Evaluation of RNAi and CRISPR technologies by large-scale gene expression profiling in the Connectivity Map

PLOS BIOLOGY

METHODS AND RESOURCES

Evaluation of RNAi and CRISPR technologies by large-scale gene expression profiling in the Connectivity Map

Ian Smith[1‡], Peyton G. Greenside[1‡], Ted Natoli[1], David L. Lahr[1], David Wadden[1], Itay Tirosh[1], Rajiv Narayan[1], David E. Root[1], Todd R. Golub[1,2,3,4], Aravind Subramanian[1]*, John G. Doench[1]*

1 Broad Institute of MIT and Harvard, Cambridge, Massachusetts, United States of America, **2** Harvard Medical School, Boston, Massachusetts, United States of America, **3** Department of Pediatric Oncology, Dana Farber Cancer Institute, Boston, Massachusetts, United States of America, **4** Howard Hughes Medical Institute, Chevy Chase, Maryland, United States of America

‡ These authors share first authorship on this work.
* aravind@broadinstitute.org (AS); jdoench@broadinstitute.org (JGD)

Citation: Smith I, Greenside PG, Natoli T, Lahr DL, Wadden D, Tirosh I, et al. (2017) Evaluation of RNAi and CRISPR technologies by large-scale gene expression profiling in the Connectivity Map. PLoS Biol 15(11): e2003213. https://doi.org/10.1371/journal.pbio.2003213

Academic Editor: Tom Freeman, University of Edinburgh, UNITED KINGDOM OF GREAT BRITAIN AND NORTHERN IRELAND

Received: June 5, 2017

Accepted: November 9, 2017

Published: November 30, 2017

Data Availability Statement: The portal to access the Connectivity Map is https://clue.io. This website also provides detailed information on the experimental protocols used to generate these data. Additional information can be found at the Library of Network-Based Cellular Signatures (LINCS) Program website, http://www.lincsproject.org. All data used in this study can be extracted from the overall CMAP dataset deposited in GEO, accession numbers GSE92742 and GSE70138; alternatively, only the data from this study are

Abstract

The application of RNA interference (RNAi) to mammalian cells has provided the means to perform phenotypic screens to determine the functions of genes. Although RNAi has revolutionized loss-of-function genetic experiments, it has been difficult to systematically assess the prevalence and consequences of off-target effects. The Connectivity Map (CMAP) represents an unprecedented resource to study the gene expression consequences of expressing short hairpin RNAs (shRNAs). Analysis of signatures for over 13,000 shRNAs applied in 9 cell lines revealed that microRNA (miRNA)-like off-target effects of RNAi are far stronger and more pervasive than generally appreciated. We show that mitigating off-target effects is feasible in these datasets via computational methodologies to produce a consensus gene signature (CGS). In addition, we compared RNAi technology to clustered regularly interspaced short palindromic repeat (CRISPR)-based knockout by analysis of 373 single guide RNAs (sgRNAs) in 6 cells lines and show that the on-target efficacies are comparable, but CRISPR technology is far less susceptible to systematic off-target effects. These results will help guide the proper use and analysis of loss-of-function reagents for the determination of gene function.

Author summary

The Connectivity Map (CMAP) is a large-scale resource of the gene expression changes caused by small-molecule and genetic perturbations in multiple cell types. Here, we analyze 2 types of genetic perturbations, RNAi and CRISPR, using CMAP data. RNAi and CRISPR are both used to generate loss-of-function mutants but exploit different cellular pathways. We find that the on-target efficacy of the 2 technologies is similar. However, the off-target effects of RNAi are much greater than typically appreciated, whereas CRISPR technology has negligible off-target activity. To enhance the use of CMAP data generated

collected in a separate deposit, GSE106127. The outputs of specific analyses are provides as Supplementary Data. All code necessary to replicate these analyses is available on GitHub: https://github.com/iamsinht/rnaicrispr/.

Funding: NIH (grant number 5U54HG006093). Received by AS and TRG. NIH (grant number 5U01HG008699). Received by AS and TRG. NIH (grant number U54HG008699). Received by AS and TRG. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: We have read the journal's policy and the authors of this manuscript have the following competing interests: A.S. is a shareholder of Genometry, Inc.

Abbreviations: Cas9, CRISPR-associated 9; CGS, consensus gene signature; CMAP, Connectivity Map; CRISPR, clustered regularly interspaced short palindromic repeat; CSS, consensus seed signature; dsDNA, double-strand DNA; FDR, false discovery rate; GWAS, genome wide association studies; LINCS, Library of Integrated Network-Based Cellular Signatures; miRNA, microRNA; NHEJ, nonhomologous end-joining pathway; nt, nucleotide; PC1, first principal component; RNAi, RNA interference; sgRNA, single guide RNA; shRNA, short hairpin RNA; siRNA, small interfering RNA.

with genetic reagents, we develop the consensus gene signature (CGS) and demonstrate that it can improve the identification of on-target activity, which will be particularly critical for analyzing data generated with RNAi. Overall, the large-scale data generated as part of CMAP allow us to robustly characterize the on- and off-target properties of 2 common approaches for genetic manipulation of cells.

Introduction

The Connectivity Map (CMAP), a component of the Library of Integrated Network-Based Cellular Signatures (LINCS), aims to build a comprehensive lookup table of the mRNA expression consequences of perturbing a cell. Conceptually, the pattern of mRNA changes serves as a signature of the perturbation, and correlations between these signatures allows insight into connections between genes, drugs, and disease states [1–5]. The original CMAP dataset (build 01) examined 164 small molecules, and build 02 increased that number to over 6,000. Even with this limited-scale dataset, the query of CMAP has provided biological insight across diverse disease areas [6–9]. Genetic perturbations are an attractive companion to small molecules, allowing the direct interrogation of gene function in order to understand how gene dysfunction leads to disease states. Here, we provide the first systematic exploration of the use of both RNA interference (RNAi) and clustered regularly interspaced short palindromic repeat (CRISPR) loss-of-function technologies in the next build CMAP [5].

Soon after its discovery as an endogenous biological process, RNAi became the leading technology for the disruption of any gene of interest, especially in mammalian systems that were previously refractory to genetic manipulation [10]. The combination of scalable reagent creation, facile cellular delivery, and potent gene knockdown has enabled genome-wide RNAi screens, leading to a wealth of information on gene function [11–13]. More recently, components derived from CRISPR loci in prokaryotes have been developed as an orthogonal means of perturbing gene function in mammalian cells [14–16]. In this system, the CRISPR-associated 9 (Cas9) nuclease is programmed with a single guide RNA (sgRNA) to create a targeted double-strand DNA (dsDNA) break [17]. When this break occurs in protein coding regions, indels arising from the error-prone nonhomologous end-joining pathway (NHEJ) can result in frameshift mutations and a null allele. Like RNAi, CRISPR technology can be used at genome scale for phenotypic screens [18–21].

The specificities of RNAi and CRISPR technologies are a critical concern for the reliable interpretation of experimental results. That molecular triggers of the RNAi pathway, while designed to silence a single gene of interest, can in fact affect multiple off-target genes has long been documented [22–24]. Indeed, in some cases, independent RNAi screens ostensibly examining the same phenotype have reported widely different hit lists [25]. And while some of this difference may be due to variations in experimental conditions, it likely also reflects, in part, the contamination of hit lists by genes erroneously nominated due to off-target effects. In mammalian cells, in which the trigger is usually a synthetic small interfering RNA (siRNA) or a transcribed short hairpin RNA (shRNA), there are 2 major modes of off-target activity. First, the RNA sequence intended to be the passenger strand can instead be selected by AGO2 as the targeting strand; in this case, all activity from this unintended loading of AGO2 will be off target [26]. Fortunately, the rules determining strand selection have been well studied and proper design can minimize this mode of off-target activity [27,28]. The second major mode of off-target activity, however, is not easily dealt with during the design of RNAi triggers: the small RNA intended to target a unique transcript can instead enter the miRNA pathway and

potentially contribute to the silencing of dozens if not hundreds of transcripts [22,23]. Indeed, for endogenous miRNAs, both computational algorithms to predict mRNA targets and experimental approaches to detect them en masse have shown that many transcripts are subject to miRNA-based regulation [29,30]. While the mRNA targets of a miRNA are still not fully predictable, it is clear that the miRNA seed sequence—nucleotides (nts) 2 through 7 or 8, counting from the 5′ end—is a major determinant of activity.

Studies of CRISPR technology have shown that sgRNAs intended to target one specific genetic locus can lead to detectable cleavage of off-target DNA sites, although the reported promiscuity of Cas9 from *Streptococcus pyogenes* has varied widely [31,32]. Importantly, isolation of clones arising from CRISPR-mediated gene editing followed by whole-genome sequencing has shown that off-target rates can be reduced below the limit of detection, although this gold-standard assay was performed with only a small number of sgRNAs and is not practical for routine testing [33]. Nevertheless, these results offer the exciting possibility that, as understanding of the on- and off-target parameters of sgRNA design increases, CRISPR technology may be deployed with negligible off-target effects [34,35].

To build the CMAP dataset, 978 transcripts designated as landmarks are detected using Luminex bead-based technology, allowing high-throughput, low-cost data collection across many perturbations and cell types [36]. Here, we have analyzed the gene expression consequences of about 13,000 shRNAs profiled across 9 cell lines, allowing for a rigorous and deep exploration of the biological effects of RNAi reagents. Additionally, we generated gene expression signatures of CRISPR-Cas9 knockout with 373 sgRNAs in 6 cell lines, allowing a direct comparison of the relative potency and specificity of these loss-of-function technologies. Finally, we compared the gene-level results obtained for CRISPR and RNAi to each other. These results will guide the proper use and analysis of genetic perturbations and demonstrate the value of systematic, large-scale lookup tables, such as CMAP.

Results

Widespread off-target effects with RNAi

RNAi reagents that do not have a sequence-matched target, such as scrambled sequences or those designed against reporter genes, are commonly used as experimental controls. By definition, on-target effects of a perturbation are the reproducible gene expression changes relating to the inhibition of the intended target gene, while off-target effects are all other reproducible changes specific to a reagent. From the CMAP dataset, we examined the correlations among signatures of 48 shRNAs targeting nonexpressed control genes and observed that biological replicates of the same shRNA correlate while shRNAs of different sequences do not correlate (Fig 1A, S1A and S1B Fig). This correlation indicates reproducible gene expression changes that are specific to the shRNA perturbation despite the absence of any on-target activity, which implies the presence of a sequence-specific off-target signal.

Molecular triggers of the RNAi pathway are known to enter the miRNA pathway, in which a seed sequence of 6–7 nts directs the repression of off-target transcripts [22,23]. If shRNAs were entering the miRNA pathway to a substantial degree, then we would expect the gene expression profiles of shRNAs sharing a seed sequence to correlate to each other. We examined all 7-nt windows along the shRNA and saw that shRNAs sharing a heptamer starting at positions 11 or 12 of the sense strand correlated strongly (S1C Fig, S1 Data). These positions correspond to nts 2–8 of the antisense strand of the most abundant Dicer-processing products of this shRNA design (S1D Fig) [37], suggesting that most shRNAs do indeed enter the miRNA pathway and produce reproducible gene expression consequences, as has been observed previously [22,23,38,39].

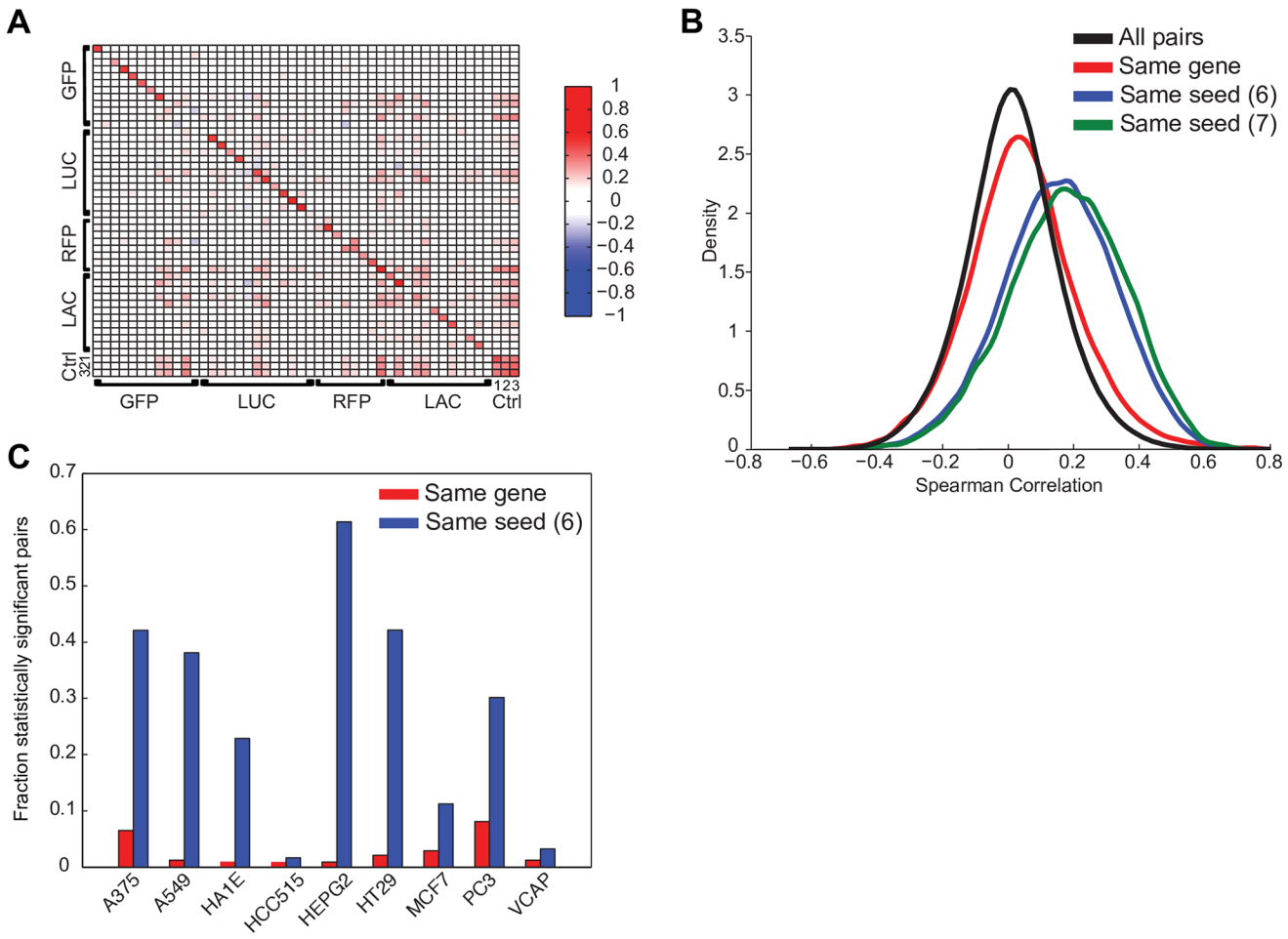

Fig 1. RNAi reagents have widespread off-target effects. (A) Heat map of Spearman correlations among pairs of shRNAs targeting control genes. Correlation on the diagonal reveals a gene expression signal that is reproducible and specific to each shRNA, despite the absence of a target. Control genes are labeled as follows: GFP, LUC, RFP, and LAC. Additional control treatments are grouped under Ctrl; 1: pgw, a lentivirus with no U6 promoter and no shRNA; 2: empty_vector, a lentivirus with a run of 5 thymidines immediately after the U6 promoter, to terminate transcription; 3: UnTrt, wells that did not receive any lentivirus. (B) Distribution of pairwise correlations of shRNA signatures with the same gene target, the same 6- and 7-mer seed sequence, and all pairs of shRNAs. Data shown are from HT29 cells. Pairs of shRNAs with the same seed correlate much better than those with the same gene, which correlate only marginally better than random pairs. (C) The fraction of pairs of shRNA signatures with the same target gene (red) or the same 6-mer seed (blue) that are statistically significant ($q < 0.25$) in each cell line. In all cell lines, correlation due to seed is more often significant than correlation due to gene. See S2 Data. Ctrl, control; GFP, green fluorescent protein; LAC, beta-galactosidase; LUC, firefly luciferase; pgw, puromycin-GFP-WPRE; RFP, red fluorescence protein; RNAi, RNA interference; shRNA, short hairpin RNA; U6, human U6 polymerase III promoter; UnTrt, untreated.

https://doi.org/10.1371/journal.pbio.2003213.g001

We next compared the consistency of gene expression changes elicited by 2 classifications of shRNAs: (i) shRNAs with different seed sequences that target the same gene and (ii) shRNAs designed to target different genes but sharing the same seed sequence. In this set of 18,263 shRNAs, fewer than 1% share both a target gene and a seed sequence. Compared to the null distribution of all pairwise correlations, shRNAs targeting the same gene are more correlated to each other, but the increase in correlation is small compared to that of shRNAs sharing the same seed sequence (Fig 1B). Interestingly, these results were more pronounced in some cell lines, although, in all cases, the correlation between same-seed pairs was greater than same-gene pairs (Fig 1C). These results prompted an examination of seed effects in large-scale viability screens, and a strikingly similar result was observed [40]. This finding means that a

larger component of gene expression changes due to shRNA treatment is a consequence of the seed sequence rather than the target gene knockdown. Furthermore, this suggests that the reliance on an individual shRNA for assessing the phenotypic consequences of gene silencing will often lead to erroneous conclusions, as seed sequence effects are prevalent.

Consensus gene signatures

To attempt to accurately measure the on-target signature, we sought to combine gene expression information from multiple shRNAs targeting the same gene—but with different seeds—to produce a consensus gene signature (CGS). As a first approach, we used a weighted average of multiple perturbations with weights based on a pairwise correlation matrix (Fig 2A). To evaluate the fidelity of a CGS, we introduce the holdout method: divide at least 6 shRNA signatures targeting the same gene into 2 disjoint groups, make a CGS from each, and find the correlation of the 2 CGSs (S2A Fig). For genes with enough signatures to generate 2 disjoint consensus signatures ($\geq$6), holdout analysis checks whether the 2 solutions agree better than nulls generated by permutation with cross-gene shRNAs. If on-target activity is a sufficiently large component of each CGS, then they will correlate with each other better than with a null population of CGSs made from random draws of shRNAs.

Interestingly, examination of the permutation null CGSs revealed that increasing the number of shRNAs led to an increase in the mean and variance of correlation, that is, a generic increase in correlation implying a common, gene-target-independent effect (Fig 2B). We then performed principal component analysis for all classes of perturbations in CMAP assayed by L1000, as well as for small molecule compounds that had previously been assayed by Affymetrix microarray in the previous build of CMAP [1]. We observed that the fraction of variance explained by the first principal component (PC1) was noticeably larger for shRNA CGSs than for individual perturbations (Fig 2C), as may have been anticipated by the generic increase in correlation seen with randomly chosen shRNAs (Fig 2B). Furthermore, although the magnitude of PC1 between both Affymetrix and L1000 is similar, they are distinct, as PC1 is well correlated across 3 different types of perturbations assayed by L1000 and much less correlated to PC1 from Affymetrix (Fig 2D).

Supervised and unsupervised analysis of principal components has been used previously in gene expression analysis, surrogate variable analysis, and population genetics to characterize and remove unwanted variation [41,42]. Thus, we sought to test the effect of removing PC1 when combining multiple shRNA signatures and performing holdout analysis. Across a range of FDR thresholds, we observed consistently improved results when removing PC1 (Fig 2E), and this performance improvement is largely driven by the decrease in variance of the null distribution when PC1 is removed (S2B and S2C Fig). Holdout analysis further suggests that cell context is pertinent for observing gene function, as most genes have a statistically significant holdout result in a fraction of cell lines, but not all of them (S2D Fig, S4 Data). Notably, removal of PC1 did not lead to a decrease in the seed effect, which is still much larger than the gene effect (Fig 2F, compare to Fig 1B). We next compared the results of retaining or removing PC1 when performing queries, using compounds profiled by both Affymetrix and L1000. We observed no discernable difference in recall, that is, the ability of an Affymetrix profile to connect to its L1000 counterpart (Fig 2G). Thus, PC1 may not have a large impact on query results between individual perturbations in CMAP or from external signatures that query the CMAP dataset, but removing PC1 will be beneficial when combining multiple signatures to create new signatures within CMAP, such as the CGS.

To evaluate the effectiveness of the CGS for extracting on-target effects, we first examined the suppression of the target genes when they were directly measured in the expression

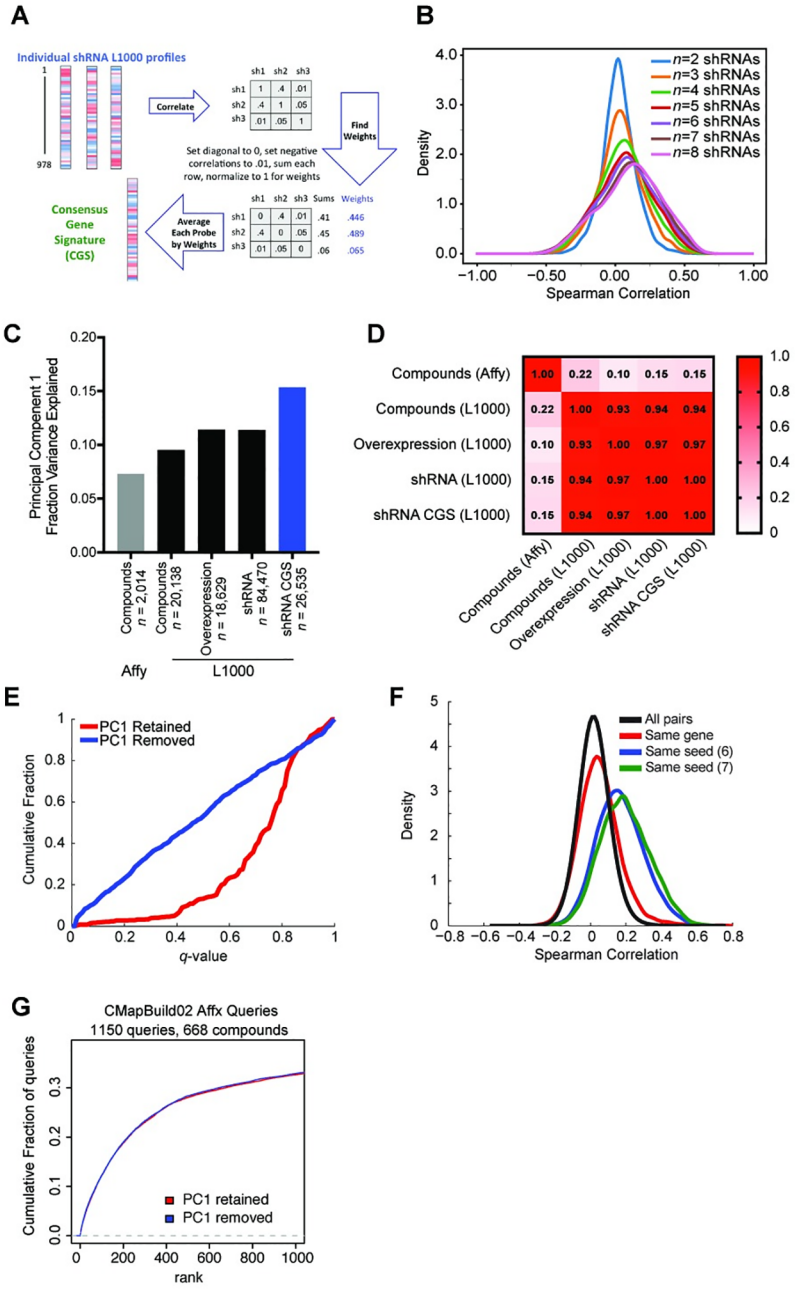

Fig 2. CGSs and investigation of PC1. (A) Schematic of the weighted average procedure for combining individual shRNA signatures targeting the same gene into a CGS. The shRNAs are weighted by the sum of their correlations to other same-gene shRNAs and then averaged. (B) CGSs made from random groups of shRNAs show increasing variance of Spearman correlation with larger numbers of component shRNAs. Because these are random groups, there should not be a consistent signal; the increasing probability of very large correlations reveals a spurious signal that we attribute to the PC1 of the data. (C) Comparison of the fraction of variance explained by PC1 for either CMAP build 02, which used Affymetrix arrays to profile small molecules [1], or the expansion of CMAP, which uses L1000 technology [5] with different types of perturbation. Level 5 data were used. The shRNA CGS has a notably larger PC1. See S3 Data. (D) Pearson correlation of PC1 across RNA measurement platforms and perturbation types in level 5 data. (E) For genes with 6 or more shRNAs, a fraction of statistically significant holdout results at different *q*-value-corrected false discovery rate thresholds, comparing PC1 retained or PC1 removed. Analysis was performed separately for each cell line and data for all cell lines are shown as a single distribution. Because holdout analysis combines multiple shRNA signatures, removal of PC1 decreases the background caused by the general increase in correlations shown in panel (B) and thus improves the performance of this particular analysis. (F) Removal of PC1 does not

diminish the magnitude of the seed effect. After removal of PC1, distribution of pairwise Spearman correlations in HT29 (as a representative cell line) for pairs of shRNAs with the same gene target, the same 6- and 7-mer seed sequence, and all pairs of shRNAs. Compare to Fig 1C. (G) Effect of PC1 of CMAP queries. For small molecules previously profiled in CMAP build 02 by Affymetrix technology, the rank of the matched compound when queried against small molecule L1000 data, with either PC1 retained or removed. CGS, consensus gene signature; CMAP, Connectivity Map; PC1, first principal component; shRNA, short hairpin RNA.

https://doi.org/10.1371/journal.pbio.2003213.g002

signatures, i.e., targeting of landmark transcripts. We saw that 60.6% of shRNA target genes rank in the top 1% of most-down-regulated transcripts when using the CGS, compared to 36.0% when using data from individual shRNAs ($p < 10^{-10}$ by Kolmogorov-Smirnov test, Fig 3A). When we removed PC1, the result did not change appreciably: 61.6% and 36.4% of target transcripts ranked in the top 1% with the CGS and individual shRNA, respectively. It is unsurprising that PC1 does not substantially affect the rank of the target knockdown gene because it is a broad signal with changes across many measured genes, whereas the knockdown of the target gene is large relative to that gene's background distribution. Thus, in the special case in which we know that transcript modulation is on-target, the CGS improves signal. We next evaluated the effectiveness of the CGS for the entire signature: using a leave-one-out approach, we found that 94% of individual shRNAs correlate better to the CGS made from remaining shRNAs than they do to other individual shRNAs targeting the same gene (Fig 3B). By comparison, 77% of shRNAs correlate more strongly with an individual shRNA sharing the same seed than to the CGS containing that shRNA, indicating that use of the CGS partially mitigates the seed-based off-target signal of shRNA signatures (Fig 3B).

To further examine individual shRNAs, we used the leave-one-out approach to create a CGS from all remaining shRNAs targeting the same gene and an analogous consensus seed signature (CSS) from shRNAs with the same seed sequence but different gene targets. We observed that the majority of shRNAs exhibit both on- and off-target activity—quantified by the Spearman correlation of the shRNA signature to the CGS and CSS, respectively (Fig 3C). Interestingly, in any particular cell line, a minority of shRNAs exhibited substantial on-target activity with minimal off-target activity. While the activities of these shRNAs showed variability in other cell lines, the performance across cell lines was correlated (Fig 3D, mean odds ratio of the pairwise contingency tables of 5.6 and $p < 10^{-10}$ by Fisher's exact test). Thus, while shRNAs that are predominantly on target in 1 cell line are more likely to be predominantly on target in another, individual shRNAs that provide an on-target phenotype in 1 cell context must be reconfirmed in other cellular contexts. The use of explicit seed-matched controls has proven useful to experimentally assess the seed contribution to a phenotype [43].

Finally, we asked whether the magnitude of gene expression changes due to a particular seed sequence was consistent across cell lines. For each seed sequence, we calculated the effect magnitude by computing the mean leave-one-out CSS correlation across all shRNAs with that seed for each cell line. We then compared the set of cell line means for each seed to the collection of all seed cell line means, with the null hypothesis that each seed has the same distribution of mean effect magnitude across cell lines. Using the F-test, we found that 66% of seeds had smaller variances of mean leave-one-out correlation than the bulk population ($q < 0.25$), implying that some seed sequences have similar magnitudes of effect across multiple cell lines (S3A Fig)—i.e., that some seeds induce consistently small or large gene expression changes across lines (S5 Data). By comparison, 16% of genes had smaller variance ($q < 0.25$) than the bulk population (S3B Fig). This analysis reveals that many seeds show consistent magnitude of effect across cell lines and that it may be possible to predict whether a particular shRNA will be predominantly off target using empirical data from other cell lines. This also suggests that

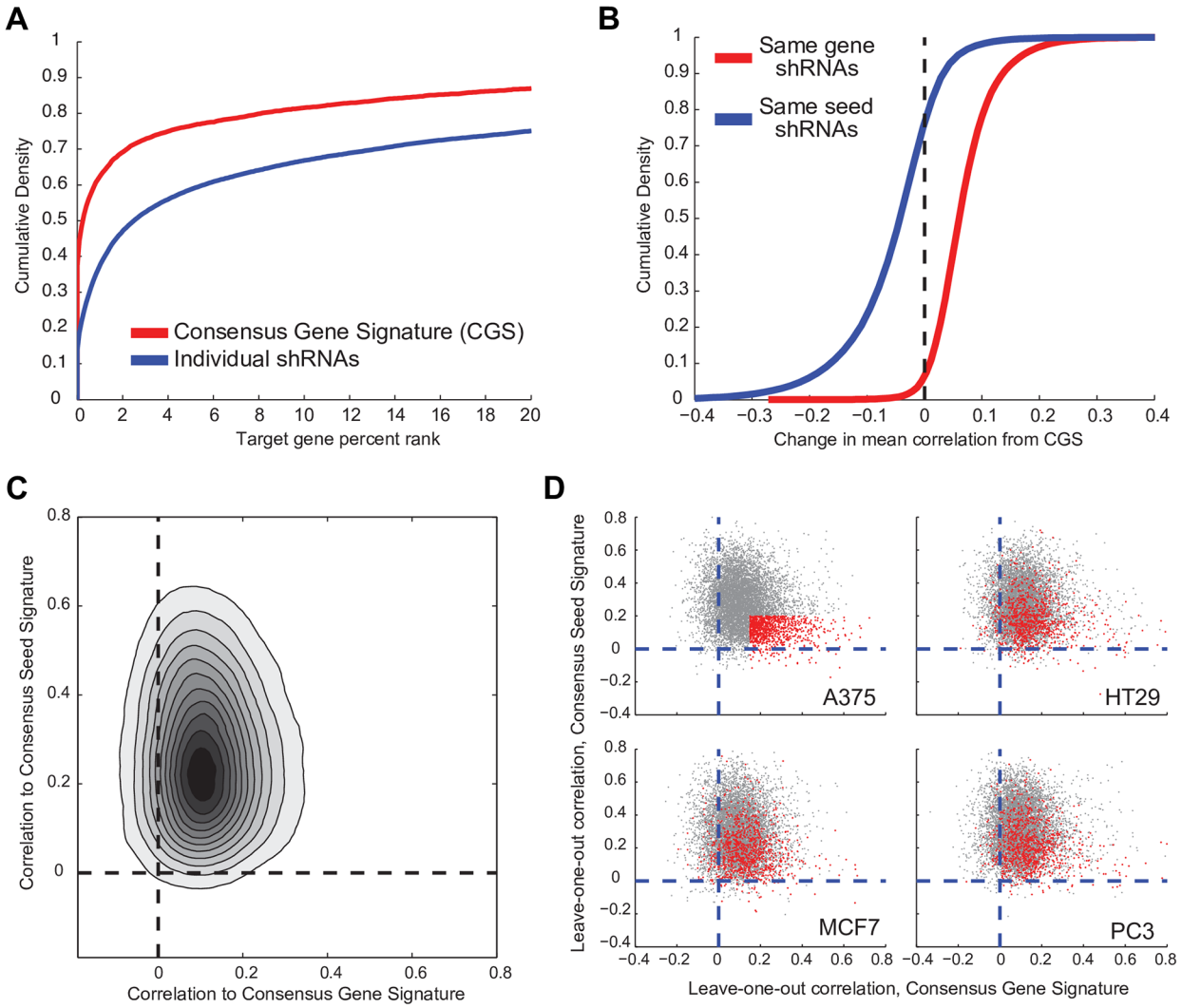

Fig 3. CGS enhances on-target signal and mitigates off-target effects. (A) For directly-measured landmark transcripts targeted by RNAi, cumulative distribution of the rank of that transcript in the resulting signature when using either the signature produced by an individual shRNA or the CGS. (B) Cumulative distribution plot of the change in the correlation to the CGS for either individual shRNAs that target the gene or for shRNAs that share the same seed as one of the shRNAs that contributed to the CGS. (C) For individual leave-one-out shRNAs, comparison of the correlation to the CGS and the analogous CSS. The density color scale is linear. (D) Comparison of on- and off-target activity across cell lines. Top left: for each shRNA in A375 cells, the plot shows the correlation between CGS (x-axis) and CSS (y-axis). Those in red have minimal off-target effects (CSS Spearman correlation <|0.2|) and substantial on-target effects (CGS Spearman correlation >0.15). Remaining panels: the red-highlight shRNAs from A375 cells are highlighted in red in 3 other cell lines. CSS, consensus seed signature; CGS, consensus gene signature; RNAi, RNA interference; shRNA, short hairpin RNA.

https://doi.org/10.1371/journal.pbio.2003213.g003

certain seeds—those with consistently large gene expression changes—should be avoided in shRNA library design, while those seeds with smaller effect sizes should be preferred.

Evaluation of CRISPR-Cas9 signatures

CRISPR technology provides an alternative means of examining the effect of gene silencing. To measure the performance of sgRNAs in gene expression space, we used on-target activity rules [44] to design 6 or 7 sgRNAs against 53 genes, 20 of which were directly measured landmark transcripts, along with negative controls targeting EGFP, beta-galactosidase, and luciferase, for a total of 373 sgRNAs. These were profiled at 96 hours postinfection in 6 cell lines

engineered to express Cas9. To directly compare RNAi results to the CRISPR dataset, we also evaluated from the CMAP database the 395 unique shRNAs targeting the same genes in the same cell lines, for a total of 3,568 matched signatures. Fifty of the 53 target genes from the CRISPR dataset had analogous RNAi data measured at the same time point and the same cell line; we note that this set of genes was biased towards those that showed evidence of a gene expression signature when assessed by RNAi.

We first investigated the change in mRNA of the targeted landmark genes perturbed by RNAi and CRISPR technologies by measuring the z-scores for the differential expression of the target gene relative to the population (Fig 4A). We observe comparable reduction in the

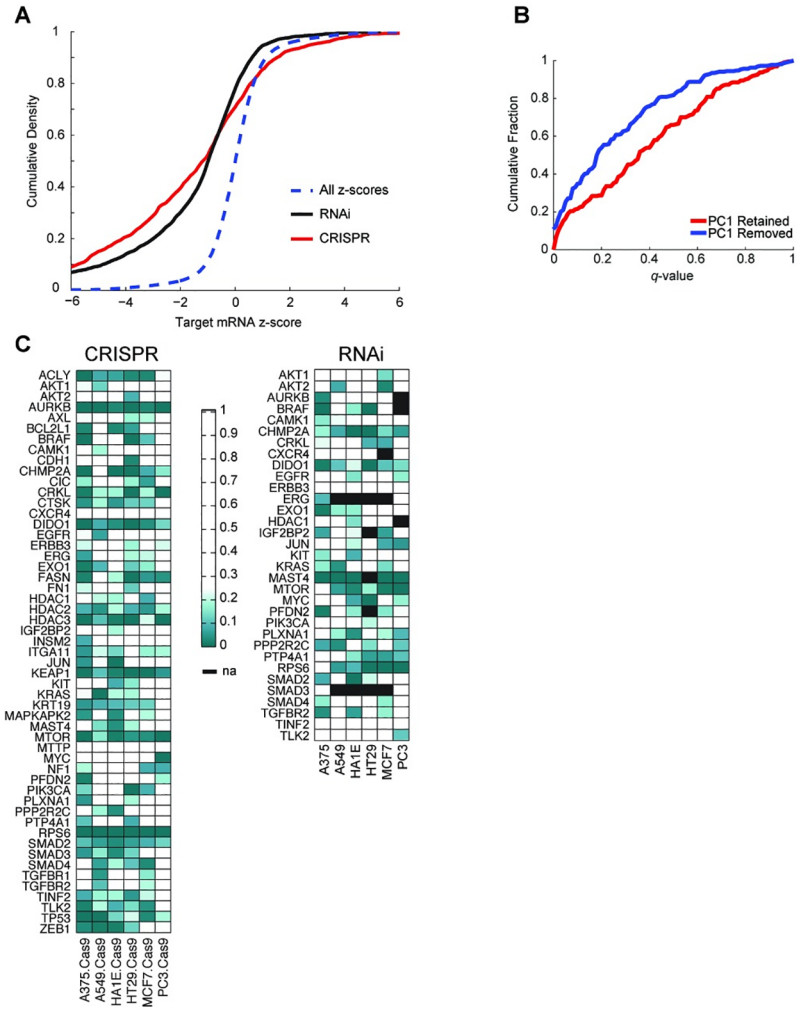

Fig 4. Gene expression analysis of CRISPR-Cas9 reagents. (A) Analysis of landmark transcript reduction, comparing CRISPR and RNAi for genes targeted by both technologies. The dotted line (blue) is a null distribution of the set of all z-scores. Both technologies show significant down-regulation of directly measured target transcripts. (B) As in Fig 2E, comparison of holdout results either retaining or removing PC1 for the CRISPR dataset. (C) Holdout analysis for genes assayed by CRISPR (left) and RNAi (right). Genes are shown for RNAi only if they were also assayed by CRISPR; furthermore, because holdout analysis requires at least 6 independent reagents, not all of the genes assayed by CRISPR have sufficient coverage by RNAi; missing values in some cell lines are indicated by black boxes. Smaller *q*-values (green) indicate greater statistical significance, i.e., that the CGS is valid. See S6 Data. Cas9, CRISPR-associated 9; CGS, consensus gene signature; CRISPR, clustered regularly interspaced short palindromic repeat; PC1, first principal component; RNAi, RNA interference.

https://doi.org/10.1371/journal.pbio.2003213.g004

transcript abundance of the target gene from the 2 technologies. Presumably, while mRNA is not the direct target of an sgRNA, the nonsense-mediated decay pathway reduces the steady-state level of mRNAs carrying frameshifts caused by Cas9-mediated cleavage and NHEJ-mediated repair. We next employed holdout analysis to gauge the efficacy and consistency of sgRNAs, as previously done with shRNAs, including a comparison of results with and without PC1. The PC1 direction among sgRNA signatures was similar to that among shRNAs—the 2 have a Pearson correlation of 0.84 in the landmark basis. We next performed holdout analysis and, as with shRNAs, observed that removal of PC1 led to more statistically significant connections across a range of FDRs (Fig 4B). In the gene-matched shRNA data, of 297 gene-cell line pairs, 183 had 6 or more independent signatures, which is the minimum necessary to run the holdout analysis (Fig 4C). These results show that both RNAi and CRISPR technologies can generate reproducible, on-target results.

A large number of DNA cuts by CRISPR-Cas9 in a cell can produce viability effects, either because the target site is amplified or because of cutting at off-target sites [45–47]. We investigated systematic gene expression changes due to copy number amplification by training a linear model to predict the copy number of target genes. We first attempted to use the hallmark DNA repair gene set from MSigDB but observed no enrichment of that gene set with increasing copy number [48]. Because the measured landmark gene set contains only a fraction of genes, it may be too small to show gene set enrichment. To search for a copy number signal with an unbiased approach, we trained a lasso-regularized linear regression model using the set of all reproducible sgRNA signatures to predict copy number annotations from the Cancer Cell Line Encyclopedia [49]. Using 20-fold cross validation, we found a small but statistically significant correlation; the predicted vector had a Pearson correlation of 0.30 with the copy number data ($p < 4 \times 10^{-15}$). Adding credence to the model, nontargeting control sgRNAs had less variance along this axis than did reproducible sgRNAs with a gene target ($p = 6 \times 10^{-6}$ by the F-test). Importantly, the variance explained by this axis is about 10-fold smaller than the variance explained by on-target activity of reproducible sgRNAs; in the future, application of a nonlinear model to a larger dataset may be informative to better quantify this effect. This result suggests that there is indeed a copy number signal in gene expression space that should be considered when interpreting experimental results but that, in most cases, it is small enough that it should not typically obscure the on-target gene signal.

Projection analysis

Unlike the seed-based off-target effects of shRNA reagents, which comprise a large fraction of the gene expression change, the penetrance of systematic off-target effects of sgRNAs is unknown. To interrogate the magnitude of on- and off- target effects in both sgRNA and shRNA signatures, we formulated a novel decomposition analysis using projection. A gene expression signature is a combination of 3 components: on-target activity, off-target effects, and assay noise. Projection estimates on-target activity by comparing the signature of an individual perturbation to signatures from other perturbations targeting the same gene and quantifies off-target activity by measuring the similarity among replicates of a signature after subtracting the on-target activity. Because the method estimates off targets by searching for reproducible elements of the signature that are not related to on-target activity, it does not require prior knowledge of the off-target mechanism to accurately detect an off-target signal.

We first assessed the consistency of projection decomposition to detect on-target effects. Of 2,231 sgRNA signatures, 1,723 individual sgRNAs (77%) had significant on-target activity ($q < 0.25$) by this analysis; for the matched 2,183 shRNA signatures, 931 (43%) were significant. To evaluate whether projection corroborates holdout analysis—an alternate method for

measuring on-target activity—we then examined signatures belonging to genes with statistically significant holdout results and saw these were preferentially enriched by projection: 91% of individual sgRNAs and 61% of individual shRNAs from genes with significant holdout results were found to have significant on-target projection ranks. Because the holdout method requires 2 disjoint sets to arrive at the same answer, it is stringent, requiring a majority of the signatures to demonstrate on-target activity. Therefore, individual signatures with on-target activity are necessary but not sufficient to give a statistically significant holdout result.

We next examined the ability of projection decomposition to reveal known off-target effects: in this case, the miRNA-seed effect in shRNAs. For each shRNA, we compared its projection estimate of off-target activity to its correlation to the leave-one-out CSS and found a significant relationship (Pearson = 0.4275, $p < 10^{-10}$, Fig 5A). While these 2 metrics are not strictly identical, their correspondence indicates that decomposition by projection—by modeling on-target activity with independent perturbations and off-target effects with assessment of technical replicates—can reveal off-target effects without prior knowledge of their source.

We then compared the relative magnitude of on- and off-target effects for RNAi and CRISPR reagents. Considering only the 1,723 sgRNA signatures with statistically significant on-target activity, we observed an average on-target magnitude of 0.211 and off-target magnitude of 0.062, with 97.4% having a larger on-target than off-target component (Fig 5B and 5C). For the 924 shRNAs targeting the same genes with significant on-target projection ranks, the mean on-target magnitude was comparable to sgRNA-based knockout, 0.197; however, the mean off-target magnitude was 0.230, and only 41.8% shRNAs had a larger on-target than off-target component (Fig 5B and 5C). Thus, while RNAi reagents necessarily engage the miRNA pathway to cause off-target effects, these results suggest the exciting possibility that CRISPR technology can frequently produce faithful on-target signatures with low levels of off-target signal.

Connections between RNAi and CRISPR

We compared the signatures produced by RNAi and CRISPR technology to each other utilizing query-based metrics, as this is, ultimately, the most important measure of the quality of these data in CMAP. For each gene in each cell line, we created a CGS using all sgRNAs (with PC1 removed) and queried the CMAP database of about 4,000 shRNA-derived CGSs (also with PC1 removed). Of 297 sgRNA CGSs for which there was a same-gene shRNA CGS, 116 (39%) had statistically significant connectivity ($q < 0.25$, Fig 6A). Furthermore, when we require that both the sgRNA and shRNA CGSs separately pass holdout analysis, 74% (37/50) connect with $q < 0.25$ (Fig 6B). These results show that for many of the genes examined here, knockdown and knockout produce similar signatures when each is independently verified for on-target signal.

Interestingly, we find examples in which holdout analysis suggests the resulting CGSs are an accurate measure of the on-target effects of gene knockdown and gene knockout, but these fail to find each other in a query, such as SMAD2 in 3 different cell lines; these may be cases in which there are phenotypic differences as a function of gene dosage. It is also possible that different kinetics of gene depletion between the 2 technologies may affect connectivity results, or that the signatures could be false positives from the holdout analysis. These results suggest that, going forward in CMAP, CRISPR technology should be the first choice for conducting loss-of-function studies, with RNAi used as a complementary or secondary assay. The high on-target activity and low prevalence of off-target effects will enable the use of smaller numbers of perturbations per gene with CRISPR technology, which in turn will allow greater coverage across both genes and cell types.

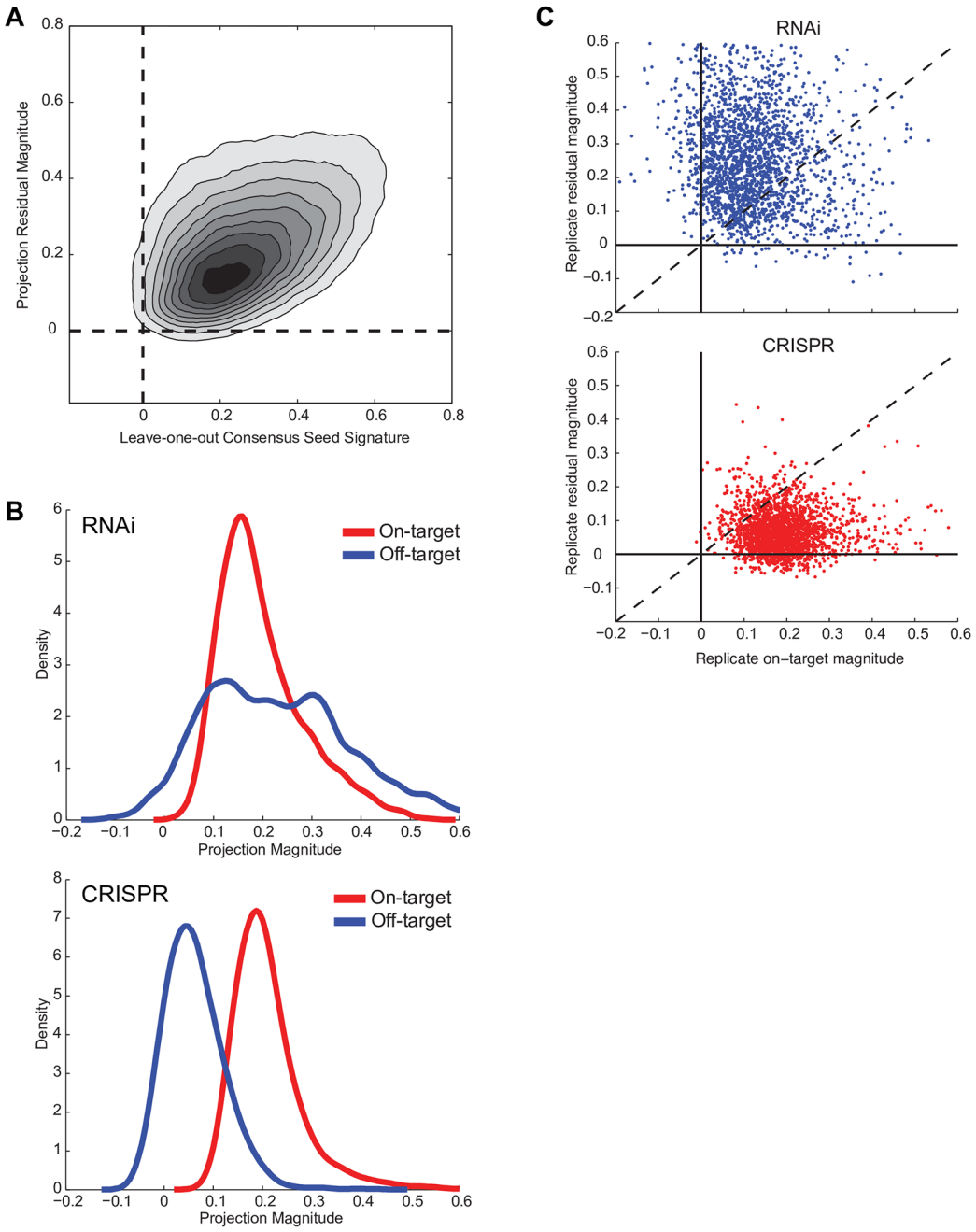

Fig 5. Decomposition by projection. (A) For shRNAs, the magnitude of the off-target effect comparing the leave-one-out CSS to projection. Pearson correlation coefficient = 0.43. (B) For individual shRNAs (top) and sgRNAs (bottom), in which the on-target magnitude passes FDR < 25%, distribution of on- and off-target magnitudes, as assessed by projection decomposition. (C) Scatter plots of on-target and off-target projection magnitudes for RNAi (top) and CRISPR (bottom) for all of the signatures of reagents in the dataset. While the 2 technologies show similar on-target activities, RNAi shows large off-target effects. CRISPR, clustered regularly interspaced short palindromic repeat; CSS, consensus seed signature; FDR, false discovery rate; RNAi, RNA interference; sgRNA, single guide RNA; shRNA, short hairpin RNA.

https://doi.org/10.1371/journal.pbio.2003213.g005

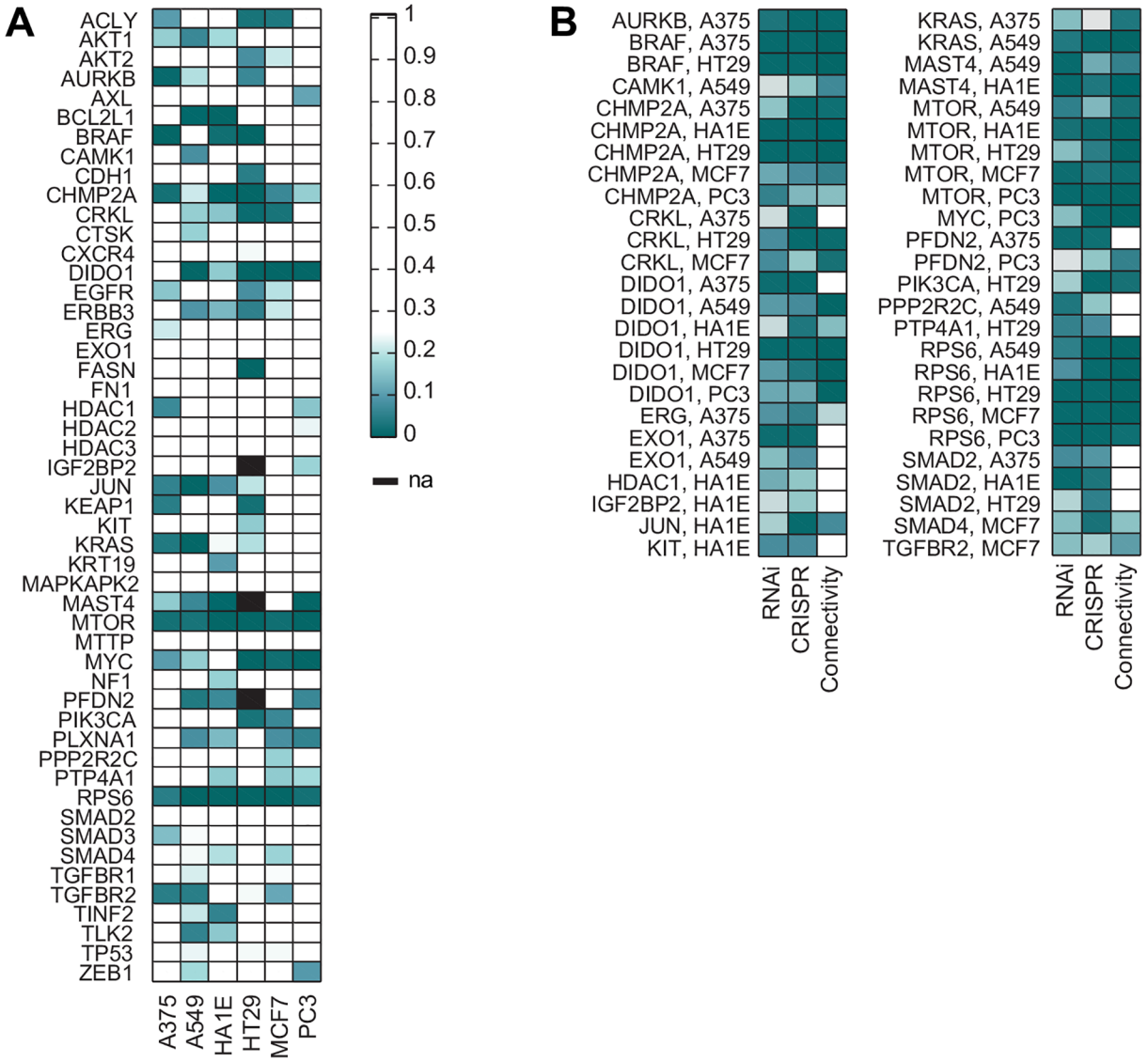

Fig 6. CMAP queries for RNAi and CRISPR reagents. (A) For all genes assessed by both CRISPR and RNAi technologies, the *q*-values for querying CMAP with the CRISPR CGS and its resulting connectivity to the RNAi CGS. See S7 Data. (B) Same as in (A), but only for genes passing holdout analysis (*q*-values < 0.25) by both technologies individually in a cell line, the *q*-values for connectivity. Holdout analysis *q*-values are plotted for each technology in the first 2 columns; connectivity *q*-values are plotted in the third column. See S7 Data. CGS, consensus gene signature; CMAP, Connectivity Map; CRISPR, clustered regularly interspaced short palindromic repeat; RNAi, RNA interference.

https://doi.org/10.1371/journal.pbio.2003213.g006

Discussion

For over a decade, RNAi has been widely used to induce gene suppression, especially in mammalian cells. More recently, CRISPR technology has emerged as an exciting new tool for loss-of-function perturbational studies. Using RNAi and CRISPR datasets generated as part of CMAP, we have characterized the gene expression changes induced by these 2 perturbation types and shown that the on-target efficacy is similar. Analysis of off-target activity, however, shows that miRNA-like seed effects of RNAi are widespread, while CRISPR technology produces few systematic off-target effects. In many cases, we observe that RNAi and CRISPR produce comparable signatures to each other once nonspecific and off-target effects are mitigated.

Analysis of large-scale gene expression data required the development of several analytical approaches. We have introduced the notion of a CGS to combine information from multiple independent perturbations, as is commonly done in the analysis of genetic screens using these technologies: concordance of multiple perturbations of different sequences targeting the same gene suggests that the observed phenotype is due to on-target silencing [50]. While the CGS approach is effective, future work will refine the CGS concept, as has been done with small-molecule gene expression profiles [51], including using inference to model the on- and off- target components of a signature. Likewise, we developed holdout analysis to assess the reliability of a CGS. This approach, however, is stringent, requiring consistent on-target effects among a majority of the individual perturbations targeting the same gene; while disjoint CGSs might fail to correlate, the CGS from the entire set of perturbations could still be valid even if only a minority of perturbations are effective. Finally, we used projection to quantify the magnitude of on- and off- target effects of a perturbation. This method is powerful because it does not require prior knowledge or characterization of off-target mechanisms to observe their effect. In the course of these analyses, we observed that the PC1 was consistent, irrespective of perturbation type, and found that removal of PC1 led to improved results when combining multiple signatures. Removal of principal components to improve signal has precedent in many circumstances, for example, genome wide association studies (GWAS) and image processing. For large-scale, high-throughput gene expression datasets, employing and improving these methods will be critical to discriminate valid gene perturbation signatures.

This is an exciting time for loss-of-function screens with genetic perturbations. RNAi experiments, when properly controlled and analyzed, have a proven record of generating discoveries. The prevalence of off-target effects detected at genome scale with RNAi, however, necessitates careful interpretation of results to accurately determine loss-of-function phenotypes—in particular, either computationally or experimentally accounting for the penetrance and heterogeneity of seed-based effects. Improved design of RNAi reagents can help decrease off-target effects, but the entry of small RNAs into the endogenous miRNA pathway is largely unavoidable [28,52,53]. Furthermore, these results emphasize the long-standing recommendation on the use of multiple reagents to gain confidence in a reproducible phenotype. CRISPR technology, when analyzed by the same metrics for consistency of reagents targeting the same gene, tends to produce predominantly on-target activity. The results presented here, assaying several hundred unique sgRNAs, reinforce critiques of a recent study that suggested massive, sequence-independent off targets triggered by CRISPR-Cas9 [54]. Given the minimal off-target activity of CRISPR-Cas9 in these assays, agreement among a small number of independent sgRNAs may be sufficient evidence of on-target activity in a majority of cases—an important consideration when it comes to building a genome-wide map across many cellular contexts of the expression consequences of loss-of-function reagents. The proper use of these 2 technologies can synergistically bring new value to the use of genetic perturbations for gene function discovery and therapeutic development.

Materials and methods

L1000 platform

In brief, the L1000 Luminex assay explicitly measures 978 specific mRNA transcripts designated as landmarks. Samples are run in 384 well plates, and the expression values are normalized z-scores across the distribution of expression values for each gene across a single plate. A signature of a perturbation—a small molecule/compound, an shRNA, an sgRNA, or an ORF—is an average of the 978 differentially expressed transcripts of 3 biological replicates of cells treated with that perturbation. The signature consists of 978 z-scores indicating gene

expression changes in the measured landmark genes due to treatment by the perturbation. Similarities between signatures were evaluated using Spearman correlation between z-scores in only the landmark space of 978 genes.

Lentivirus reagents and infections

For RNAi reagents, shRNAs were from the TRC collection, cloned into pLKO (Addgene #10878). For CRISPR reagents, sgRNAs were cloned into pLentiGuide (Addgene #52963). Constructs were made in virus as described previously [13]. Following virus production, we first perform a lentiviral infection efficiency assay to determine the proper volume of virus to use going forward. Various volumes of virus are added in duplicate to cells, and the next day, puromycin is added to one replicate and no selection is added to the other replicate. The cells are assayed by Cell Titer Glo 3–5 days postinfection, and the ratio of viability with and without puromycin is used to determine infection efficiency. We then select for further use the volume of virus that gives about 75% infection efficiency via this assay, which corresponds to an MOI of about 1.4. This same protocol is used for both shRNAs and sgRNAs.

For Cas9 lines, the Cas9-expressing cell lines were generated by infection with the pLX_311-Cas9 lentivirus (Addgene #96924). Because this is a low-titer virus, the MOI was low (0.2 or less), and thus the majority of the cells that survive blasticidin selection will have a single integrant.

Data curation

We performed analyses on experiments from 8 cancer cell lines (A375, A549, HCC515, HEPG2, HT29, MCF7, PC3, VCAP) and 1 immortalized cell line (HA1E); all lines were lysed 96 hours post viral infection except for VCAP, which was lysed at 120 hours to account for its slower growth rate. All shRNAs were run in triplicate, with each specific shRNA L1000 signature representing a weighted average between the 3 technical replicates.

Relative on- and off-target strengths

To quantify the relative strength of on-target and off-target effects, we compared the distribution of Spearman correlations for all pairs of shRNAs sharing the same target gene and all shRNAs sharing the same 6-mer or 7-mer seed sequence, compared to correlations of all pairs of shRNAs. Within each cellular context, we identified all pairs of shRNAs that share either seed sequence or target gene and took a rank-based Spearman correlation between their expression profiles. We compared the distribution of same-seed shRNA correlations both for the 6-mer and 7-mer definitions of seed region and found comparable effects.

FDRs

To correct for multiple hypothesis testing, we use the Storey's FDR correction to calculate a q-value given the p-values. First, we assign p-values in general by comparing observed statistics to permutation nulls. For example, in holdout analysis, we calculate the mean correlation between 2 consensus signatures iteratively generated from a set of signatures targeting the same gene. We generate a null distribution by applying the same procedure to permuted random groups, and we assign p-values to the measured values by ranking relative to the permutation null values. The interpretation of Storey's q-value is that from a large collection of hypotheses, the set of hypotheses (numbering N) with q-value $< t$ will have, at most, a fraction t of type 1 errors or false positives. That is, on average, under the null hypothesis modeled by the permutation nulls, the expected number of type 1 errors is at most $N^* t$. We use a q-value

threshold of 0.25 for most of our analyses; as CMAP is designed to be a screening tool leading to hypotheses for follow-up experiments, we are willing to tolerate a relatively higher number of false positives.

PC1

The analysis of the principal components of the data used the standard PCA Matlab code, which in turn uses SVD to linearly transform the data into a new orthogonal basis. The data that were used in the analysis are the differentially expressed genes—the z-scored signatures—in which the means of each gene over all signatures are approximately zero, and the sign and magnitude in a signature indicates the response of that gene to the perturbation, relative to the population. Note that PCA first mean centers the data, i.e., the mean of each gene is set to 0 before calculating the variance. In the analysis, we initially ran PCA on each cell line and perturbation type independently, although our final analysis considered the dataset as a whole. The PC1 is the linear projection of the data with maximal variance—the axis with the greatest variation in the data. The weighting of PC1 for a signature is the magnitude of the signature in the first principal component direction, and by definition, the PC1 weighting across the signatures has larger variance than any other principal component weighting. PC1 for the dataset as a whole is provided in the data repository.

Our findings from PC1 are surprising for several reasons. First, the direction of the PC1—its representation in landmark gene space—is virtually identical across cell lines and perturbation types. The direction of PC1 among small molecule treated A375 signatures is the same as that among shRNA signatures in MCF7, for instance. This suggests that, independent of treatment and cell line, there is an axis along which genes are consistently modulated—implying a generic response. The weightings of PC1 are surprisingly large, accounting for 10%–20% of the variance of signatures. Furthermore, the weights of PC1 increase among consensus signatures—in particular, the fraction of variance from PC1 increases, even among null consensus signatures, as more shRNAs are averaged together. This necessitates using a size-matched permutation null distribution—with the same number of shRNAs in each CGS—for each size of CGS, but it also suggests that the increase in correlation variance among CGSs of increasing size may be due to PC1.

Removing the PC1 is straightforward. First, we ran PCA on the entirety of the CMAP data, constituting over 400,000 signatures, and used the PC1 from this global calculation as the axis on which all signatures were projected. The alternative would be to remove the local PC1, but while our analysis has shown this to be very consistent, running PCA on smaller datasets would introduce batch-specific uncertainty in the PC1 removal. To remove PC1, the dataset is mean centered, the component of the signatures in the direction of PC1 is subtracted, and the mean is added back on. The signal may be further improved by mean centering as part of the normalization process—as opposed to median centering—but this possibility has not been exhaustively explored. Finally, the direction of PC1 depends on both normalization and the calculation of differential expression. Our approach computes the z-score relative to each measured gene independently, so it should be equivalent to the quantile normalized data if each gene is scaled to variance 1.

CGS

To generate a CGS, we group all shRNA signatures targeting the same gene within the same cell line and apply the modz algorithm. We create a pairwise Spearman correlation matrix between the expression profiles of all signatures in this group, explicitly setting the diagonal of the correlation matrix to 0. The weight assigned to each shRNA signature is given by the sum

across its corresponding row of the correlation matrix, with the weights normalized to sum to 1. The CGS is then given by a linear combination—or a weighted average—of the shRNAs, with coefficients given by the weights.

Target gene rank. To evaluate the rank of the target gene, we considered only the shRNA signatures for which the targeted gene was measured explicitly as one of the 978 landmark genes. For each of these signatures, we ranked the 978 genes from most down-regulated to most up-regulated. We identified the rank (1 to 978) of the targeted gene and plotted the distribution of target gene ranks for each individual shRNA signature. We then formed CGSs from the individual shRNAs and similarly identified the rank of the specifically targeted gene. We plotted the cumulative distribution of the individual shRNA target gene ranks compared to the CGS target gene ranks as a function of percent rank.

CGS-shRNA correlations. We evaluated the enrichment of on-target effects by evaluating whether an individual shRNA signature correlates better to a disjoint CGS than to individual shRNAs used to create that CGS. For each gene and cell line with 2 or more shRNA signatures, for each signature in the group, we generated a CGS using all but one of those signatures. We compared the correlation between the CGS and the excluded shRNA to the mean correlation between the excluded shRNA and the shRNAs comprising the CGS. The difference between these 2 quantities is the improvement from using a CGS compared to a typical individual shRNA signature.

Holdout analysis and CGS–CGS correlations. We evaluated the enrichment in overall correlation between CGSs for the same gene. Within each cell line, we partitioned the shRNAs into 2 groups for genes with 6 more shRNAs. We generated 2 CGSs with the 2 groups and computed the correlation between those 2 consensus signatures. We partitioned the groups 30 times and recalculated the CGS correlation to avoid bias by a particular combination of shRNAs and summarized these partitions by taking the median correlation. To mitigate any artificial increases in group correlation introduced by the CGS process, we introduced a permutation null by generating 10,000 identically sized groups of shRNAs targeting different genes and performing the same partition correlation calculation. For each gene, we assigned a p-value of enrichment by counting what fraction of permutation null groups had higher median correlation than the observed median correlation for that gene and cell line. To correct for multiple hypothesis testing, we assigned q-values using Storey's FDR procedure and considered significant groups with q-values <0.25. A gene in a cell line with a statistically significant holdout result indicates that independent CGSs agree better than would be expected assuming no common on-target effects among the shRNAs. Such a result indicates that the CGS of all of the shRNAs is representative of the on-target gene expression consequence of the target gene knockdown.

Leave-one-out consensus signature correlations

We identified the group of shRNAs with at least one same-gene shRNA and at least one same-6mer-seed shRNA. For each of these shRNAs, we generated a CGS with the same-gene shRNAs and a CSS with the same-6mer-seed shRNAs (excluding the shRNA itself in both groups). We found the Spearman correlation between the shRNA and its CGS and the shRNA and its CSS. We plotted the overall shRNA-CGS correlations on the x-axis and the shRNA-CSS correlations on the y-axis to estimate the magnitudes of the on- and off-target effects within a given shRNA.

Projection analysis

The projection method decomposes a signature of a perturbation into 2 orthogonal components representing the on-target and reproducible off-target effects without incorporating

information about the mechanism of the off targets. The most important outputs from projection are 2 quantities representing the normalized length of the on-target component and reproducible off-target component of the unit vector of the signature. First, a reference signature is constructed from other perturbations with the same purported on target—i.e., a CGS of other shRNAs or sgRNAs targeting the same gene. We calculate the cosine similarity of the given signature to the reference and assign p-values by ranking that similarity relative to a number of permutation null reference signatures. The permutation nulls are CGSs constructed from unrelated groups of the same number of shRNAs or sgRNAs. For the projection magnitudes to be meaningful, the signature must have statistically significant similarity to the reference ($q < 0.25$). Without statistically significant similarity, either (1) the signature lacks measurable on-target activity or (2) the reference signature is not a good estimate of the true on-target signature. In the latter case, the true on-target component of the test signature will be incorrectly assessed as reproducible off target.

If the test signature has significant similarity to the reference, residuals are calculated from the test signature's replicates by projecting the replicates onto the subspace orthogonal to the reference signature. The on-target magnitude is given by the mean of the length of the replicates projected onto the reference divided by the length of the replicates; equivalently, the on-target magnitude is the mean inner product of the replicate unit vectors with the unit vector reference in the standard landmark basis. The off-target magnitude is given by the mean of the pairwise inner product of the replicate residuals normalized by the length of the replicates. It follows from the Pythagorean theorem that for a, the on-target magnitude, and b, the off-target or residual magnitude, that $a^2 + b^2 < 1$. In fact, $a^2 + b^2$ for a given signature is identically the fraction of the signature that is reproducible, and this quantity correlates very well with the standard CMAP metric of reproducibility for a signature—the 75th quantile of Spearman correlation among replicates of that signature (S4 Fig).

The interpretation of projection is straightforward: any reproducible component of a signature orthogonal to the on-target activity is an unintended off target—whether systematic biology, batch effect, or other artifact. The reproducibility of off targets can be measured by considering the components of replicates of a signature orthogonal to an estimated reference. In general, we expect small off targets because of batch effects and errors in calculating the reference signature—the on-target activity is not perfectly known, which is the original problem. That the calculated off-target activity correlates with the magnitude of the known seed effects in shRNA signatures, as measured by correlation with the CSS, validates the approach.

Supporting information

S1 Fig. Control shRNAs and the seed effect. (A) An example of strong correlation: 2 replicates of the same control shRNA show strong reproducibility as measured by the Spearman correlation of the differential expression of the 978 measured landmark genes. Each point represents 1 landmark transcript. (B) An example of no correlation between signatures of different shRNAs. (C) 7-mers beginning at positions 11 and 12 of the annotated sense strand show the greatest correlation, corresponding to the seed sequence of the antisense/targeting strand, and reflecting heterogeneity of Dicer processing. See S1 Data. (D) Schematic of shRNAs used in CMAP. The 21-nt sense strand, which has the same sequence as the target mRNA, is underlined and numbered. The blue highlight indicates the major siRNA product produced after Dicer processing. The bolded, yellow nts indicate the seed sequence of the antisense/targeting strand. CMAP, Connectivity Map; nt, nucleotide; shRNA, short hairpin RNA; siRNA, small interfering RNA.
(PDF)

S2 Fig. The consensus signature and the PC1. (A) Schematic of holdout analysis. For genes with 6 or more shRNAs, the CGS is calculated from subsets of shRNAs, and the resulting CGSs are correlated. This procedure is repeated with different random partitions of shRNAs. (B) The distributions for null signatures, comprising random draws of shRNAs, with PC1 removed. Compare to Fig 2B. (C) Comparison of holdout correlations to permutations nulls in an example cell line. Top: unmodified signatures; Bottom: PC1 removed. Removing PC1 increases the fraction of genes with a statistically significant correlation. (D) For each gene, the number of cell lines in which the CGS passes statistical significance for holdout analysis for unmodified data (black) and for data with the PC1 removed (gray). See S4 Data. CGS, consensus gene signature; shRNA, short hairpin RNA; PC1, first principal component.
(PDF)

S3 Fig. Leave-one-out consensus correlation. For each seed sequence (A) or gene target (B), the mean correlation to the leave-one-out CSS or CGS is calculated within each cell line. The vector of means by cell line for each seed or gene can be compared to the collection of means by cell line for all seeds or genes. Seeds and genes that cause gene expression changes of the same magnitude (not necessarily direction) will tend to have smaller variance than the population, reflecting that consistency. These scatter plots show the mean (x-axis) and standard deviation (y-axis) of the vector of means across the 9 cell lines. Those in red are different than the population per the F-test at an FDR of <25%. Note that the threshold for significance is different for genes and seeds because genes and seeds have different leave-one-out correlation distributions. CGS, consensus gene signature; CSS, consensus seed signature; FDR, false discovery rate.
(PDF)

S4 Fig. Projection magnitudes recapitulate other measures of reproducibility. The projection algorithm decomposes any particular signature into on-target and residual relative magnitudes. These represent the entirety of the reproducible part of the signature; the remainder is noise. The Pythagorean theorem gives the length of that reproducible part as the square root of the sum of the squares of the 2 projection magnitudes. This projection length correlates very well with the independently measured replicate correlation. For the set of all shRNAs, the x-axis is the calculated projection length, and the y-axis is the 75th quantile of Spearman correlation among replicates. The 2 quantities correlate with a Pearson correlation of 0.78, providing a sanity check that the projection outputs are meaningful. The nonlinearity is due to the use of the 75th quantile for replicate correlation by convention, compared to the use of the mean for projection length. shRNA, short hairpin RNA.
(PDF)

S1 Data. Defining the window of seed effects. All windows of 7 nucleotides of the sense strand of the shRNA were examined for correlation between pairs sharing the same seed sequence. shRNA, short hairpin RNA.
(TXT)

S2 Data. Comparison of same-gene and same-seed effects. The fraction that are significant at a false discovery rate <0.25 is shown. For same-seed effects, a 6-nucleotide seed window was used.
(TXT)

S3 Data. Correlation of the first principle component across different types of perturbations.
(TXT)

S4 Data. The number of genes that pass holdout analysis with a false discovery rate <0.25, binned by the number of cell lines in which they pass holdout analysis.
(TXT)

S5 Data. Bad seeds. Please see ReadMe tab in the file for a description.
(XLSX)

S6 Data. Holdout analyses for RNAi and CRISPR. For genes assayed by both technologies, holdout analysis results. CRISPR, clustered regularly interspaced short palindromic repeat; RNAi, RNA interference.
(XLSX)

S7 Data. Connectivity between consensus gene signatures for RNAi and CRISPR. CRISPR, clustered regularly interspaced short palindromic repeat; RNAi, RNA interference.
(TXT)

Acknowledgments

We thank Oana Enache, Roger Hu, Larson Hogtrom, Jackie Rosains, Aviad Tsherniak, and Mudra Hegde (Broad Institute) for analytical insight and helpful conversations. We thank the entire Genetic Perturbation Platform (GPP) and Connectivity Map (CMAP) teams for support and advice.

Author Contributions

Conceptualization: Ian Smith, Peyton G. Greenside, Todd R. Golub, Aravind Subramanian, John G. Doench.

Data curation: Ian Smith, Ted Natoli, David L. Lahr, Rajiv Narayan, John G. Doench.

Formal analysis: Ian Smith, Peyton G. Greenside, David Wadden, Itay Tirosh, Aravind Subramanian, John G. Doench.

Funding acquisition: Todd R. Golub, Aravind Subramanian.

Investigation: Ian Smith, Peyton G. Greenside, Ted Natoli, David Wadden.

Methodology: Ian Smith, Peyton G. Greenside, Ted Natoli, David L. Lahr, David Wadden, Rajiv Narayan, John G. Doench.

Project administration: Todd R. Golub, Aravind Subramanian, John G. Doench.

Resources: David L. Lahr, Rajiv Narayan.

Software: Ian Smith, Ted Natoli, David L. Lahr, Rajiv Narayan.

Supervision: David E. Root, Todd R. Golub, Aravind Subramanian, John G. Doench.

Validation: David L. Lahr.

Visualization: Ted Natoli, John G. Doench.

Writing – original draft: Ian Smith, Peyton G. Greenside, John G. Doench.

Writing – review & editing: Ian Smith, Peyton G. Greenside, David E. Root, John G. Doench.

References

1. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science. American Association

for the Advancement of Science; 2006; 313: 1929–1935. https://doi.org/10.1126/science.1132939 PMID: 17008526

2. Belcastro V, Siciliano V, Gregoretti F, Mithbaokar P, Dharmalingam G, Berlingieri S, et al. Transcriptional gene network inference from a massive dataset elucidates transcriptome organization and gene function. Nucleic Acids Research. 2011; 39: 8677–8688. https://doi.org/10.1093/nar/gkr593 PMID: 21785136
3. Wang Z, Monteiro CD, Jagodnik KM, Fernandez NF, Gundersen GW, Rouillard AD, et al. Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd. Nature Communications. Nature Publishing Group; 2016; 7: 12846. https://doi.org/10.1038/ncomms12846 PMID: 27667448
4. Pirhaji L, Milani P, Leidl M, Curran T, Avila-Pacheco J, Clish CB, et al. Revealing disease-associated pathways by network integration of untargeted metabolomics. Nature Methods. Nature Research; 2016; 13: 770–776. https://doi.org/10.1038/nmeth.3940 PMID: 27479327
5. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A Next Generation Connectivity Map: L1000 Platform And The First 1,000,000 Profiles. bioRxiv. Cold Spring Harbor Labs Journals; 2017;: 136168. https://doi.org/10.1101/136168
6. Zimmer M, Ebert BL, Neil C, Brenner K, Papaioannou I, Melas A, et al. Small-Molecule Inhibitors of HIF-2a Translation Link Its 5′UTR Iron-Responsive Element to Oxygen Sensing. Mol Cell. Elsevier Ltd; 2008; 32: 838–848. https://doi.org/10.1016/j.molcel.2008.12.004 PMID: 19111663
7. Hahn CK, Berchuck JE, Ross KN, Kakoza RM, Clauser K, Schinzel AC, et al. Proteomic and Genetic Approaches Identify Syk as an AML Target. Cancer Cell. Elsevier Ltd; 2009; 16: 281–294. https://doi.org/10.1016/j.ccr.2009.08.018 PMID: 19800574
8. Liu J, Lee J, Hernandez MAS, Mazitschek R, Ozcan U. Treatment of Obesity with Celastrol. Cell. Elsevier Inc; 2015; 161: 999–1011. https://doi.org/10.1016/j.cell.2015.05.011 PMID: 26000480
9. Kunkel SD, Suneja M, Ebert SM, Bongers KS, Fox DK, Malmberg SE, et al. mRNA expression signatures of human skeletal muscle atrophy identify a natural compound that increases muscle mass. Cell Metab. 2011; 13: 627–638. https://doi.org/10.1016/j.cmet.2011.03.020 PMID: 21641545
10. Echeverri CJ, Beachy PA, Baum B, Boutros M, Buchholz F, Chanda SK, et al. Minimizing the risk of reporting false positives in large-scale RNAi screens. Nature Methods. 2006; 3: 777–779. https://doi.org/10.1038/nmeth1006-777 PMID: 16990807
11. Paddison PJ, Silva JM, Conklin DS, Schlabach M, Li M, Aruleba S, et al. A resource for large-scale RNA-interference-based screens in mammals. Nature. 2004; 428: 427–431. https://doi.org/10.1038/nature02370 PMID: 15042091
12. Berns K, Hijmans EM, Mullenders J, Brummelkamp TR, Velds A, Heimerikx M, et al. A large-scale RNAi screen in human cells identifies new components of the p53 pathway. Nature. 2004; 428: 431–437. https://doi.org/10.1038/nature02371 PMID: 15042092
13. Moffat J, Grueneberg DA, Yang X, Kim SY, Kloepfer AM, Hinkle G, et al. A Lentiviral RNAi Library for Human and Mouse Genes Applied to an Arrayed Viral High-Content Screen. Cell. 2006; 124: 1283–1298. https://doi.org/10.1016/j.cell.2006.01.040 PMID: 16564017
14. Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, et al. RNA-Guided Human Genome Engineering via Cas9. Science. 2013; 339: 823–826. https://doi.org/10.1126/science.1232033 PMID: 23287722
15. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, et al. Multiplex Genome Engineering Using CRISPR/Cas Systems. Science. 2013; 339: 819–823. https://doi.org/10.1126/science.1231143 PMID: 23287718
16. Jinek M, East A, Cheng A, Lin S, Ma E, Doudna J. RNA-programmed genome editing in human cells. eLife. 2013; 2: e00471. https://doi.org/10.7554/eLife.00471 PMID: 23386978
17. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. Science. 2012; 337: 816–821. https://doi.org/10.1126/science.1225829 PMID: 22745249
18. Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelsen TS, et al. Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. Science. 2014; 343: 84–87. https://doi.org/10.1126/science.1247005 PMID: 24336571
19. Wang T, Wei JJ, Sabatini DM, Lander ES. Genetic Screens in Human Cells Using the CRISPR-Cas9 System. Science. 2014; 343: 80–84. https://doi.org/10.1126/science.1246981 PMID: 24336569
20. Koike-Yusa H, Li Y, Tan E-P, Velasco-Herrera MDC, Yusa K. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. Nat Biotechnol. 2013; 32: 267–273. https://doi.org/10.1038/nbt.2800 PMID: 24535568
21. Gilbert LA, Horlbeck MA, Adamson B, Villalta JE, Chen Y, Whitehead EH, et al. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. Cell. 2014; 159: 647–661. https://doi.org/10.1016/j.cell.2014.09.029 PMID: 25307932

22. Jackson AL, Bartz SR, Schelter J, Kobayashi SV, Burchard J, Mao M, et al. Expression profiling reveals off-target gene regulation by RNAi. Nat Biotechnol. 2003; 21: 635–637. https://doi.org/10.1038/nbt831 PMID: 12754523

23. Doench JG, Petersen CP, Sharp PA. siRNAs can function as miRNAs. Genes Dev. 2003; 17: 438–442. https://doi.org/10.1101/gad.1064703 PMID: 12600936

24. Sigoillot FD, Lyman S, Huckins JF, Adamson B, Chung E, Quattrochi B, et al. A bioinformatics method identifies prominent off-targeted transcripts in RNAi screens. Nature Methods. 2012; 9: 363–366. https://doi.org/10.1038/nmeth.1898 PMID: 22343343

25. Kaelin WG. Molecular biology. Use and abuse of RNAi to study mammalian gene function. Science. 2012; 337: 421–422. https://doi.org/10.1126/science.1225787 PMID: 22837515

26. Schwarz DS, Hutvágner G, Du T, Xu Z, Aronin N, Zamore PD. Asymmetry in the assembly of the RNAi enzyme complex. Cell. 2003; 115: 199–208. PMID: 14567917

27. Reynolds A, Leake D, Boese Q, Scaringe S, Marshall WS, Khvorova A. Rational siRNA design for RNA interference. Nat Biotechnol. 2004; 22: 326–330. https://doi.org/10.1038/nbt936 PMID: 14758366

28. Knott SRV, Maceli AR, Erard N, Chang K, Marran K, Zhou X, et al. A Computational Algorithm to Predict shRNA Potency. Mol Cell. 2014; 56: 796–807. https://doi.org/10.1016/j.molcel.2014.10.025 PMID: 25435137

29. Friedman RC, Farh KKH, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. Genome Research. 2008; 19: 92–105. https://doi.org/10.1101/gr.082701.108 PMID: 18955434

30. Chi SW, Zang JB, Mele A, Darnell RB. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature. 2009; 460: 479–486. https://doi.org/10.1038/nature08170 PMID: 19536157

31. Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, et al. DNA targeting specificity of RNA-guided Cas9 nucleases. Nat Biotechnol. 2013; 31: 827–832. https://doi.org/10.1038/nbt.2647 PMID: 23873081

32. Fu Y, Foden JA, Joung JK. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. Nat Biotechnol. 2013; 31: 822–826. https://doi.org/10.1038/nbt.2623 PMID: 23792628

33. Veres A, Gosis BS, Ding Q, Collins R, Ragavendran A, Brand H, et al. Low Incidence of Off-Target Mutations in Individual CRISPR-Cas9 and TALEN Targeted Human Stem Cell Clones Detected by Whole-Genome Sequencing. Cell Stem Cell. 2014; 15: 27–30. https://doi.org/10.1016/j.stem.2014.04.020 PMID: 24996167

34. Tsai SQ, Zheng Z, Nguyen NT, Liebers M, Topkar VV, Thapar V, et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. Nat Biotechnol. 2014; 33: 187–197. https://doi.org/10.1038/nbt.3117 PMID: 25513782

35. Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol. 2016; 34: 184–191. https://doi.org/10.1038/nbt.3437 PMID: 26780180

36. Peck D, Crawford ED, Ross KN, Stegmaier K, Golub TR, Lamb J. A method for high-throughput gene expression signature analysis. Genome Biol. 2006; 7: R61. https://doi.org/10.1186/gb-2006-7-7-r61 PMID: 16859521

37. Guda S, Brendel C, Renella R, Du P, Bauer DE, Canver MC, et al. miRNA-embedded shRNAs for Lineage-specific BCL11A Knockdown and Hemoglobin F Induction. Mol Ther. 2015; 23: 1465–1474. https://doi.org/10.1038/mt.2015.113 PMID: 26080908

38. Singh S, Wu X, Ljosa V, Bray M-A, Piccioni F, Root DE, et al. Morphological Profiles of RNAi-Induced Gene Knockdown Are Highly Reproducible but Dominated by Seed Effects. PLoS ONE. 2015; 10(7): e0131370. https://doi.org/10.1371/journal.pone.0131370 PMID: 26197079

39. Sigoillot FD, King RW. Vigilance and Validation: Keys to Success in RNAi Screening. ACS Chem Biol. 2011; 6: 47–60. https://doi.org/10.1021/cb100358f PMID: 21142076

40. Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, et al. Defining a Cancer Dependency Map. Cell. 2017; 170: 564–576.e16. https://doi.org/10.1016/j.cell.2017.06.010 PMID: 28753430

41. Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci USA. 2000; 97: 10101–10106. PMID: 10963673

42. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38: 904–909. https://doi.org/10.1038/ng1847 PMID: 16862161

43. Buehler E, Chen Y-C, Martin S. C911: A bench-level control for sequence specific siRNA off-target effects. PLoS ONE. 2012; 7(12): e51942. https://doi.org/10.1371/journal.pone.0051942 PMID: 23251657

44. Doench JG, Hartenian E, Graham DB, Tothova Z, Hegde M, Smith I, et al. Rational design of highly active sgRNAs for CRISPR-Cas9–mediated gene inactivation. Nat Biotechnol. 2014; 32: 1262–1267. https://doi.org/10.1038/nbt.3026 PMID: 25184501

45. Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, et al. Identification and characterization of essential genes in the human genome. Science. 2015; 350: 1096–1101. https://doi.org/10.1126/science.aac7041 PMID: 26472758

46. Aguirre AJ, Meyers RM, Weir BA, Vazquez F, Zhang CZ, Ben-David U, et al. Genomic Copy Number Dictates a Gene-Independent Cell Response to CRISPR/Cas9 Targeting. Cancer Discovery. 2016; 6: 914–929. https://doi.org/10.1158/2159-8290.CD-16-0154 PMID: 27260156

47. Munoz DM, Cassiani PJ, Li L, Billy E, Korn JM, Jones MD, et al. CRISPR Screens Provide a Comprehensive Assessment of Cancer Vulnerabilities but Generate False-Positive Hits for Highly Amplified Genomic Regions. Cancer Discovery. 2016; 6: 900–913. https://doi.org/10.1158/2159-8290.CD-16-0178 PMID: 27260157

48. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. Oxford University Press; 2011; 27: 1739–1740. https://doi.org/10.1093/bioinformatics/btr260 PMID: 21546393

49. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. Nature Publishing Group; 2012; 483: 603–307. https://doi.org/10.1038/nature11003 PMID: 22460905

50. Shao DD, Tsherniak A, Gopal S, Weir BA, Tamayo P, Stransky N, et al. ATARiS: Computational quantification of gene suppression phenotypes from multisample RNAi screens. Genome Research. 2013; 23: 665–678. https://doi.org/10.1101/gr.143586.112 PMID: 23269662

51. Duan Q, Reid SP, Clark NR, Wang Z, Fernandez NF, Rouillard AD, et al. L1000CDS(2): LINCS L1000 characteristic direction signatures search engine. NPJ Syst Biol Appl. 2016; 2: 257. https://doi.org/10.1038/npjsba.2016.15 PMID: 28413689

52. Fellmann C, Zuber J, McJunkin K, Chang K, Malone CD, Dickins RA, et al. Functional Identification of Optimized RNAi Triggers Using a Massively Parallel Sensor Assay. Mol Cell. Elsevier Inc; 2011; 41: 733–746. https://doi.org/10.1016/j.molcel.2011.02.008 PMID: 21353615

53. Watanabe C, Cuellar TL, Haley B. Quantitative evaluation of first, second, and third generation hairpin systems reveals the limit of mammalian vector-based RNAi. RNA Biol. Taylor & Francis; 2016; 13: 25–33. https://doi.org/10.1080/15476286.2015.1128062 PMID: 26786363

54. Schaefer KA, Wu W-H, Colgan DF, Tsang SH, Bassuk AG, Mahajan VB. Unexpected mutations after CRISPR-Cas9 editing in vivo. Nature Methods. 2017; 14: 547–548. https://doi.org/10.1038/nmeth.4293 PMID: 28557981