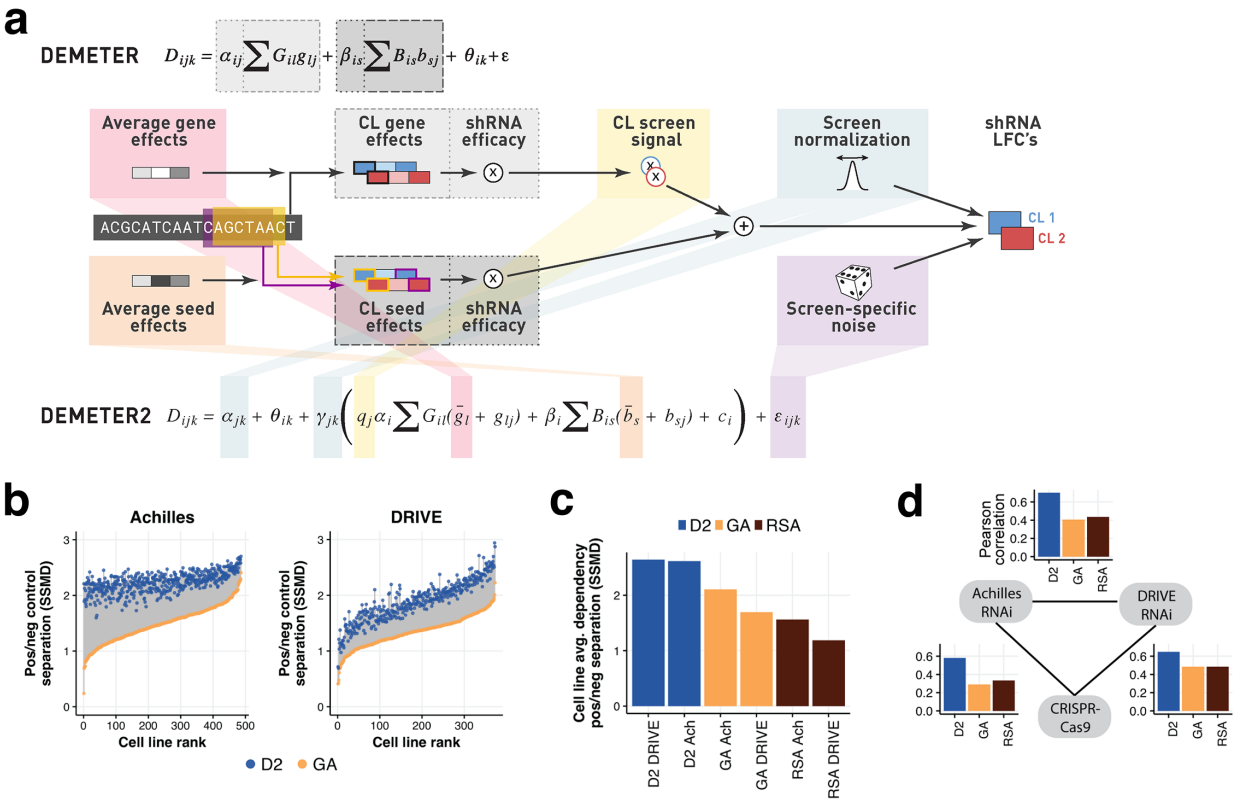

Figure 1: DEMETER2 improves identification of essential genes

a) Both D1 and D2 represent the observed shRNA log fold change (LFC) depletion values in each cell line (CL) as a combination of gene knockdown and off-target seed effects. D2 introduces a number of additional model components highlighted in the schematic diagram. **b)** Separation of gene dependency distributions for known common essential genes and non-essential (unexpressed) genes is measured by the strictly standardized mean difference (SSMD). Positive/negative control separation was much better for DEMETER2 gene dependency scores (blue dots) compared with per-gene averaging of shRNA depletion scores (GA; yellow dots) in both the Achilles (left) and DRIVE (right) datasets. **c)** D2 estimates of across-cell-line average gene dependency showed improved separation of positive and negative control genes compared with previous methods. **d)** Across-cell-line average gene dependency scores were in better agreement between datasets (Achilles RNAi, DRIVE RNAi, and CRISPR-Cas9 data) when using D2 estimates compared with previous methods. Each barplot shows the correlation of average dependency scores between a pair of datasets. Colors represent agreement when using different models for estimating dependencies from RNAi data.

differences in the distributions of shRNA depletion levels (such as those produced by differences in the number of population doublings and passages between measurements). Furthermore, D2 estimates the noise level associated with each screen to account for variable data quality.

Another key addition in D2 is the use of a hierarchical model for the gene and seed effects, allowing for efficient pooling of information across cell lines. Combined with shRNA-specific terms designed to capture measurement errors in the initial shRNA abundance and unaccounted-for off-target effects, these additions allow the model to accurately estimate gene dependency on an absolute scale (where a zero score represents no dependency) rather than a relative scale as in D1 (where a zero score represents the average dependency across all cell lines). Finally, D2 utilizes a Bayesian inference approach for parameter estimation which provides uncertainty estimates for the gene effects and other model parameters. In addition to facilitating comparisons of gene effects across cell lines and genes, where the precision of estimates can vary widely, these uncertainty estimates can be directly utilized in downstream analyses to improve their statistical power, as we demonstrate below.

D2 provides accurate estimates of absolute gene dependency

We first sought to compare D2 with existing methods in terms of its ability to identify genes which are essential in individual cell lines, as well as genes which are 'common essential' across cellular contexts. To this end we utilized two recently published RNAi datasets: the Broad Institute Project Achilles dataset (Tsherniak et al. 2017), which consists of 501 cell lines screened with 94k shRNAs targeting 17k genes (with a median coverage of 5 shRNAs per gene), and the Novartis DRIVE dataset (McDonald et al. 2017), which consists of 397 cell lines screened with 158k shRNAs targeting 8k genes (with a median coverage of 20 shRNAs per gene).

First, we measured the accuracy of dependency scores from each model by computing positive/negative control separation (measured by the strictly standardized mean difference, or SSMD), using a curated list of 'gold standard' common-essential genes (Hart et al. 2015) as positive controls, and genes that were unexpressed in each cell line as negative controls. Since DEMETER (Tsherniak et al. 2017) and ATARiS (Shao et al. 2013) only estimate relative differences in gene dependency across cell lines, they cannot be used for such analyses. Thus, we compare the performance of D2 with a simple approach that averages the depletion scores across shRNAs targeting each gene (gene averaging; GA).

Compared with GA, D2 provided much more accurate identification of essential genes in the Achilles dataset (SSMD increased by 58% on average, with improvement for all 486 cell lines tested; **Fig. 1b**). Less dramatic, though similarly consistent, improvements were also observed for the DRIVE dataset (SSMD increased by 42% on average, with improvement for all 373 cell lines tested). Furthermore, the improvements observed with D2 for both datasets were even larger when compared with the redundant siRNA activity (RSA) method (König et al. 2007) (**Fig. S1**), which was employed by the DRIVE study for identifying essential genes (McDonald et al. 2017).

The use of ~ 20 shRNAs per gene in the DRIVE dataset (compared to ~ 5 in Achilles) could ostensibly permit very effective extraction of the common on-target activities of same-gene shRNAs by average or RSA statistics. The robust improvement provided by D2 in this case highlights the benefits of model-based normalization, information-pooling across cell lines, as well as inference of shRNA efficacy, for identification of essential genes. Perhaps surprisingly, positive/negative control separation was lower overall for the DRIVE dataset compared to the Achilles dataset (**Fig. S2a-b**). This difference is likely due to the lower average on-target efficacy of shRNAs in the DRIVE library (**Fig. S2c**), which is consistent with the necessarily less selective design criteria needed to create a library with 20 shRNAs per gene. The extra information provided by additional shRNAs per gene nonetheless reduces false positive signals and improves estimates of differential gene dependency across cell lines (**Fig. S2d,e**).

We next evaluated the ability of these models to identify common essential genes, using the average dependency score across cell lines for each gene. For both the Achilles and DRIVE datasets, the application of D2 resulted in a much better separation of positive and negative control genes when assessing the average dependency of each gene (Hart et al. 2015), compared with both the GA and RSA methods (**Fig. 1c**). As a further test of the accuracy of average gene dependency estimates, we compared them with estimates obtained in a genome-wide CRISPR-Cas9 screening dataset (n = 391 cell lines), using the CERES model to correct for gene-independent DNA-cutting toxicity effects (Meyers et al. 2017). For the Achilles data, D2 estimates showed a two-fold increased correlation with CRISPR-based estimates compared with GA (Pearson r; D2 = 0.58; GA = 0.29). There was a similar, though less pronounced, improvement for the very high-shRNA coverage DRIVE dataset (D2 = 0.65; GA: 0.49; **Fig. 1d**). Furthermore, the agreement between Achilles and DRIVE estimates of average gene dependency was much higher with D2 (r = 0.70) compared with either GA (r = 0.41) or RSA (r = 0.44).

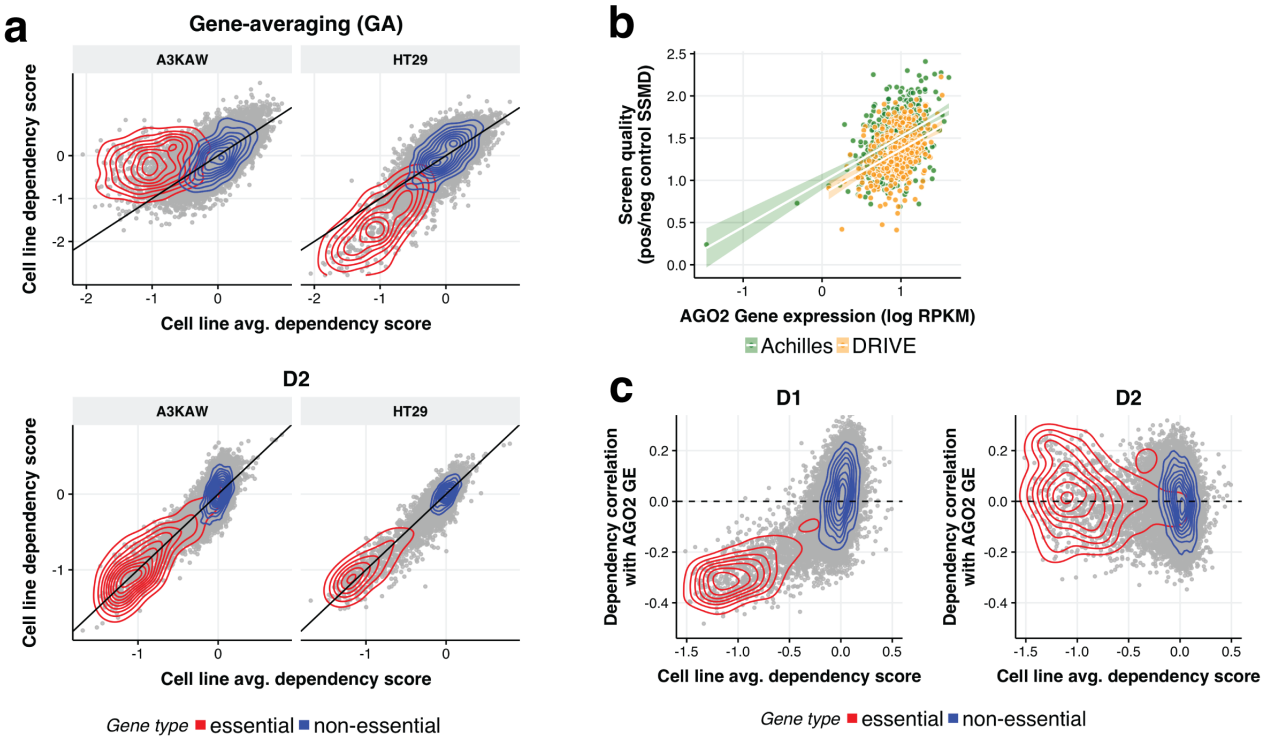

Figure 2: D2 corrects biases related to variable screen quality

a) Comparison of across-cell-line average gene dependency scores with scores estimated for individual example low- (left) and high- (right) quality screens. Density estimates for the set of gold standard common essential and non-essential genes are highlighted by the red and blue contours respectively. Estimates using gene-averaging (GA; top plots) show broad systematic differences across all essential genes in these cell lines compared with the population average. These systematic differences are corrected for by D2 (bottom plots). **b)** The screen quality estimated for each cell line (SSMD of positive/negative control gene dependencies, using GA) was correlated with the expression level of *AGO2* for both the Achilles (Spearman's *rho* = 0.39; $p<2.2\text{x}10^{-16}$; *green*) and DRIVE (*rho* = 0.37; $p<2.2\text{x}10^{-16}$; *gold*) datasets. **c)** Correlation between each gene's dependency profile and mRNA expression of *AGO2* is plotted against the across-cell-line average dependency score for the gene, with curated common-essential and non-essential genes indicated with red and blue dots respectively. Using D1 (left), gene dependency profiles were systematically (negatively) correlated with *AGO2* expression for more common essential genes. This correlation was eliminated using D2 (right).

Thus, D2 addresses a key limitation of previous methods (Shao et al. 2013; Tsherniak et al. 2017) by providing estimates of gene dependency on an absolute scale, allowing direct comparison across genes. Furthermore, D2 greatly improves identification of common-essential genes compared with existing approaches.

D2 corrects screen-quality biases

As shown in **Fig. 1b**, there were large differences in the quality of screening data across cell lines (in terms of the separation of positive and negative control gene dependencies), in both the Achilles and DRIVE datasets. When using existing methods, cell lines with lower screen quality appear to be systematically less dependent on nearly all common essential genes, and conversely for cell lines with high screen quality (as illustrated for example low- and high-quality screens in **Fig. 2a**). Such global differences are likely due to assay-specific technical factors, rather than real differences in genetic dependencies. Indeed, differences in screen quality were associated with expression of *AGO2* (**Fig. 2b**), the catalytic component of the RNA-induced silencing complex (RISC), suggesting they reflect variation in the efficacy of the underlying RNAi machinery across cell lines.

As demonstrated below, these differences can lead to substantial confounding effects in downstream analyses. To address this problem, D2 infers a 'screen signal' parameter for each cell line, and effectively removes this

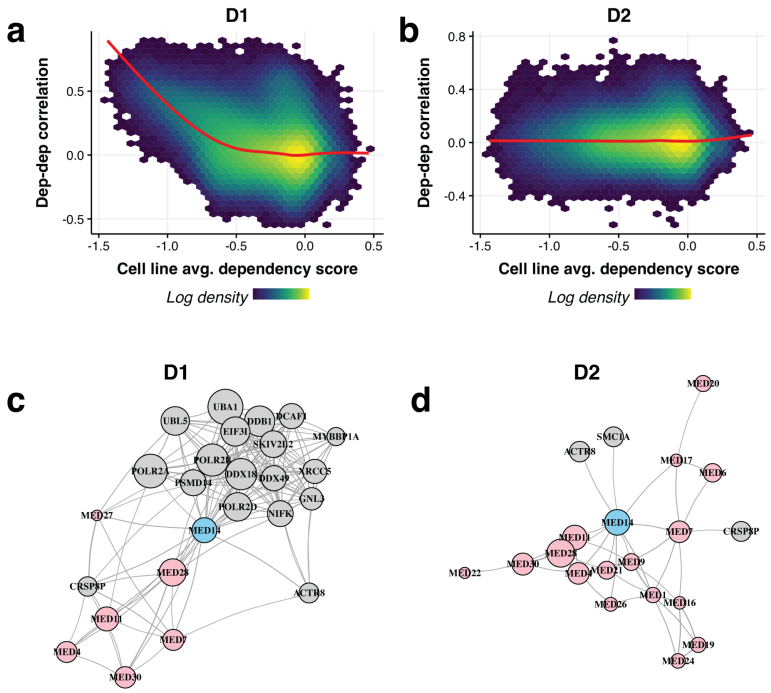

Figure 3: Screen quality biases impair dependency correlation analyses

a) Correlation between pairs of gene dependency profiles, estimated using D1 applied to the Achilles data, increased systematically for gene pairs that were more essential on average. Red line shows the smoothed average. Color shows the density of data points in each region (log color scale). **b)** Same as **a** but using D2, showing that the systematic upward bias in pairwise correlation between pairs of common essential genes is removed. **c)** Gene dependency correlation network surrounding *MED14*, constructed using D1 dependency estimates applied to the DRIVE dataset. Each node represents a gene, with edges depicting strong pairwise correlations (see Methods). Red nodes indicate genes that form protein complexes with *MED14*. Node size indicates the across-cell-line average dependency score for each gene (larger nodes representing more common-essential genes). **d)** Same as **c** using D2, showing that the local dependency correlation network for *MED14* is more enriched for co-complex members.

source of bias from the estimated gene dependency scores. The model-inferred screen signal parameters are closely related to measured differences in screen quality (**Fig. S3a**). They also show good agreement when estimated independently from the Achilles and DRIVE datasets (**Fig. S3b**), suggesting that they capture robust differences in how different cell lines behave in RNAi screens. By estimating and accounting for these differences in screen signal, the gene dependency scores estimated by D2 for the same example cell lines no longer show systematic deviations from the all-cell-line average (**Fig. 2a, bottom**).

To show the magnitude of the effect of screen-quality related biases on estimated gene dependencies, as well as their successful removal with D2, we calculated the correlation across cell lines between *AGO2* mRNA expression and dependency scores for each gene. When using D1, the dependency profiles of many genes showed strong anticorrelation with *AGO2* expression, and the strength of anticorrelation was systematically increasing for genes that were more essential on average (**Fig. 2c**). In contrast, D2 dependency profiles showed little correlation with *AGO2* expression, even for common essential genes. These screen-quality related biases were not specific to D1, but rather were present to similar degrees using other methods, as well with both the Achilles and DRIVE datasets (**Fig. S4**). While screen-quality related biases could in principle be removed by simply rescaling each cell line's gene dependency scores to better align data across cell lines for positive and negative control genes, this process dramatically magnifies the effects of noise in the lowest quality cell lines, resulting in substantially poorer results for downstream analyses (**Fig. S5**).

As an illustration of how such biases can impact downstream analyses, we computed correlations between the patterns of dependency across cell lines for each pair of genes. Such dependency correlation analyses have recently been shown to provide a powerful mechanism for identifying functional relationships as well as physical interactions among genes (Wang et al. 2017; McDonald et al. 2017; Tsherniak et al. 2017). Since the screen-quality related biases are shared across genes, they can significantly influence estimates of dependency correlations, artificially inflating correlations between common essential genes. Indeed, we found that

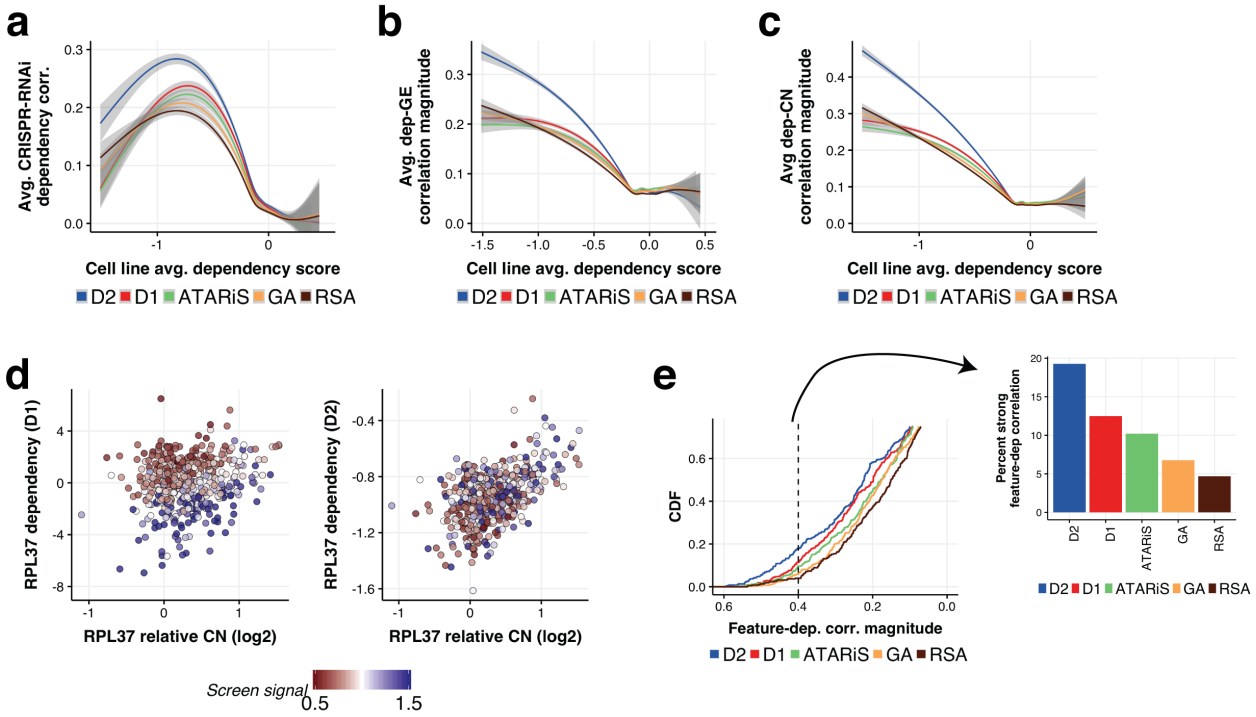

Figure 4: D2 improves estimated dependency profiles, particularly for essential genes

a) Average correlation between RNAi and CRISPR-Cas9 gene dependency profiles as a function of the across-cell-line average dependency score, using the Achilles dataset. Different colored curves and shaded regions show the smoothed conditional mean correlation, and standard error estimates, obtained using different models for estimating RNAi gene dependencies. D2 gene dependency estimates show better average agreement with CRISPR-Cas9 dependency profiles compared to existing methods. **b)** Average magnitude of pairwise correlations between gene dependency and mRNA expression profiles for each gene (again for the Achilles dataset), plotted as a function of average gene dependency as in **a**. D2 dependency scores showed stronger correlation with the gene's own expression levels compared with existing methods. **c)** Similar to **b**, showing stronger correlations between D2 dependency profiles and the genes' own relative copy number, particularly for genes which are more essential on average. **d)** Scatterplot of *RPL37* dependency vs. *RPL37* relative copy number using D1 (left) and D2 (right) dependency scores. Color represents the screen signal parameter estimated (from D2) for each cell line. **e)** A benchmark set of dependency-genomic feature relationships identified from CRISPR-Cas9 data (see Methods) was used to evaluate the extent to which Achilles RNAi dependency estimates recapitulated the same associations. Colored curves show the empirical distributions of correlation magnitude across these dependency-feature pairs for each model. D2 dependency estimates showed better agreement with benchmark genomic feature associations compared to existing methods. Barplot at right shows the fraction of dependency-feature pairs with correlation magnitude greater than 0.4 for each model.

dependency correlations estimated using D1 increased systematically for gene pairs that were more pan-essential (**Fig. 3a**). Again, this relationship was not particular to D1, being present with other approaches such as ATARiS and GA (**Fig. S6**). In contrast, dependency correlations estimated using D2 showed no such bias (**Fig. 3b**). To highlight how D2 can improve identification of functional interactions between genes, we consider an example co-dependency network computed for the mediator complex gene *MED14* (see Methods). When using D1, *MED14* was connected with several other mediator complex genes, as expected. However, as *MED14* is essential in many cell lines it also connected strongly with a large group of common essential genes (**Fig. 3c**). When using D2, connectivity with the group of common essential genes was removed, and the co-dependency network of *MED14* was much more selective for other members of the mediator complex (**Fig. 3d**).

D2 improves estimates of relative gene dependency

Of particular interest for understanding cancer genetic dependencies is the ability to identify differences across cell lines, such as dependencies associated with a particular subtype of cancer, or with a particular biomarker. To assess the ability of D2 to accurately

estimate relative differences in dependencies across cell lines, we compared the dependency profiles across cell lines for each gene with dependency estimates derived from the Achilles CRISPR-Cas9 dataset (Meyers et al. 2017) for the same genes and cell lines. While there may be some differences in the consequences of shRNA-mediated versus sgRNA-mediated effects, in general better agreement with CRISPR-based gene dependency profiles should reflect an improved ability to isolate and quantify on-target gene-knockdown effects. We found that D2-based estimates of gene dependency were in better agreement with those derived from CRISPR data compared with D1, ATARiS, and GA in both the Achilles (**Fig. 4a**) and DRIVE data (**Fig. S7a**). As expected regardless of method, correlations between the RNAi and CRISPR dependency estimates were better for genes with stronger viability effects overall. Consistent with the greater impact of screen-quality related biases for essential genes (**Fig. 2**), the improvements in D2 compared with other methods were systematically larger for genes that were more essential on average.

We also tested the ability of different models to recover expected relationships between genetic dependencies and genomic features using several approaches (see Methods). First, we computed the correlation between each gene's dependency and its mRNA expression levels, as well as its relative copy number. D2-based dependency estimates showed significantly higher correlation with the gene's own expression (**Fig. 4b**), and copy number (**Fig. 4c**). In both cases, differences between D2 and other models were again most pronounced for genes that were more essential on average. These common essential genes typically show a strong positive correlation between dependency scores and the genes' copy number and expression levels, reflecting the CYCLOPS relationships (Nijhawan et al. 2012; Tsherniak et al. 2017), where partial loss of an essential gene renders a cell more sensitive to further suppression of the gene. As an example, we consider dependency profiles estimated for the common-essential ribosomal gene *RPL37*. When using D1, *RPL37* dependency is weakly correlated with *RPL37* copy number (Peason's r = 0.18), instead showing a much stronger correlation with inter-cell line differences in screen quality (r = -0.69; **Fig. 4d**). In contrast, D2 largely removes the spurious association with screen signal (r = 0.12), resulting in an estimated *RPL37* dependency profile that is much more highly correlated with copy number variations (r = 0.50).

Finally, we tested the ability of D2 to recapitulate the most robust relationships between genetic dependencies and genomic features (dependency-feature pairs) identified from the CRISPR-Cas9 dataset (see Methods). When using D2, we found significantly stronger correlations between these dependency-feature pairs compared with D1 for both Achilles (p=1.4x10^{-9}, Wilcoxon signed rank test, n = 384 pairs; **Fig. 4e**) and DRIVE (p=8.2x10^{-4}, n = 231 pairs; **Fig. S7d**) data. Furthermore, the improvements provided by D2 were even more pronounced compared with other methods such as ATARiS, RSA and GA.

The ability of D2 to estimate the uncertainty of each gene dependency score allows for downstream analyses that account for the large differences in uncertainty across cell lines and genes. As expected, D2's estimates of dependency uncertainty were closely related to the screen quality and replicate agreement of each cell line, as well as differences in the number and quality of shRNAs targeting each gene (**Fig. S8a-c**). In fact, the analyses presented here were all performed by weighting each gene dependency score by its associated precision (see Methods). Given the large differences in uncertainty across cell lines and genes, there is good reason to expect improvements when using analyses that appropriately account for the uncertainty of each measurement. Indeed, we found that the agreement between RNAi and CRISPR-Cas9 dependency profiles improved when accounting for the gene dependency uncertainty (**Fig. S8d**). Nevertheless, D2-based results still provided improvements over other methods across all metrics, even without utilizing uncertainty-aware analysis methods.

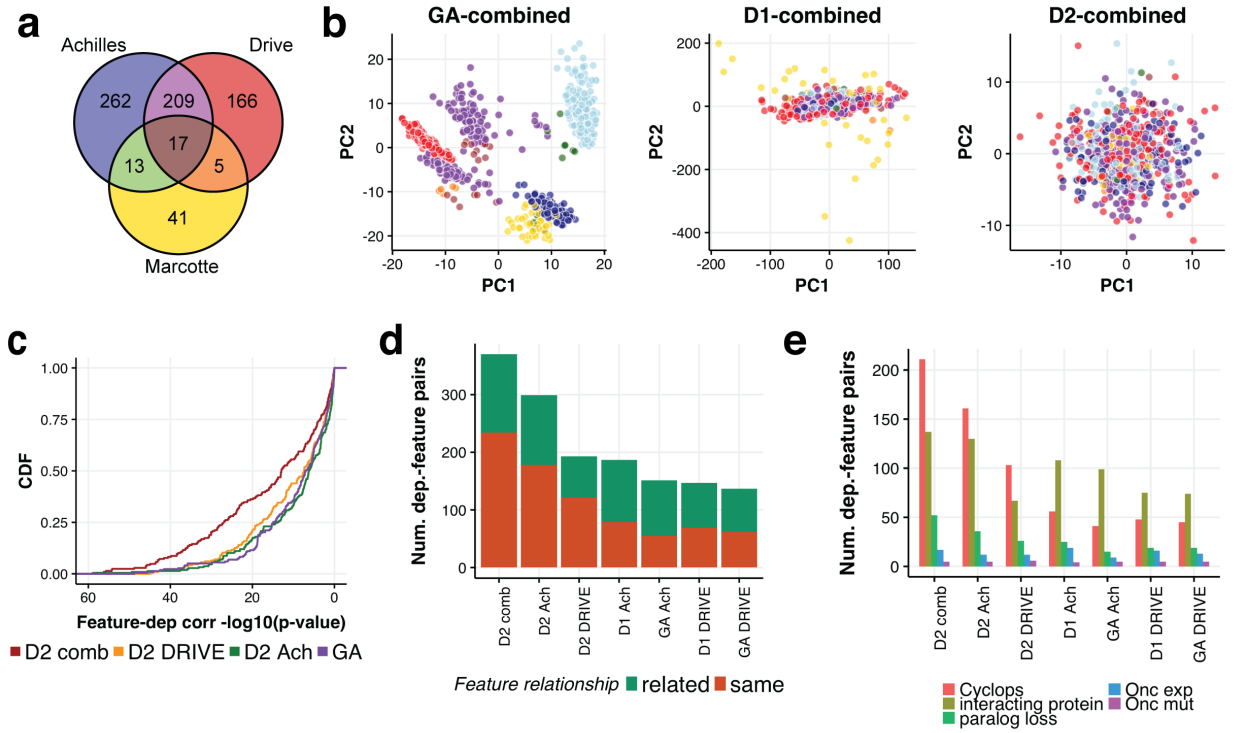

Figure 5: D2 effectively integrates multiple RNAi screens

a) Venn diagram showing the overlap of cell lines screened across the DRIVE, Achilles and Marcotte et al. datasets. There were 713 unique cell lines assayed. **b)** D2 combines datasets with minimal batch effects compared with merging separate D1 estimates or computing pooled GA estimates (see Methods). Each plot shows the first two principal components of the gene dependency data, with different colors representing which set of experiments were used to screen each cell line. Color scheme is the same as **a**, though light and dark blue indicate the cell lines screened with the '98k' and '55k' libraries in the Achilles dataset. **c)** The statistical significance of measured associations between the benchmark dependency/feature pairs (same set as in **4e**) is greatly increased when using the combined D2 dataset. Plot shows the empirical CDF of negative log p-values for dependency-feature associations using each model. **d)** The combined D2 dataset allows for identification of a greater number of biologically plausible dependency-feature relationships compared with using either the DRIVE or Achilles dataset alone. The top correlated genomic feature is computed for each gene dependency, and the bar plot depicts how many genes have a best genomic feature that is known to be biologically related (either the same gene, a sequence paralog, or a physically interacting gene; see Methods). **e)** Dependency-feature associations identified by each model are classified as CYCLOPS, oncogene expression, oncogene mutation, paralog loss, and physical interactions. The D2 combined model identifies more relationships in nearly every class compared to using the individual datasets or other models.

These observations show that D2 improves estimates of relative differences in dependency across cell lines, allowing for improved identification of genomic features associated with genetic dependencies.

Integration of multiple RNAi datasets

The availability of the Achilles and DRIVE RNAi datasets, as well as other genome-scale RNAi screens raises the possibility of integrating these data to provide a single comprehensive set of gene dependency estimates. However, such data integration is challenging due to systematic differences between screens (batch effects), variable noise levels, and partial overlaps of the cell lines and shRNAs used. The ability of D2 to estimate batch specific normalization factors and noise parameters in a statistically principled modeling framework make it well-suited to address these challenges.

We applied D2 to the combined Achilles and DRIVE datasets, along with a recently published genome-scale screen of 76 breast cancer cell lines (Marcotte et al. 2016). The combined dataset represents 713 unique cell lines, and 241k unique shRNAs (**Fig 5a**). A primary concern when merging data across multiple datasets in this way is that large dataset-related differences (batch effects) will remain in the combined dependency

estimates, confounding downstream analyses. Indeed, other methods of integrating across datasets -- such as computing per-gene averages on the normalized and pooled data, or averaging together separate D1-based estimates from each dataset -- produce results with strong screen-related batch effects (**Fig. 5b**). In contrast, these batch effects are greatly reduced when D2 is used to integrate multiple datasets, as evidenced by a large reduction in the fraction of variance captured by the first two principal components (7.3% with D2, compared with 21% and 18% for GA and D1 respectively).

We next sought to determine whether the integration of these RNAi datasets using D2 provided improved results due to the increased sample size and/or more accurate dependency estimates. Indeed, the combined D2 dataset showed small but consistent improvements in the dependency estimates for genes and cell lines screened in both Achilles and DRIVE, despite the fact that the total number of shRNAs targeting these genes was only marginally increased in the combined dataset compared with using the DRIVE dataset alone. The combined D2 model nevertheless showed better agreement with CRISPR-Cas9 estimates of per-gene average gene dependencies (**Fig. S9a**), as well as dependency profiles across cell lines (**Fig. S9b**). The most significant advantage offered by the integration of these RNAi datasets, however, is the increased coverage of cell lines, and the resulting increase in statistical power to identify relationships in the dependency data. For example, the same benchmark set of dependency - genomic feature relationships identified using CRISPR-Cas9 data (as in **Fig. 4e**) were identified with much higher statistical significance using the combined D2 dataset compared with using the individual datasets (**Fig. 5c**).

As a further test of the power of the combined D2 dataset for identifying relationships between dependencies and genomic features, we performed a simple global analysis whereby we determined the genomic feature that was most strongly correlated to each gene's dependency profile (considering mRNA expression, copy number and mutation features). We then asked whether the top correlated genomic feature was associated with the same gene as the dependency, or with a gene that was known to be biologically related (considering physical interactions, CORUM protein complex membership, and sequence paralogues; see Methods), using such known biological relationships as a proxy for correctly identified dependency-feature relationships.

Overall, when using the combined D2 dataset the number of such dependency-feature relationships identified was 92% and 24% larger respectively compared with applying D2 to either the DRIVE or Achilles datasets alone (**Fig. 5d**). The number of relationships identified with the combined D2 dataset was 2.0-2.7 fold larger compared with applying either D1 or gene-averaging to the Achilles or DRIVE datasets. To better understand the source of these improvements, we categorized these dependency-feature relationships as 'CYCLOPS', 'oncogene expression', 'oncogene mutation', 'paralog deficiency' and 'interacting protein' (Methods). The combined D2 dataset provided improvements in identifying nearly all categories of relationships compared with using D2 on the individual datasets (**Fig. 5e**). Compared with the D1 and GA datasets, the combined D2 model provided the most dramatic improvements in identifying CYCLOPS relationships (e.g. 3.8-fold and 4.4-fold increases respectively compared to D1 Achilles and DRIVE data), reflecting the fact that correction of screen-quality bias had the largest impact on common essential genes which tend to show such relationships. However, the combined D2 model identified substantially more relationships in other categories as well (e.g. 2.1-and 2.7-fold more 'interacting protein' relationships compared with the D1 Achilles and DRIVE datasets respectively). Thus, the combined D2 dataset can substantially increase the utility of existing large-scale RNAi datasets for identifying genetic dependencies and their associated genomic features.

Discussion

We present an improved model (DEMETER2) for inferring cancer cell line genetic dependencies from RNAi screens, and show that it provides significant improvements over existing methods across a range of performance measures when applied to both the Broad Institute Achilles (Tsherniak et al. 2017) and Novartis DRIVE (McDonald et al. 2017) datasets. The D2 model also allows for effective data integration, and we apply it to combine three recently released RNAi screening datasets (McDonald et al. 2017; Tsherniak et al. 2017; Marcotte et al. 2016) to produce the largest compilation of cancer cell line genetic dependencies to date, comprised of 713 unique cell lines. We provide these data, along with the source code used to generate them, as a resource at https://depmap.org/R2-D2.

The predecessor of D2, DEMETER, was designed to address the strong off-target effects, and variable shRNA quality, which are well-known to confound interpretation of RNAi screening data. Here we show that differences in screen quality between cell lines can pose additional challenges for efforts to map genetic vulnerabilities in cancer (Tsherniak et al. 2017; McDonald et al. 2017), by confounding comparisons of dependencies across cell lines. For example, when using existing methods, common essential genes appear to be systematically stronger dependencies in cell lines with higher screen quality, biasing downstream analyses such as estimation of gene-gene co-dependencies, and identification of molecular features predictive of dependency. D2 addresses this problem by incorporating explicit estimation of multiple screen normalization parameters from the data. Estimated 'screen signal' parameters capture differences in the strength of gene suppression achieved in each cell line, and were remarkably reproducible when estimated from independent RNAi datasets. These parameters were also correlated with expression of *AGO2*, a key component of the RNAi pathway, suggesting they reflect intrinsic differences in RNAi efficiency among cell lines (Vickers et al. 2007; Grimm et al. 2010; McDonald et al. 2017). Additional model-inferred normalization parameters captured inter-screen differences in the overall scale of shRNA depletion measurements which were correlated with differences in the cell lines' measured growth rate (**Fig. S10**). The D2 model thus identifies and removes multiple sources of cell line- and screen-related systematic bias in order to facilitate direct comparisons of genetic dependencies across cell lines. Surprisingly, the improvements gained by modeling such screen quality differences were often as large as those gained by accounting for off-target effects and variable shRNA efficacy (see, e.g., **Figs. 4**, **5**, and **S7**), which have been the focus of nearly all previous attempts to model RNAi screening data. This general approach – using a hierarchical modeling framework to pool information across cell lines, coupled with model-based normalization – could be used to correct for similar sources of systematic bias in other functional screening assays, including CRISPR-Cas9 knockout screens.

Another limitation of previous models designed to address RNAi off-target effects (Tsherniak et al. 2017; Shao et al. 2013) is that they only provide estimates of the relative differences in gene dependency across cell lines, precluding identification of common essential genes and direct comparisons of dependency scores across genes. On the other hand, methods such as RSA (König et al. 2007) are more targeted towards calling essential genes in individual screens, such that multiple methods can be required to assess different aspects of RNAi screening data (McDonald et al. 2017). D2 addresses both of these use-cases by directly estimating gene dependency on an absolute scale that is comparable across genes and cell lines. Furthermore, we found that identification of essential genes was much improved using D2 compared to previous methods across all cell lines tested, and its estimates of across-cell-line average gene dependencies showed better agreement with curated common essential genes (Hart et al. 2015), as well as with CRISPR-Cas9 based estimates.

We previously processed the Achilles dataset of 501 RNAi screens with D1 to identify 769 genes of interest that show a strong differential dependency pattern across the cell lines (Tsherniak et al., 2017). As D1 generates only relative dependency scores, we used a cut-off of six global standard deviations from the mean to define this set ("six-sigma dependencies"). With the availability of absolute dependency scores from D2, more refined and stable approaches can be used to identify genes showing dependency patterns of interest. Notably, when we comparably analyzed the D2 Achilles dataset (using a matched threshold of 5.2 sigma to give an equal proportion of outlier genes) 57% of the previously reported genes that have D2 scores are re-identified. The main reasons for the discrepancy are the the instability of the six-sigma metric used (a gene is called a six-sigma dependency even if it is a six-sigma dependency in a single cell line), and the screen quality bias-correction utilized by D2. We thus anticipate that the application of D2, and the availability of large loss-of-function datasets from CRISPR-Cas9 screens will permit further improvements in the identification of differential dependencies.

Finally, we note that the D2 model bears several similarities to the siMEM method (Marcotte et al. 2016) which uses a mixed-effects modeling approach to account for differential gene dependencies across cell lines and variable shRNA efficacy, and can model datasets with more than two time points. In contrast to D2, however, siMEM is designed to test associations between a given gene's dependency and a particular cell line phenotype. D2 facilitates a broader range of possible analyses by providing accurate and directly-comparable dependency estimates for each gene and cell line in a unified model. Furthermore, by directly modeling shRNA off-target effects and accounting for screen-quality related biases, D2 also provides an accurate assessment of any dependency-phenotype associations of interest (see, e.g., **Figs. 4, 5**).

In addition to improving gene dependency estimates compared with previous methods, D2 allows for an effective integration of data across multiple RNAi datasets. The resulting integrated RNAi dataset both improves the quality of gene dependency estimates, and also maximizes the coverage of cellular contexts and genes assayed, increasing the power of these datasets for discovering patterns in cancer cell genetic dependencies. We illustrate this by showing that the combined D2 dataset substantially increases the number of dependency-feature relationships identified compared with previous models, as well as with using the individual RNAi datasets. In summary, our results show that the combined RNAi dataset produced by D2 is a valuable resource that will greatly extend the utility of existing RNAi screening data.

Methods

Data processing

For maximal consistency, we reprocessed raw shRNA read counts data for the DRIVE and Achilles datasets using the same pipeline. This consisted of first normalizing the counts data for each sample by computing the log counts per million, using the function 'cpm' from the R package edgeR with a 'prior counts' value of 10. Any shRNAs that did not have a log counts per million of at least 1 in the plasmid DNA were removed from analysis. We then normalized each shRNA abundance by its associated value in the plasmid DNA sample to get log-fold change (LFC) estimates for each shRNA in each sample. These values were median-collapsed across replicates for each sample to get the LFC data serving as input to the models.

Achilles RNAi dataset

For Achilles data, the plasmid DNA measures were shared across samples within each of the three batches (Tsherniak et al. 2017). Hence, we used the replicate-collapsed plasmid abundances for each batch to estimate LFC values, as well as to identify shRNAs with insufficient plasmid representation, for all samples in the batch.

DRIVE dataset

The raw DRIVE shRNA counts data (v4) were downloaded from https://data.mendeley.com/datasets/y3ds55n88r/4 (McDonald et al., 2017). The cell line 'f36p' was first removed from all analyses because it was a clear outlier in the number of sample read counts, with about 10 times the number of shRNAs exhibiting 0 counts compared to any other cell line. Any plasmid DNA measurement with insufficient counts was replaced by a 'virtual library', calculated as described in (McDonald et al. 2017). LFC values for shRNAs with the same targeting sequence in the same experiment were median-collapsed, along with any technical replicates, to obtain unique plasmid and sample counts for a given sequence in each pool and cell line.

Marcotte et al. dataset

The raw data used to generate LFC estimates for each shRNA and cell line were downloaded from the link provided in (http://neellab.github.io/bfg/; files used: "Normalized ExpressionSet", "updated shRNA annotations"). The file mapping probes to their sequences was downloaded from GEO (GSE74702. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL21133). Probe sequences were mapped to their ids using the updated shRNA annotations file.

LFC values were computed by taking the difference of t_2 and t_0 log2 measurements per replicate, and median-collapsing across replicates. For cases where a replicate had a t_2 measurement with no matching t_0 measurement, t_0 for this replicate was inferred by taking the mean t_0 measurement across all cell lines for that shRNA. We subsequently filtered out 8960 hairpins that Marcotte et al. identified with a t_0 measurement below a noise threshold (Tsherniak et al. 2017; McDonald et al. 2017; Marcotte et al. 2016). In addition, the cell line HCC1428 did not have t_2 measurements recorded and was therefore excluded from the analysis.

Mapping shRNAs to genes

Gene mappings were found by performing an exact string search through all RefSeq transcript RNA sequences (both protein-coding and non-coding) downloaded on June 28, 2017 from:

ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/mRNA_Prot/human.*.rna.gbff.gz

For a given shRNA, the query sequence used in this search was the initial 19mer of the 21mer "target sequence". The final 1-2 bases of the 21mer tend to be cleaved off in the cell as an shRNA is processed into an siRNA, and thus they do not contribute to its targeting specificity. A shRNA was mapped to a gene if its initial 19mer was an exact match to *any* of the gene's transcripts in this search.

DEMETER2

Model description

DEMETER2 (D2), as with DEMETER1 (D2), seeks to explain the observed shRNA depletion in each sample as a combination of gene knockdown effects and off-target seed effects. D2 expands the D1 model in several ways, as described below (and illustrated schematically in **Fig. 1a**). Let the D1 model in several ways, as described below (and illustrated schematically in **Fig. 1a**). Let D_{ijk} represent the depletion score measured for shRNA *i* in cell line *j* and dataset *k*, the D2 model is then given by the following equation (a complete description of the model parameters is given in **Table S1**):

$$D_{ijk} = a_{jk} + \theta_{ik} + \gamma_{jk}\left(q_j \alpha_i \sum_l G_{il}(\bar{g}_l + g_{lj}) + \beta_i \sum_s B_{is}(\bar{b}_s + b_{sj}) + c_i \right) + \epsilon_{ijk}$$

In D2, the gene knockdown effect in a given cell line is explicitly modeled as a sum of two components: $\bar{g}_l$, the across-cell-line average effect for gene *l*, and g_{lj} a component specific to cell line *j*. Similarly, the effects associated with seed sequence *s* are represented by the across-cell-line average ($\bar{b}_s$) and cell-line specific (b_{sj}) effects of seed sequence *s*. This hierarchical model structure allows for information sharing across cell lines when estimating across-cell-line average gene and seed effects, while still effectively capturing inter-cell line variation. The set of genes $\{l\}$ and seeds $\{s\}$ targeted by a given shRNA are determined by the elements (G_{il}, B_{is}) of fixed binary matrices that encode the shRNA-to-gene and shRNA-to-seed mappings.

As with the D1 model, each shRNA is assigned two seed sequences, given by positions 1-7 and 2-8 on the antisense strand. Similar to the D1 model, the efficacy of each shRNA in eliciting a given on- or off-target effect is modeled by parameters α_i and β_i respectively, which are constrained to be in the unit interval [0,1]. Note that in the D1 model, separate efficacy parameters are estimated for each gene and seed targeted by a given shRNA. We found that these approaches gave very similar results, and using a single gene- and seed-efficacy parameter per shRNA allowed for better computational efficiency.

An important addition in D2 is the introduction of a 'screen signal' parameter q_j for each cell line, which scales how the cell line's gene effects are translated into shRNA-level depletion scores. This allows the model to account for global differences in the gene knockdown effects measured for different cell lines, such as arising from variable RNAi efficacy.

Additionally, overall scale and offset parameters γ_{jk} and a_{jk} capture differences in the distribution of LFC values between screens. Both sets of scale parameters (q_j and γ_{jk}) are constrained so that their average is one, ensuring that they capture relative differences in scale across screens, and that the model is identifiable. Note that we assume that each cell line is characterized by a fixed screen signal parameter q_j across datasets, while the parameters γ_{jk} and a_{jk} vary across different screens from a given cell line in order to capture batch effects.

Systematic shifts in the LFC values for an shRNA that are not captured by the gene and seed effect predictions, are modeled by additive components θ_{ik} and c_i. The former, which is a fixed offset across cell lines for each shRNA *i* in dataset *k*, is designed to capture errors in the initial plasmid DNA measurements (which are shared across samples in the Achilles dataset). The c_i on the other hand, are intended to model off-target effects not captured by the model-predicted seed-effects (shared across datasets using shRNA *i*).

Finally, the ϵ_{ijk} represent noise terms associated with each depletion measurement, which are assumed to be independently, normally distributed with screen-specific noise variance σ^2_{jk}.

Parameter estimation

We use a hybrid approach for parameter estimation, where Bayesian inference is used to estimate posterior distributions for the gene effects ($\overline{g}_l$, g_{lj}), seed effects ($\overline{b}_s$, b_{sj}), and intercept terms (θ_{ik}, c_i, and a_{jk}), while point estimates (maximum a posteriori, MAP) are used for the remaining model parameters (the scale terms (q_j and γ_{jk}), shRNA efficacies (α_i and β_i), and noise variances (σ^2_{jk})). Initial point estimates of all parameters are constructed using an alternating block-wise coordinate ascent approach, after which a final stage of variational Bayesian inference is used to approximate the posterior distribution over $\Theta = \left\{\overline{g}_l, g_{lj}, \overline{b}_s, b_{sj}, \theta_{ik}, c_i, a_{jk}\right\}$. The full procedure is outlined below:

1) Initialize shRNA efficacy terms: α_i, β_i, along with screen signal (q_j) and noise variance (σ^2_{jk}) parameters to 1.
2) Initialize screen-specific scale terms γ_{jk} by regressing the LFC data for each cell-line/batch on the average LFC across cell lines for that batch: $D_{ijk} = \hat{\gamma}_{jk}\overline{D}_{ik} + c$.
3) Estimate Θ given current estimates of remaining parameters.
4) Estimate shRNA efficacies (α_i, β_i) given current estimates of remaining parameters.
5) Estimate screen signal parameters q_j given current estimates of remaining parameters.
6) Estimate overall scale parameters γ_{jk} given current estimates of remaining parameters.
7) Repeat steps [3-6] until convergence of the log-posterior.
8) Initialize noise variance parameters σ^2_{jk} by estimating the average residual variance of the model for each cell line/batch.
9) Apply variational inference to estimate the posterior distribution of Θ, along with the noise variances σ^2_{jk}, given point estimates of other parameters.

In steps 3, 4, and 6 we use SciPy's L-BFGS-B numerical optimization routine (Byrd et al. 1995) to maximize the conditional posterior with respect to each parameter set. When estimating the shRNA efficacies we use bound

constraints to ensure they are restricted to the interval [0,1]. To fit the overall scale parameters γ_{jk}, we maximized the posterior with the cell-line specific gene and seed effects (g_{lj}, b_{sj}) set to zero. This ensures that overall scale differences between samples were absorbed by γ_{jk}, rather than being incorporated in the estimates of g_{lj} and b_{sj}.

The screen signal terms q_j are updated by estimating the relative differences in measured gene effects for predefined positive and negative control gene sets. In particular, $\hat{q}_j$ are given by the difference between median positive control and negative control gene effects:

$$\hat{q}_j = \left(\underset{l \in L_{neg}}{\text{median}}(\overline{g}_l + g_{lj}) - \underset{l \in L_{pos}}{\text{median}}(\overline{g}_l + g_{lj}) \right)$$

We then normalize the q_j to have an average value of 1 across cell lines. For the positive and negative control sets, we used the curated sets created by Hart et al. (Hart et al. 2015). We found largely similar results when updating the q_j by maximizing the conditional posterior, as with the γ_{jk} (and we provide this MAP estimation as an option in the open source version of the code). However, estimates obtained using positive/negative control gene separation provided more robust correction of systematic differences in screen quality between cell lines, particularly with the Achilles dataset. Note that while we use predefined sets of positive and negative control genes as part of the parameter estimation procedure, this does not create biases in the gene effect estimates for these genes. Rather, these gene sets are only used for estimating a global scaling of each cell line's gene effects relative to other cell lines, and hence do not affect, for instance, the rank order of gene effects for a given cell line. Furthermore, by using the medians across large gene sets (217/926 positive/negative control genes respectively), the estimates of q_j are insensitive to inclusion of individual genes. Nevertheless, we performed cross-validation experiments where we split the positive and negative control gene sets into separate 'train' and 'test' sets to verify that this procedure does not introduce bias in our downstream model performance evaluation (**Fig. S11**).

For the final stage of model-fitting, we used a variational approximation to estimate the posterior distribution $p(\Theta \mid \mathbf{D}; \widehat{\Psi})$, where $\Theta = \left\{ \overline{g}_l, g_{lj}, \overline{b}_s, b_{sj}, \theta_{ik}, c_i, a_{jk} \right\}$ is the set of parameters for which we estimate the posterior, $\mathbf{D}$ is the observed LFC data, and $\widehat{\Psi}$ is the fixed vector of point estimates (MAP) for the remaining model parameters. We use a fully-factorized Gaussian (mean-field) model $q(\Theta; \lambda)$ to approximate the posterior, which is parameterized by λ: the set of marginal means and variances for each parameter in Θ. The λ, along with the noise variances σ^2_{jk} for each cell line/batch are then estimated by minimizing the KL-divergence $\mathrm{KL}\left(q(\Theta; \lambda) \mid p(\Theta \mid \mathbf{D}; \sigma^2_{jk}, \hat{\Psi}) \right)$. To accomplish this, we utilized Edward (Tran et al. 2017), a probabilistic modeling language built on top of TensorFlow. In particular, we used the Edward function 'KLqp', which uses stochastic variational expectation-maximization to simultaneously optimize λ and σ^2_{jk}, by alternating between minimizing $\mathrm{KL}(q \mid p)$ with given σ^2_{jk} and maximizing $\mathbb{E}_{q(\Theta;\lambda)} \left[p(\Theta, \mathbf{D}; \sigma^2_{jk} \mid \Psi) \right]$ with respect to σ^2_{jk}.

Priors and hyperparameter selection

We use zero-mean Gaussian priors for the set of parameters $\Theta = \left\{ \overline{g}_l, g_{lj}, \overline{b}_s, b_{sj}, \theta_{ik}, c_i, a_{jk} \right\}$ for which we estimate approximate posteriors. For the remaining parameters we assume uniform priors. The model thus uses hyperparameters that specify the prior variance associated with each parameter in Θ: $\sigma^2_{\overline{g}}$, σ^2_{g}, $\sigma^2_{\overline{b}}$, σ^2_{b}, σ^2_{θ}, σ^2_{c}

, and σ_a^2. In general, the results of the model were largely robust towards the precise choices of these hyperparameters. The most important hyperparameter is σ_g^2, which (along with $\sigma_{\bar{g}}^2$) controls the proportion of variance in the estimated gene effects attributed to between-gene vs. within-gene differences. Even changes in σ_g^2 however, left the relative patterns across cell lines for each gene, as well as the across-cell-line averages gene effects, largely unchanged, and hence have minimal effect on most downstream analyses.

To select values for the prior variances $\sigma_{\bar{g}}^2$, $\sigma_{\bar{b}}^2$, σ_c^2, and σ_θ^2 we performed a coarse grid search, choosing the values that produced average gene dependency estimates $\bar{g}_l$ with maximal separation between positive and negative control gene sets. We then performed a second grid search over values of σ_g^2 and σ_b^2, using a variety of metrics to select values providing the best results, including out-of-sample prediction accuracy, agreement with CRISPR data, and correlation with expected biomarkers. For a_{jk} we used an uninformative prior, setting σ_a^2 to an arbitrary large value.

Code availability

The full source code implementing the model, data preprocessing, and figure generation is made available at https://github.com/cancerdatasci/demeter2.

Other modeling details

Data preprocessing

As in the original DEMETER model, we exclude data for shRNAs that target more than 10 genes from analysis, as such 'promiscuous' shRNAs are likely to provide unreliable data. We also identified groups of genes that were targeted by identical sets of shRNAs. Since the models cannot distinguish the effects of knocking down individual genes within such groups, we combined them into single entities ('gene families') when estimating either DEMETER or DEMETER2 models (Tsherniak et al. 2017). For GA, RSA, ATARiS, and DEMETER, LFC values were z-score normalized per cell line and batch before model fitting. For the Achilles data, there were three different batches of cell lines screened, reflecting changes in the library and experimental methods (Tsherniak et al. 2017). For the DRIVE data, each cell line was screened using three shRNA libraries (McDonald et al. 2017), creating three batches of shRNA data per cell line.

DEMETER

The DEMETER model was fit using the R source code provided at: https://github.com/cancerdatasci/demeter. Achilles LFC data were provided as input in three separate batches of cell lines (Tsherniak et al. 2017). For the DRIVE data, LFC values from different 'pools' were all combined into a single matrix as input. shRNAs from different pools were considered distinct, even if they shared the same targeting sequence. The following parameters were determined by performing separate hyperparameter searches on the DRIVE and Achilles data.

	Achilles	DRIVE
randseed	1	1
G.S	1.67e-5	2.38e-4
alpha.beta.gamma	0.583	0.033
max.num.iter	500	500
learning.rate	0.005	5e-5

When applying DEMETER to the Marcotte et al. dataset we used the same hyperparameters as used for analyzing the Achilles dataset.

Note that while DEMETER was previously applied to quantile-normalized LFC data (Tsherniak et al. 2017), here we used z-score normalization to make the results more directly comparable to those produced by DEMETER2.

ATARiS

ATARiS was run using Gene Pattern (http://software.broadinstitute.org/cancer/software/genepattern/modules/docs/ATARiS/1) with the default parameters. The first solution for each gene was taken. The ATARiS algorithm was run separately on the '98k' and '55k' Achilles data. Note that we combined the two '55k' Achilles batches for running ATARiS, because they used nearly identical shRNA libraries, and one of the batches had too few cell lines to get reliable results.

RSA

RSA was run using the R implementation provided by the Genomics Institute of the Novartis Research Foundation (http://winzeler.ucsd.edu/supplemental/KonigNatureMethod-2007/RSA.html) with the following parameters:

- No Bonferroni correction
- Not reversed
- Lower bound: -1000
- Upper bound: 1000

The bounds were set to extreme values so that no gene would automatically be considered a hit, and genes targeted by only one hairpin were removed. Data were combined across batches for each cell line (after z-score normalization). Separate input files were created for each cell line and fed into the RSA algorithm.

Processing gene dependency estimates

Gene dependency scores for RSA, GA, and D2 were normalized using a uniform scaling and offset (applied to all cell lines) so that the median of the across-cell-line average dependency scores for positive and negative control gene sets (Hart et al. 2015) were set to -1 and 0 respectively. Such normalization could not be applied with ATARiS and D1, which

estimate gene dependencies on a relative scale. Hence, for these models, we applied a global z-score normalization of the dependency scores (mean-subtracting per gene, then normalizing by the global standard deviation).

For some genes the estimated dependency scores were deemed unreliable, and were excluded from all analysis. In particular, any genes that were targeted by fewer than three shRNAs were excluded. We also excluded genes that were determined (by the DEMETER2 model) to have poor quality reagents. Specifically, we removed genes where the average gene-knockdown efficacy (α_i) was less than a minimum value (0.2), or where the sum of α_i across targeting shRNAs was less than a threshold of 1.5. These criteria resulted in the removal of 571/8393 genes for the DRIVE dataset and 358/16860 genes for the Achilles dataset. Finally, genes that were part of a 'gene-family', sharing identical sets of targeting shRNAs, were excluded from the analyses, since it is not possible to distinguish the specific gene knockdown effects among genes within such a group (Tsherniak et al. 2017). For comparisons across models, we analyze only those genes for which we obtained valid gene effect estimates with D2, based on the above criteria, to ensure fair comparisons.

Genomic features

Gene expression data were taken from the the file: CCLE_DepMap_18Q1_RNAseq_RPKM_20180214.gct, downloaded from the Cancer Cell Line Encyclopedia (CCLE) portal (https://portals.broadinstitute.org/ccle/data). These RPKM values were then transformed according to log10(RPKM + 0.001), and our analysis was restricted to protein coding genes only. For identifying genes that were 'unexpressed' in a given cell line, we utilized a log10(RPKM) threshold of -1.

Gene-level relative copy number data were derived from a combination of CCLE whole-exome sequencing (WES) and SNP data (https://portals.broadinstitute.org/ccle/data). We utilized WES and SNP data to achieve maximal coverage across cell lines. When multiple datasets were available for a given cell line, we prioritized WES over SNP data. Relative copy number data were also log-transformed for analysis.

For mutation data, we utilized the merged mutation calls file from the CCLE portal (CCLE_DepMap_18Q1_maf_20180207.txt), which combines information from multiple data sources and types. We considered mutations to be 'damaging' if they were marked as 'deleterious' in the maf file. A subset of missense mutations was further categorized as 'hotspot' mutations, if they were annotated as being either TCGA hotspots or COSMIC hotspots.

Dependency-feature relationship analysis

We benchmark the ability of different models to identify relationships between genetic dependencies and genomic features in several ways. First, we used CRISPR-Cas9 gene dependency estimates, based on the CERES algorithm (Meyers et al. 2017), to identify a set of top dependency-feature relationships which we could then test in the RNAi datasets using different models. In particular, for each CRISPR-Cas9 gene dependency profile we identified the most strongly correlated feature (based on the Pearson correlation) for each of four feature types (mRNA expression, relative copy number, damaging mutation, and hotspot missense mutation). We then took the top feature-dependency

correlations for each feature type (up to 200 pairs per feature type), after removing any relationships that did not have a minimum correlation magnitude of 0.4, producing a set of 417 benchmark feature-dependency relationships (193 copy number, 200 gene expression, 11 damaging mutation, 13 hotspot missense mutation).

We also employed a list of known gene-gene relationships to test how frequently the genomic feature most correlated with a gene's dependency was from a gene known *a-priori* to have some relationship with the targeted gene. We used gene-gene relationships defined in several ways:

- Physical interactions: gene pairs that were identified as CORUM protein complex co-members (Ruepp et al. 2010), or as physically interacting using protein-protein interaction data from InWeb (Li et al. 2017).
- Paralogs were defined as gene pairs which last underwent a duplication event rather than a speciation event according to Ensembl.

For the analysis in **Fig. 5e**, dependency-feature relationships were classified into the following groups:

- **CYCLOPS:** defined as cases where the top correlated genomic feature was either mRNA expression or copy number of the target gene, and the correlation was positive.
- **Oncogene mutation**: defined as cases where the top correlated genomic feature was the gene's own hotspot missense mutation status, and the correlation was negative (stronger dependency in the mutant cell lines).
- **Oncogene expression**: defined as cases where the top correlated feature was the gene's own mRNA expression, and the sign of the correlation was negative.
- **Paralog deficiency**: defined as cases where the top correlated feature was either damaging mutation, copy number, or gene expression of a sequence paralog to the target gene, and the correlation was positive.
- **Interacting protein**: defined as cases where the top correlated feature was from any gene known to physically interact with the target gene (using either positive or negative correlations).

Dependency-feature relationships that fit multiple of the above categories were prioritized in the order described above (e.g. relationships would only be classified as 'physical interactors' if they did not meet the criteria for any of the other categories).

Additional analysis details

Wherever possible we utilized weighted statistics (including Pearson correlations, means and variances) to evaluate the quality of D2 gene dependency estimates, where dependency scores were linearly weighted by their associated precision (the inverse posterior variance). Weighted Pearson correlations, and associated p-values, were computed using the R package "weights" (https://CRAN.R-project.org/package=weights).

For PCA analysis (**Fig. 5b**) we used probabilistic PCA, implemented in the R package pcaMethods (Stacklies et al. 2007), which naturally handles missing values in the gene dependency matrices. PCA was applied to the matrices of dependency estimates from each model after mean-subtracting per gene.

Figures showing conditional average correlations were created using the 'geom_smooth' function from the R package ggplot2 (Wikham 2009), which employs GAM models with cubic spline basis functions to estimate conditional means and standard errors.

Network analyses (**Fig. 3c-d**) were generated using the R package igraph (Csardi and Nepusz 2006). We first identified the top 25 gene dependency profiles most strongly correlated (magnitude of Pearson correlation) with the query gene's dependency profile, using these genes as the nodes of the graph. Edge weights between nodes were given by the magnitude of Pearson correlations between gene pairs, using only edges where the correlation magnitude was > 4 z-score above the mean (across all gene pairs for each dataset). Disconnected nodes were then trimmed from the graph before generating the plots using the 'layout_nicely' algorithm in igraph.

To generate figures 3a,b, as well as figure S6 we computed pairwise dependency correlations using a subset of genes for computational efficiency. Specifically, we used the top 2500 genes with highest across-cell-line variance (according to DEMETER), considering only genes present in both Achilles and DRIVE.

Paired two-sample comparisons were made using Wilcoxon signed rank tests. Significance of Pearson correlations was assessed using the R functions *cor.test* (with the function *wtd.cor* from the R package weights used for precision-weighted correlations). For computing dependency-feature association p-values (**Fig. 5c**) we used the Pearson correlation p-value for continuous features (gene expression, copy number), and p-values from two-sample t-tests for binary features (mutations). Two-sided p-values were used in all cases.

Data availability

All datasets used to generate the results presented here are publicly available, as described here or above. The results of the DEMETER2 model applied to the DRIVE and Achilles datasets, as well as to the combined DRIVE, Achilles and Marcotte et al. data, are available at https://depmap.org/R2-D2, and in a figshare record at: https://doi.org/10.6084/m9.figshare.6025238.v1 (Cancer Data Science, 2018a). CRISPR-Cas9 essentiality screening data processed using the CERES algorithm can be downloaded from https://doi.org/10.6084/m9.figshare.5863776.v1 (Cancer Data Science, 2018b). Gene-level copy number information, as described above, can also be downloaded from the DEMETER2 figshare record.

Acknowledgements:

This work was funded in part by the Carlos Slim Foundation. We thank Andrew Tang for help with figure design.

References

Birmingham, A., Anderson, E.M., Reynolds, A., Ilsley-Tyree, D., Leake, D., Fedorov, Y., Baskerville, S., Maksimova, E., Robinson, K., Karpilow, J., Marshall, W.S. and Khvorova, A. 2006. 3' UTR seed matches, but not overall identity, are associated with RNAi off-targets. *Nature Methods* 3(3), pp. 199–204.

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

Byrd, R.H., Lu, P., Nocedal, J. and Zhu, C. 1995. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing* 16(5), pp. 1190–1208.

Cancer Data Science (2018a): DEMETER2 data. figshare. Fileset. https://doi.org/10.6084/m9.figshare.6025238.v1

Cancer Data Science (2018b): Broad Institute Cancer Dependency Map, CRISPR Avana dataset 18Q1 (Avana_public_18Q1). figshare. Fileset. https://doi.org/10.6084/m9.figshare.5863776.v1

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., Jiang, G., Hsiao, J., Mermel, C.H., Getz, G., Barretina, J., Gopal, S., Tamayo, P., Gould, J., Tsherniak, A., Stransky, N., Luo, B., Ren, Y., Drapkin, R., Bhatia, S.N., Mesirov, J.P., Garraway, L.A., Meyerson, M., Lander, E.S., Root, D.E. and Hahn, W.C. 2011. Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Proceedings of the National Academy of Sciences of the United States of America* 108(30), pp. 12372–12377.

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., Lizotte, P.H., Jiang, G., Hsiao, J., Tsherniak, A., Dwinell, E., Aoyama, S., Okamoto, M., Harrington, W., Gelfand, E., Green, T.M., Tomko, M.J., Gopal, S., Wong, T.C., Li, H., Howell, S., Stransky, N., Liefeld, T., Jang, D., Bistline, J., Hill Meyers, B., Armstrong, S.A., Anderson, K.C., Stegmaier, K., Reich, M., Pellman, D., Boehm, J.S., Mesirov, J.P., Golub, T.R., Root, D.E. and Hahn, W.C. 2014. Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. *Scientific data* 1, p. 140035.

Csardi, G. and Nepusz, T. 2006. The igraph software package for complex network research. *InterJournal* Complex Systems, p. 1695.

Grimm, D., Wang, L., Lee, J.S., Schürmann, N., Gu, S., Börner, K., Storm, T.A. and Kay, M.A. 2010. Argonaute proteins are key determinants of RNAi efficacy, toxicity, and persistence in the adult mouse liver. *The Journal of Clinical Investigation* 120(9), pp. 3106–3119.

Hart, T., Brown, K.R., Sircoulomb, F., Rottapel, R. and Moffat, J. 2014. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Molecular Systems Biology* 10, p. 733.

Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K.R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., Mero, P., Dirks, P., Sidhu, S., Roth, F.P., Rissland, O.S., Durocher, D., Angers, S. and Moffat, J. 2015. High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell* 163(6), pp. 1515–1526.

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 (New York)* 12(7), pp. 1179–1187.

König, R., Chiang, C., 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. *Nature Methods* 4(10), pp. 847–849.

Li, T., Wernersson, R., Hansen, R.B., Horn, H., Mercer, J., Slodkowicz, G., Workman, C.T., Rigina, O., Rapacki, K., Stærfeldt, H.H., Brunak, S., Jensen, T.S. and Lage, K. 2017. A scored human protein-protein interaction network to catalyze genomic interpretation. *Nature Methods* 14(1), pp. 61–64.

Luo, B., Cheung, H.W., Subramanian, A., Sharifnia, T., Okamoto, M., Yang, X., Hinkle, G., Boehm, J.S., Beroukhim, R., Weir, B.A., Mermel, C., Barbie, D.A., Awad, T., Zhou, X., Nguyen, T., Piqani, B., Li, C., Golub, T.R., Meyerson, M., Hacohen, N., Hahn, W.C., Lander, E.S., Sabatini, D.M. and Root, D.E. 2008. Highly parallel identification of essential genes in cancer cells. *Proceedings of the National Academy of Sciences of the United States of America* 105(51), pp. 20380–20385.

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., Pe'er, D., Moffat, J. and Neel, B.G. 2016. Functional genomic landscape of human breast cancer drivers, vulnerabilities, and resistance. *Cell* 164(1–2), pp. 293–309.

McDonald, R.E., de Weck, A., Schlabach, M.R., Billy, E., Mavrakis, K.J. and Hoffman, G.R. 2017. DRIVE raw data. *Mendeley Data, v4*. http://dx.doi.org/10.17632/y3ds55n88r.4

McDonald, E.R., de Weck, A., Schlabach, M.R., Billy, E., Sellers, W.R., et al. 2017. Project DRIVE: A Compendium of Cancer Dependencies and Synthetic Lethal Relationships Uncovered by Large-Scale, Deep RNAi Screening. *Cell* 170(3), p. 577–592.e10.

Meyers, R.M., Bryan, J.G., McFarland, J.M., Weir, B.A., Sizemore, A.E., Xu, H., Dharia, N.V., Montgomery, P.G., Cowley, G.S., Pantel, S., Goodale, A., Lee, Y., Ali, L.D., Jiang, G., Lubonja, R., Harrington, W.F., Strickland, M., Wu, T., Hawes, D.C., Zhivich, V.A., Wyatt, M.R., Kalani, Z., Chang, J.J., Okamoto, M., Stegmaier, K., Golub, T.R., Boehm, J.S., Vazquez, F., Root, D.E., Hahn, W.C. and Tsherniak, A. 2017. Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nature Genetics* 49(12), pp. 1779–1784.

Nijhawan, D., Zack, T.I., Ren, Y., Strickland, M.R., Lamothe, R., Schumacher, S.E., Tsherniak, A., Besche, H.C., Rosenbluh, J., Shehata, S., Cowley, G.S., Weir, B.A., Goldberg, A.L., Mesirov, J.P., Root, D.E., Bhatia, S.N., Beroukhim, R. and Hahn, W.C. 2012. Cancer vulnerabilities unveiled by genomic loss. *Cell* 150(4), pp. 842–854.

Rämö, P., Drewek, A., Arrieumerlou, C., Beerenwinkel, N., Ben-Tekaya, H., Cardel, B., Casanova, A., Conde-Alvarez, R., Cossart, P., Csúcs, G., Eicher, S., Emmenlauer, M., Greber, U., Hardt, W.-D.,

Helenius, A., Kasper, C., Kaufmann, A., Kreibich, S., Kühbacher, A., Kunszt, P., Low, S.H., Mercer, J., Mudrak, D., Muntwiler, S., Pelkmans, L., Pizarro-Cerdá, J., Podvinec, M., Pujadas, E., Rinn, B., Rouilly, V., Schmich, F., Siebourg-Polster, J., Snijder, B., Stebler, M., Studer, G., Szczurek, E., Truttmann, M., von Mering, C., Vonderheit, A., Yakimovich, A., Bühlmann, P. and Dehio, C. 2014. Simultaneous analysis of large-scale RNAi screens for pathogen entry. *BMC Genomics* 15, p. 1162.

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 Research* 38(Database issue), pp. D497-501.

Schmich, F., Szczurek, E., Kreibich, S., Dilling, S., Andritschke, D., Casanova, A., Low, S.H., Eicher, S., Muntwiler, S., Emmenlauer, M., Rämö, P., Conde-Alvarez, R., von Mering, C., Hardt, W.-D., Dehio, C. and Beerenwinkel, N. 2015. gespeR: a statistical model for deconvoluting off-target-confounded RNA interference screens. *Genome Biology* 16, p. 220.

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., Margolin, A.A., Root, D.E., Hahn, W.C. and Mesirov, J.P. 2013. ATARiS: computational quantification of gene suppression phenotypes from multisample RNAi screens. *Genome Research* 23(4), pp. 665–678.

Stacklies, W., Redestig, H., Scholz, M., Walther, D. and Selbig, J. 2007. pcaMethods--a bioconductor package providing PCA methods for incomplete data. *Bioinformatics* 23(9), pp. 1164–1167.

Tran, D., Kucukelbir, A., Dieng, A., Rudolph, M., Liang, D. and Blei, D. 2017. Edward: A library for probabilistic modeling, inference, and criticism. *arXiv:1610.09787v3*.

Tsherniak, A., Vazquez, F., Montgomery, P.G., Weir, B.A., Kryukov, G., Cowley, G.S., Gill, S., Harrington, W.F., Pantel, S., Krill-Burger, J.M., Meyers, R.M., Ali, L., Goodale, A., Lee, Y., Jiang, G., Hsiao, J., Gerath, W.F.J., Howell, S., Merkel, E., Ghandi, M., Garraway, L.A., Root, D.E., Golub, T.R., Boehm, J.S. and Hahn, W.C. 2017. Defining a cancer dependency map. *Cell* 170(3), p. 564–576.e16.

Vickers, T.A., Lima, W.F., Nichols, J.G. and Crooke, S.T. 2007. Reduced levels of Ago2 expression result in increased siRNA competition in mammalian cells. *Nucleic Acids Research* 35(19), pp. 6598–6610.

Wang, T., Yu, H., Hughes, N.W., Liu, B., Kendirli, A., Klein, K., Chen, W.W., Lander, E.S. and Sabatini, D.M. 2017. Gene Essentiality Profiling Reveals Gene Networks and Synthetic Lethal Interactions with Oncogenic Ras. *Cell* 168(5), p. 890–903.e15.

Wikham, H. 2009. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.