ARTICLES PREPRINT

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

Correction of Off-Targeting in CRISPR Screens Uncovers Genetic Dependencies in Melanoma Cells

Alexandar R. Perez[1,2], Laura Sala[1], Richard K. Perez[3], Joana A. Vidigal[*,1]

CRISPR-based high-throughput screens are a powerful method to unbiasedly assign function to a large set of genes, but current genome-wide libraries yield a substantial number of false positives and negatives. We use a retrieval-tree based approach to accurately characterize the off-target space of these libraries and show that they contain a notable fraction of highly promiscuous gRNAs. Promiscuous gRNAs are depleted from screens in a gene-independent manner, create noise in the data generated by these libraries, and ultimately lead to low accuracy in hit identification. This extensive off-targeting also contributes to low overlap between data generated by independent libraries. To minimize these problems we developed the CRISPR Specificity Correction (CSC), a computational approach that segregates on- and off-targeting effects on gRNA depletion. We demonstrate that CSC is able to reduce the occurrence of false positives, improve hit reproducibility between different libraries, and uncover both known and novel genetic dependencies in melanoma cells.

CRISPR-Cas9 can disrupt loci at genome-scale, and holds the potential of assigning function to loci in an unbiased manner with unprecedented scope and sensitivity[1]. Though relatively recent, CRISPR screens have been successfully employed to understand disease dependencies in numerous settings[1-5] including in vivo[6]. CRISPR libraries have gone through several design iterations to increase on-target efficiency, reduce off-targeting, and consequently improve performance[7-9]. Yet, second generation libraries still yield a significant number of false positives and negatives suggesting that additional improvements to guide design may further increase screen sensitivity and specificity.

During our recent efforts to adapt CRISPR screening to the noncoding genome[10, 11], we found a systematic flaw in the way off-targets of guide RNAs (gRNAs) are identified by design algorithms[11]. The flaw stems from the application of short-read aligners to identify potential off-target loci. Aligners were developed to deal with large datasets generated by high-throughput sequencing and have a trade-off between speed and exhaustive read-matching, leading to truncation of an alignment search if an effort limit is exceeded[12, 13]. Because of this, the adoption of aligners as a strategy to identify potential off-target sites by the majority of gRNA design tools, often leads to a mis-characterization of the potential off-target space of gRNAs even when off-targets have perfect or near-perfect (1 mismatch) complementarity to the guide[11]. As a consequence, tools that rely on aligners fail to discard highly promiscuous gRNAs and often assign to them misleadingly high specificity scores[11], a problem that has been well documented by others[7, 14]. To address this issue we developed GuideScan[11], a retrieval-tree (trie) based gRNA-design algorithm that accurately enumerates all potential off-target sites up to a user-specified number of mismatches.

Here, we use GuideScan to accurately characterize the off-target space of published genome-wide libraries and determine if guide promiscuity compromises their performance. We found that all libraries we analyzed had a substantial fraction of gRNAs with perfect or near-perfect off-targets. Like gRNAs targeting amplified genomic loci[15], these promiscuous gRNAs are depleted from screens in a gene-independent manner and contribute to substantial noise in the data the libraries generate. They also contribute to the occurrence of false-positives and false-negatives following hit-calling, and low overlap between genes identified as essential by independent libraries. To overcome this problem, we trained a gradient-boosted regression tree model to learn the influence of guide specificity features on gRNA depletion in the context of dropout assays. We use data from this model to

[1] Laboratory of Biochemistry and Molecular Biology, National Cancer Institute, Bethesda, MD, USA. [2] Department of Anesthesia and Perioperative Care, University of California, San Francisco, San Francisco, California, USA. [3] School of Medicine, University of California, San Francisco, San Francisco, California, USA. *To whom correspondence should be addressed. Email: joana.vidigal@nih.gov

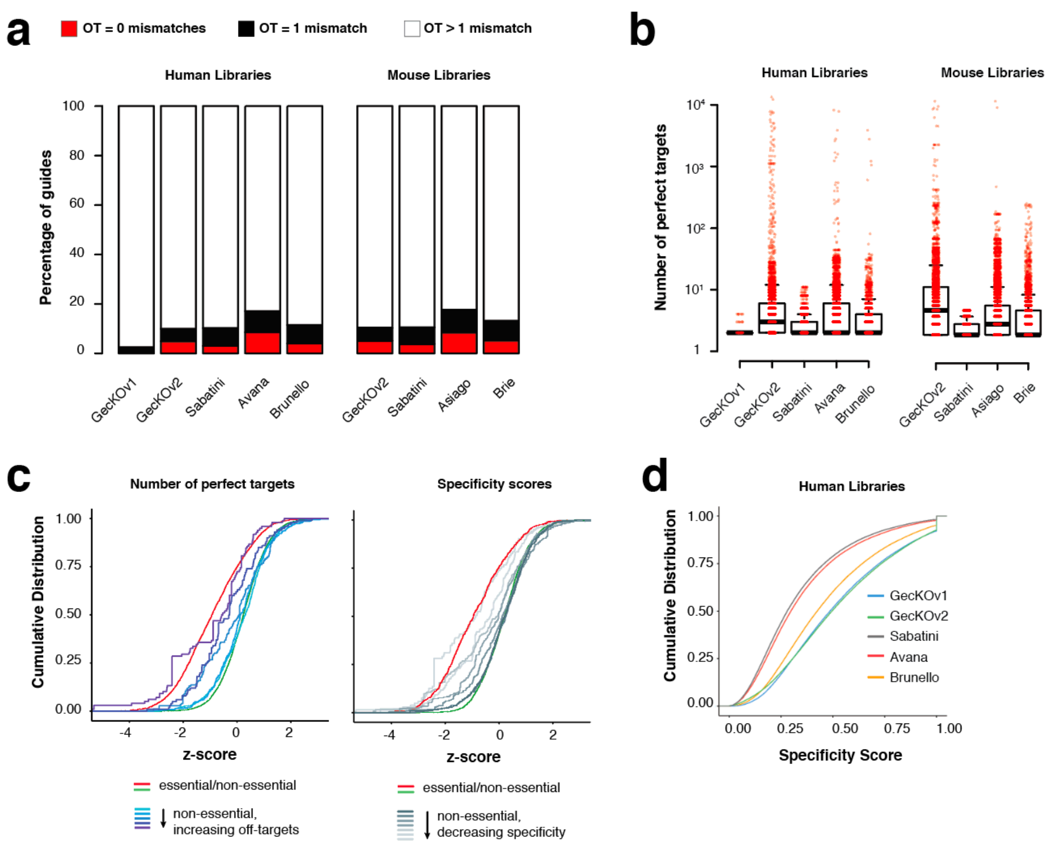


Figure 1. Pervasive off-targeting in Genome-wide CRISPR libraries. (a) Percentage of gRNAs in commonly used human and mouse genome-wide libraries whose nearest off-target (OT) alignment has zero (red), one (black), or more than one (white) mismatches to the guide. Mismatches are calculated as hamming distances. **(b)** number of perfect target sites for gRNAs that have off-targets at hamming distance 0. Each dot represents a unique gRNA. **(c)** Cumulative distributions of z-scores of gRNAs from all libraries during viability screens in A375 cells. Left, guides targeting non-essential genes were binned based on increasing number of perfect OTs (2, 3, 4, 5, >5, >10, >15; blue distributions). Distributions of gRNAs targeting essential (red) or non-essential (green) genes and with no off-target up to a hamming distance of 3 are plotted for comparison. Right, same as before but binning the gRNAs based on decreasing specificity scores (grey curves). Distributions of gRNAs targeting essential (red) or non-essential (green) genes and with specificity of 1 are plotted for comparison. **(d)** Cumulative distributions of specificity scores for the gRNAs in human libraries.

develop a CRISPR Specificity Correction (CSC) that disentangles the contribution of off-targeting from the effects of gene disruption to the depletion of a guide. We show that implementation of CSC reduces the number of false positives, increases the concordance between hits identified by different libraries, and demonstrate it uncovers both known and novel genetic dependencies in melanoma cells.

RESULTS

Characterizing the off-target space of genome-wide CRISPR libraries

We used GuideScan[11] to construct retrieval trees (tries) that index all possible Cas9-target sequences in the mouse (mm10) and human genomes (hg38) (see methods). We traversed these to retrieve all potential off-target loci of commonly used human and mouse libraries[3, 7, 8, 16, 17], allowing for up to 3 mismatches between gRNA and off-target. We found that a substantial fraction of guides in all libraries has multiple perfect or near-perfect sites in the genome (**Figure 1a**). For a single gRNA, perfect off-targets and single mismatch neighbors can occur tens-of-thousands of times (**Figure 1b** and **Supplementary Tables 1-9**). Furthermore, some of these promiscuous gRNAs are included multiple times in a single library, as previously described[18]. Finally, in each library, a substantial fraction of genes are targeted by gRNAs with multiple perfect target sites in the genome (**Supplementary Figure 1a**).

It is well established that gRNAs targeting amplified genomic regions can confound measurements of cell proliferation/viability in CRISPR loss-of-functions screens as a single gRNA can direct Cas9 to cleave multiple loci eliciting a DNA-damage response that includes cell cycle arrest[15]. We reasoned that the same could be true for highly promiscuous gRNAs present in genome-wide libraries. To assess this, we examined gRNA performance in dropout screens, which aim to identify genes whose functions are required for cell proliferation or viability[2, 7]. Guides that target such genes are expected to confer a selective disadvantage to cells and, as a consequence, be depleted from the cell population over time. These assays are particularly useful to benchmark the performance of CRISPR libraries and gRNA design rules because the availability of curated sets of “essential” and “non-essential” genes[19] allows the estimation of true positive and true negative hits.

As expected, highly-specific gRNAs (with no off-targets up to 2 mismatches) targeting essential genes were robustly depleted from the library over the course of two weeks (**Figure 1c**; red curves), while the representation of specific guides targeting non-essential genes remained roughly unchanged (**Figure 1c**; green curves). However, when we binned gRNAs targeting non-essential genes based on the number of perfect target sites in the genome (**Figure 1c**, left; blue curves), we saw a concomitant increase in the degree of their depletion until the distribution became similar to that of gRNAs targeting essential genes. The same effect was observed, though to a lesser extent, when gRNAs with a single perfect genomic target where binned based on increasing numbers of off-targets with one mismatch to the guide

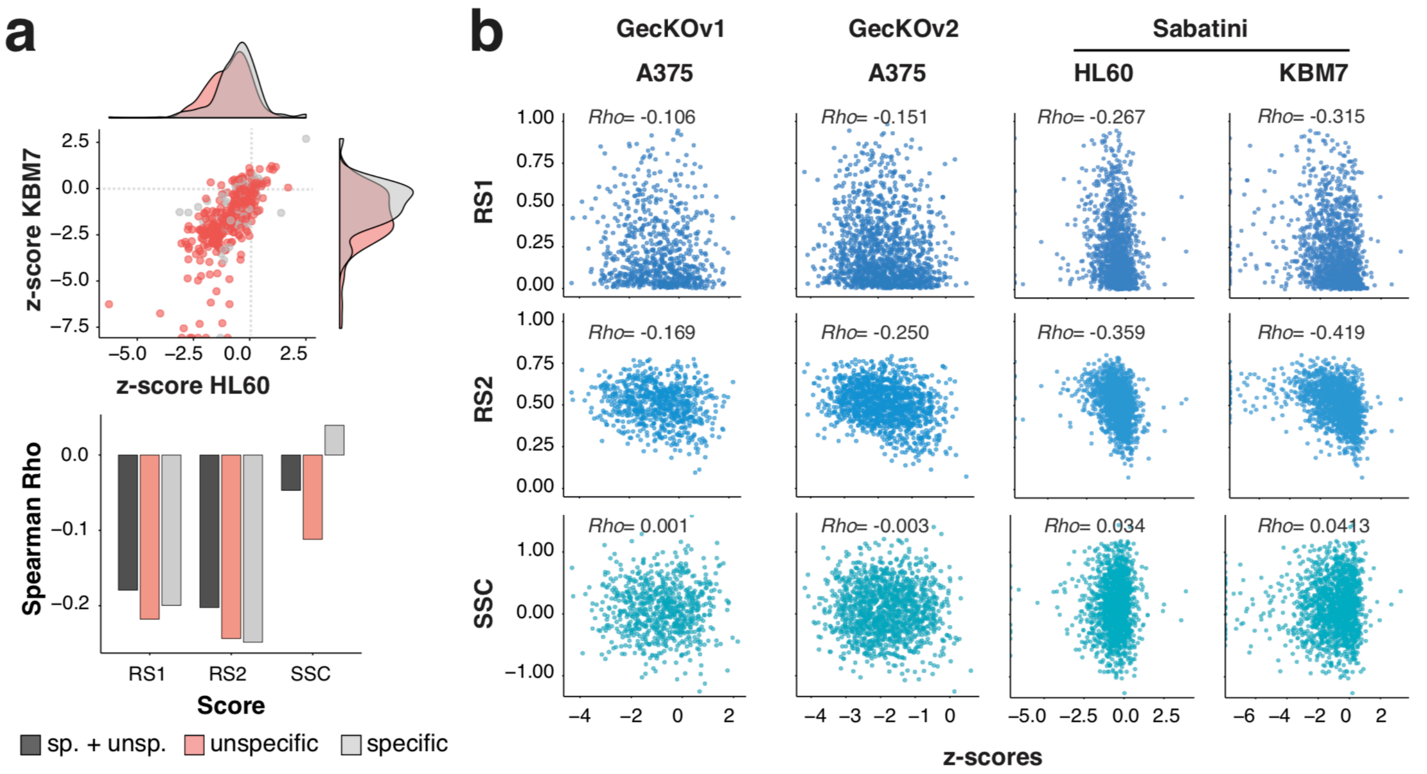


Figure 2. Influence of off-targeting on gRNA efficiency rules. (a) Top, dotplots and density plots showing the depletion of highly-specific (score = 1; grey) or highly-unspecific (score < 0.05; pink) gRNAs targeting essential genes in viability screens performed in KBM7 and HL60 cells. Bottom, spearman correlation coefficients between efficiency scores (RS1, RS2, and SSC) and depletion of specific gRNAs (pink), unspecific gRNAs (grey), or gRNAs from both groups combined (black). **(b)** Dotplots showing the correlation between the depletion of specific (score > 0.16) gRNAs targeting essential genes and RS1, RS2, and SSC scores in four distinct viability screens. Spearman correlation coefficients are shown.

(**not shown**), suggesting that, in agreement with previous reports[18, 20], a wide-range of off-targets can reduce cell fitness and confound growth measurements in CRISPR assays. To determine how the full range of off-targets contributed to gene-independent depletion, we calculated GuideScan specificity scores for all gRNAs[11]. This score ranges from 0-1 for targeting guides (with 1 denoting the most specific gRNAs) and takes into account the number, position, and type of mismatch between guide and genomic sequence[11] (**Figure 1d**, **Supplementary Figure 1b**). Binning gRNAs targeting non-essential genes based on this score showed that decreasing specificity led to increasing depletion of gRNAs (**Figure 1c**, right; **Supplementary Figure 1c**), analogous to what has been recently reported for gRNAs targeting regulatory elements[21]. Importantly, distributions of gRNAs with scores above 0.16 became indistinguishable from those of highly specific guides (specificity = 1; Kolmogorov–Smirnov test, adjusted for multiple testing), suggesting that above this specificity threshold off-targeting no longer interferes with gRNA representation in the library.

Together this data shows that off-targeting in CRISPR genome-wide libraries is more pervasive than previously anticipated. It also suggests that gRNAs with low specificity (GuideScan scores equal or below 0.16) may contribute to significant noise in data generated during genome-wide screens. In fact, for all libraries, a substantial fraction of genes is targeted by at least one (**Supplementary Figure 1d**), and often multiple (**Supplementary Figure 1e**, **Supplementary Figure 1f**), guides below our specificity threshold. Since promiscuous gRNAs are preferentially depleted during dropout assays, this may lead to the incorrect identification of genes as essential hits. In addition to contributing to false-positive hits, promiscuous gRNAs may also lead to the occurrence of false-negatives in the data by minimizing the signal-to-noise ratio. Finally, even when promiscuous gRNAs target essential genes (**Supplementary Figure 1f**) and their depletion correctly reflects genetic dependencies, off-targeting may still be problematic as it can serve as a confounder for gRNA design rules that have been learned or validated based on the degree to which gRNAs essential genes are lost from screens.

Impact of off-targeting on gRNA efficiency metrics

We first sought to determine if off-targeting affected previously published metrics of gRNA efficiency. We focused on three distinct scores that predict gRNA efficiency based on gRNA design rules defined through three distinct screening modes. Rule Set 1 (RS1) was developed based on a set of guides targeting cell surface markers, where loss of these markers was used as a proxy for gRNA activity[9]. Rule Set 2 (RS2) was developed based on enrichment of guides targeting genes whose mutation is known to confer resistance to a panel of drugs[7]. Finally, Spacer Scoring for CRISPR (SSC) was developed based on the depletion of known essential genes from dropout screens[22].

We computed RS1, RS2, and SSC scores for all guide RNAs in the Wang library[23], which does not incorporate any of these scoring metrics in its gRNA design, and used the depletion of gRNAs targeting essential genes as a surrogate measurement for gRNA activity. To test if off-targeting confounded any of the three scores, we selected gRNAs targeting essential genes and grouped them into highly-specific (specificity = 1) or highly-unspecific (specificity < 0.05) sets. As expected, the set of guides with the lowest specificity showed the strongest depletion in two independent cell lines (**Figure 2a,** top). We then calculated the correlation between all three efficiency scores and the depletion of gRNAs in each set individually or combined. Because guides that are more efficient at disrupting essential genes should have the strongest degree of depletion, we expected to see a negative correlation between depletion and each of the three efficiency metrics. While this was true for RS1 and RS2, both of which had moderate negative correlations for all groups tested (**Figure 2a**, bottom), SSC scores only correlated negatively with depletion when highly unspecific gRNAs were considered. This suggests that design rules extracted from gRNA depletion values are vulnerable to being confounded by off-target toxicity. In agreement with this, we observed mild negative correlations between RS1 and RS2 scores and the depletion of specific (specificity score > 0.16) gRNAs from two additional libraries (**Figure 2b**), but no evidence that SSC scores were able to predict gRNA activity in this setting. Overall, RS2 showed the best predictive value (**Figure 2b**, **Supplementary Figure 2**) and was therefore selected for follow up analysis.

Finally, we tested if gRNAs targeting functional pro-

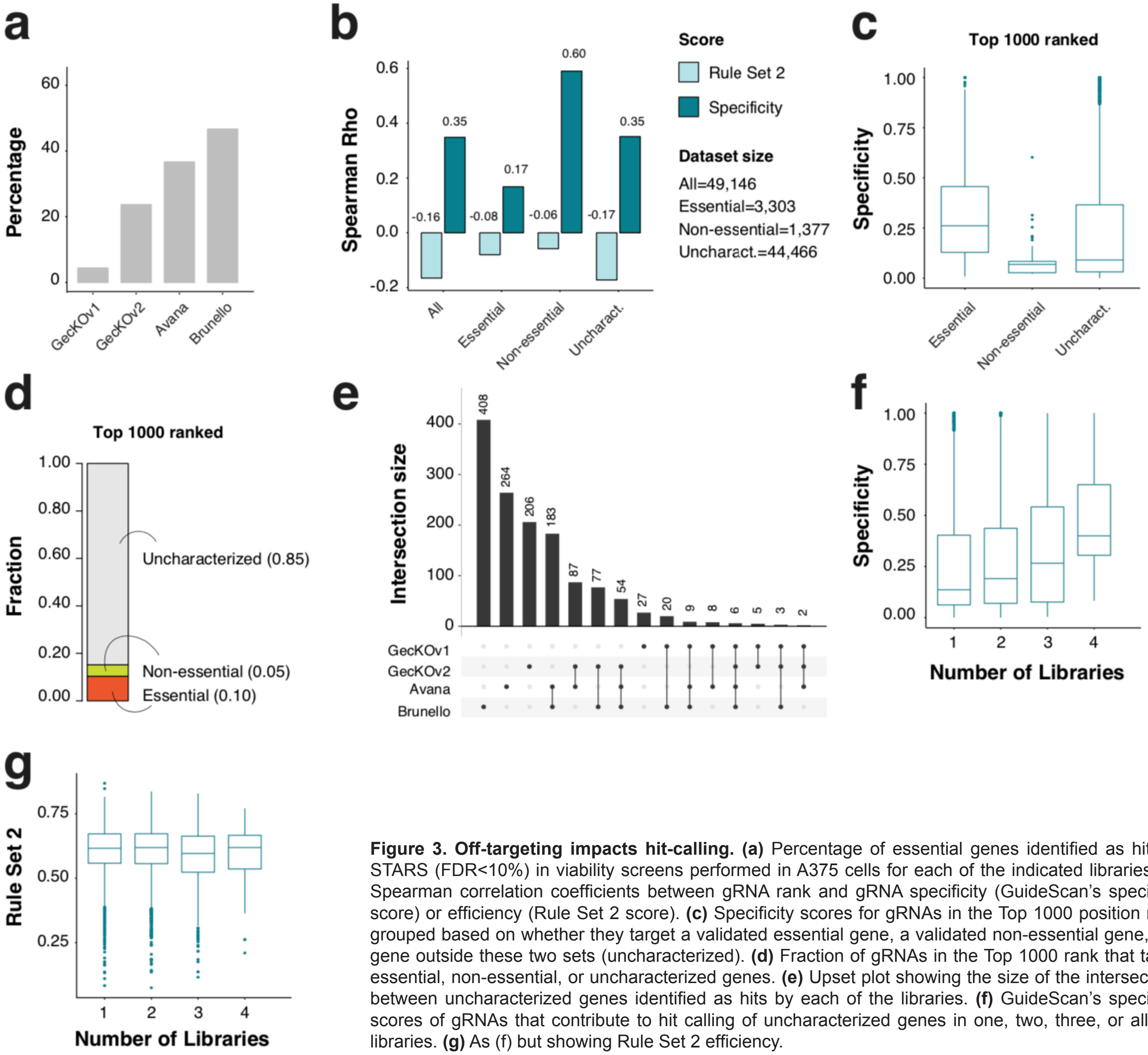


Figure 3. Off-targeting impacts hit-calling. (a) Percentage of essential genes identified as hits by STARS (FDR<10%) in viability screens performed in A375 cells for each of the indicated libraries. **(b)** Spearman correlation coefficients between gRNA rank and gRNA specificity (GuideScan's specificity score) or efficiency (Rule Set 2 score). **(c)** Specificity scores for gRNAs in the Top 1000 position rank, grouped based on whether they target a validated essential gene, a validated non-essential gene, or a gene outside these two sets (uncharacterized). **(d)** Fraction of gRNAs in the Top 1000 rank that target essential, non-essential, or uncharacterized genes. **(e)** Upset plot showing the size of the intersections between uncharacterized genes identified as hits by each of the libraries. **(f)** GuideScan's specificity scores of gRNAs that contribute to hit calling of uncharacterized genes in one, two, three, or all four libraries. **(g)** As (f) but showing Rule Set 2 efficiency.

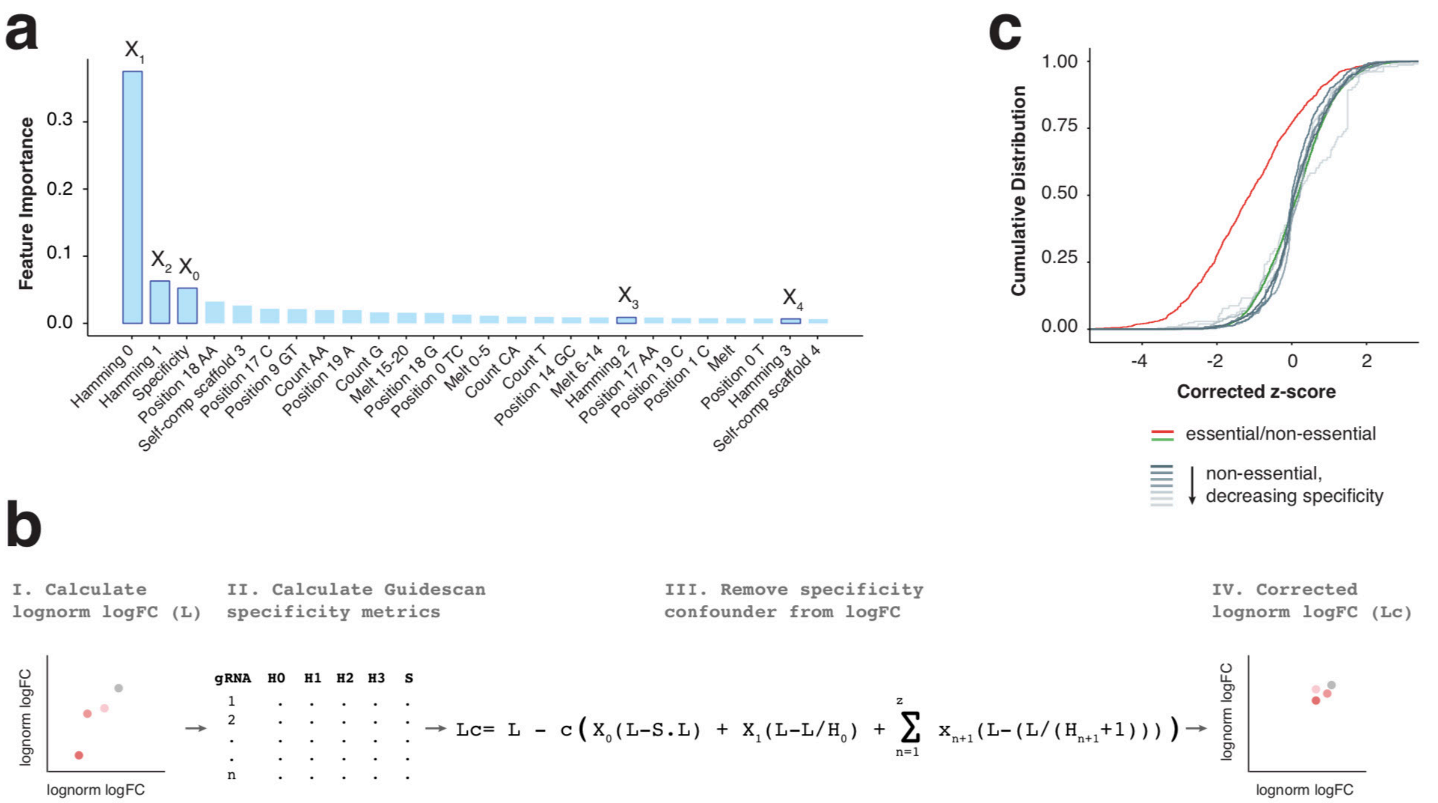


Figure 4. A CRISPR Specificity Correction (CSC) to minimize noise in CRISPR screens. (a) Importance of the top 26 features for the boosted gradient regression tree model. The five features used in CSC are highlighted. **(b)** Schematic representation of CSC. **(c)** Cumulative distributions of z-scores of gRNAs from GecKO1, GecKO2, Avana, and Brunello libraries during viability screens in A375 cells following CSC implementation. Distributions are plotted as in Figure 1c.

tein domains are preferentially depleted from dropout assays[4] by virtue of being more promiscuous—since sequences of functional domains are under strong evolutionary constraints and are often shared by multiple genes. We found that gRNAs used to define this rule[4] where highly specific, with no evidence that their dramatic depletion was driven by off-targeting (**Supplementary Figure 2b**, **Supplementary Figure 2c**), lending further support to the idea that targeting functional domains is a useful strategy to generate true loss-of-function alleles.

Off-targeting affects library performance and hit-call reproducibility

Next, to determine if off-targeting affects hit-calling, we used the STARS algorithm[7] to identify essential hits from high-throughput screens performed under identical conditions using GecKOv1, GecKOv2, Avana, and Brunello libraries[7,2]. The Brunello library identified the highest percentage of genes from a curated set of core essentials (n=291) at FDR<10% (46.7%), followed by Avana (36.7%), GecKOv2 (23.7%), and GecKOv1 (4.4%) (**Figure 3a**).

STARS, like other hit-calling algorithms[24, 25], is a gRNA-ranking system and rewards genes for which multiple guides score. We reasoned that both gRNA efficiency and gRNA specificity could influence the position of a guide within this ranking and thus help explain the differences between the hits identified by each of the libraries, or the inability to identify the majority of core essential genes as hits. To examine this, we pooled the rankings of all four libraries and examined if the gRNA's position in these lists correlated best with its efficiency (as determined by RS2 scores) or specificity (as determined by GuideScan's score). We looked at all guides combined or grouped by the targeting of essential, non-essential, or uncharacterized genes (i.e. genes outside the two pre-validated sets). In all cases, the specificity score showed the strongest correlation with guide rank (**Figure 3b**). This was most noticeable for gRNAs targeting non-essential and uncharacterized genes. In fact, within the top 1000 rank, guides from these two sets tended to have the lowest specificities (**Figure 3c**), suggesting their ranking was largely driven by off-targeting. This is problematic because together, gRNAs targeting uncharacterized and non-essential genes occupy 90% of top rank positions during hit calling (**Figure 3d**), which may relegate gRNAs targeting true essential genes to lower positions.

Because gRNAs targeting uncharacterized genes tend to have relatively low specificities (**Figure 3c**) which correlate with rank (**Figure 3b**) we examined the genes from this set that were identified as essential by each of the four libraries. Again, Brunello returned the highest number of hits, followed by Avana, GecKOv2, and GecKOv1 (**Figure 3e**). Despite the indication that these genes play essential roles in A375 cells, the majority of the hits did not overlap amongst libraries (**Figure 3e**). To see if this lack of hit-calling reproducibility is driven by differences in specificity or gRNA efficiency we retrieved for each library all gRNAs that lead uncharacterized genes to be identified as hits, and grouped them based on how many libraries had identified the gene they target as a hit (**Figure 3f**, **3g**). We reasoned that if the discrepancies in hit calling are caused by off-targeting, this should be reflected in the specificity of the gRNAs in these groups. Indeed, we found that gRNAs that lead to genes being identified as hits in a single library had the lowest specificity (**Figure 3e**) and that the specificity scores increased with the number of libraries that reproduced the hit-call (**Figure 3e**). In contrast, the RS2 scores were identical for all groups regardless of the agreement amongst libraries (**Figure 3g**).

Based on these results, we conclude that off-targeting causes substantial noise in the data generated by genome-wide libraries. This includes the high ranking of promiscuous gRNAs during hit-calling which compromises the identification of true positives hits and the reproducibility of the data generated by independent libraries.

A computational correction for off-target mediated gRNA depletion

Our data suggests that CRISPR loss-of-function screens should include only gRNAs with a specificity score above 0.16 in their pools to prevent toxicity caused by off-targeting. Yet, current CRISPR libraries have not been designed using this specificity threshold, and as a consequence generate data with low signal-to-noise ratios. In an effort to minimize these issues, we set out to decouple the gene-knockout effect from the off-target effect by developing a CRISPR Specificity Correction (or CSC) that adjusts for the contribution of off-target cleavage to the total depletion of the gRNA.

GuideScan computes specificity metrics that include the enumeration of potential off-targets up to a Hamming distance of 3, as well as the gRNA's specificity score. These five metrics, along with the sequence features used in RS2[7], and novel features quantifying the gRNA's self-complementarity (including complementarity to the scaffold sequence) were computed for all gRNAs in the Avana library[7, 8]. Overall these 426 engineered features (**Supplementary Table 10**) were used as covariates in the regression models. Previous studies attempted to learn cutting efficiency rules based on the performance of gRNAs targeting a subset of genes with known functional relevance[7, 9, 22]. In contrast, our model aims to predict gRNA depletion from the totality of gRNAs in the library, under the assumption that most of genes do not play a functional role in cell viability or proliferation and therefore depletion of the majority of guide RNAs will reflect the influence of off-targeting.

We trained three regression models (linear regression, random forest regression, and gradient-boosted regression tree) on 90% of the guides and held out the remaining 10% for testing. Gradient-boosted regression trees performed the best at predicting gRNA depletion as assessed by mean squared error. All five specificity metrics computed by GuideScan were within the top 25 most important features in predicting $\log_2$ fold depletion (**Figure 4a**, **Supplementary Table 10**) representing 50.6% of the overall feature importance. This supports our previous analysis that off-target toxicity is a major driver of gRNA depletion in CRISPR screens. The remaining contribution came from features that impact the activity of the gRNA including those described in RS2 (**Figure 4a**). Importantly, self-complementarity of different lengths were amongst the most important features for the model (4%) outside the specificity metrics. Self-complementarity with the scaffold region and complement segment of the gRNA are both expected to affect gRNA activity by interfering with its association with the target. Therefore, the identification of these novel features suggests that current sequence specific cutting efficiency rules can be further enhanced.

The five specificity features of the model represent confounding variables in CRISPR screens as they contribute to gene-independent gRNA depletion. CSC adjusts for this to generate a corrected $\log_2$ fold-change (Lc) for each guide in the library using the following formula (**Figure 4b**):

$$L - c(x_0(L - sL) + x_1\left(L - \frac{L}{h_0}\right) + \Sigma_{n=1}^{z} x_{n+1}\left(L - \frac{L}{h_n+1}\right))$$

Where (L) is the original $\log_2$ fold-change of the gRNA, (H0, H1, H2, H3, and s) represent the five specificity metrics computed by GuideScan (i.e., occurrences at Hamming distance 0, 1, 2, and 3, and specificity score), and (X0 to X4) represent the numeric values of the relative feature importance learned for each of the specificity metrics **(Figure 4a**). The cumulative depletion effects are scaled by a computationally derived dynamic coefficient (c), ranging from 0 to 10, to capture and minimize the influence of experimentally derived batch effects (see Methods).

We applied CSC to GecKOv1, GecKOv2, Avana, and Brunello datasets derived from viability screens in A375 melanoma cells[2, 7], and plotted the resulting depletion values for specific gRNAs targeting essential and non-essential genes, as well as gRNAs binned based on decreasing specificity. We found that CSC successfully removed the influence of off-targeting from the depletion of gRNAs targeting non-essential genes, bringing all distributions together regardless of gRNA specificity (compare **Figure 4c**, **Supplementary Figure 3a** with Figure 1d, Supplementary Figure 1b). Of note, this was accomplished without compromising the signal of gRNAs targeting essential genes (**Figure 4c**, **Supplementary Figure 3a**; red distribution).

In summary, our data suggests that off-targeting causes gRNAs to be inappropriately lost from libraries and that this confounder can be corrected for using CSC.

Testing CSC performance

To benchmark CSC, we applied STARS to the corrected $\log_2$ fold-change of all four libraries and examined how that affected the identification of gold-standard essential genes as hits (FDR<10%). When comparing the number of true-positives that scored, we found that CSC adjustment uncovered validated essential genes that were previously missed (**Figure 5a**). However, large numbers of essential genes that were

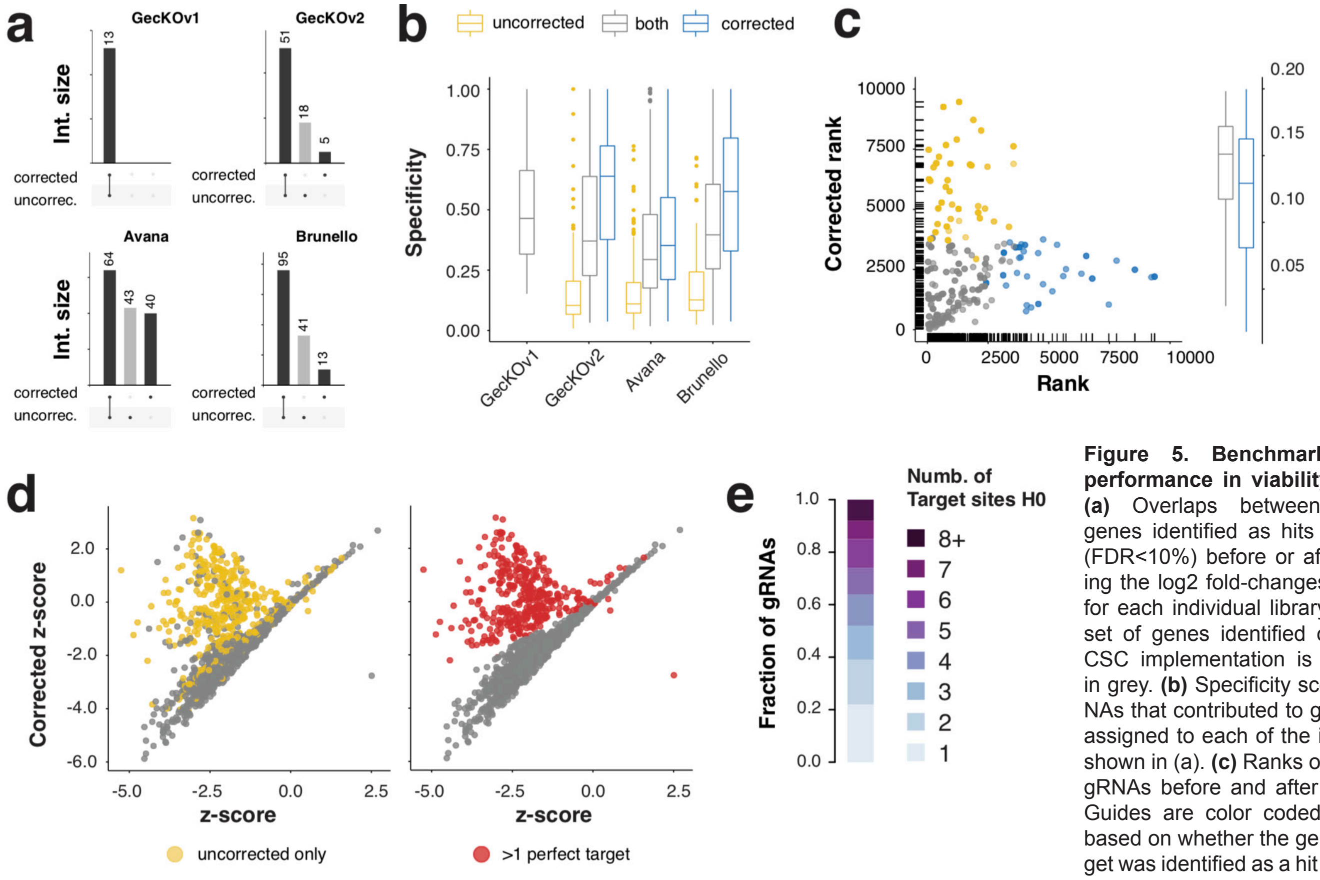


Figure 5. Benchmarking CSC performance in viability screens. (a) Overlaps between essential genes identified as hits by STARS (FDR<10%) before or after correcting the log2 fold-changes with CSC for each individual library. The subset of genes identified only before CSC implementation is highlighted in grey. **(b)** Specificity scores of gRNAs that contributed to genes being assigned to each of the interactions shown in (a). **(c)** Ranks of unspecific gRNAs before and after correction. Guides are color coded as in (b), based on whether the gene they target was identified as a hit only before CSC (yellow), only after CSC (blue), or in both (grey). Boxplots on the right show the specificity scores for gRNAs that lead to hit calling after CSC. **(d)** Z-scores of gRNAs targeting essential genes before and after CSC implementation, where gRNAs are labeled based on whether they led to the identification of essential genes as hits only before CSC (left, yellow), or whether they have more than one perfect target (right, red). **(e)** Fraction of gRNAs that identify essential genes only before CSC that have only 1 (light blue) or multiple (2-8+, graded colors) perfect target sites in the genome. H0, hamming 0.

previously identified as hits failed to score after correcting the $\log_2$ fold-change with CSC (**Figure 5a**).

To investigate the reasons behind this observation we retrieved all gRNA sequences targeting the validated set of essential genes and, for each library, grouped them based on whether their target had scored as essential only before the correction, only after the correction, or in both cases. We found that essential genes identified as hits only before correction were targeted by the most unspecific gRNAs. In contrast, hits identified only after CSC implementation were on average the most specific (**Figure 5b**). The observation that essential genes uncovered as hits only after CSC are targeted by the most specific gRNAs lends further support to the notion that, before correction, these gRNAs were outcompeted by promiscuous gRNAs (see **Figure 3**). However, the inability to identify essential genes targeted by highly unspecific gRNAs could suggest that CSC cannot retrieve information from these gRNAs and its implementation is theoretically equivalent to discarding all gRNAs with low specificity scores from hit-calling analysis. To determine if this was the case, we selected gRNAs targeting essential genes with low specificity and plotted their ranks before and after CSC correction (**Figure 5c**). We reasoned that if CSC was not able to retrieve information from these guides, then they would fail to rank after their $\log_2$ fold-change had been corrected. Instead, we found that even after correction with CSC, highly unspecific gRNAs occupied some of the highest ranks of STARS (**Figure 5c**). In fact, ranking gRNAs could have specificities scores as low as 0.018. Thus, CSC does not simply discard gRNAs with low specificity.

To further explore what features led known essential genes to be lost after CSC application, we analyzed how the depletion values of gRNAs changed before and after its implementation (**Figure 5d**). We found that for the majority of gRNAs, z-scores were similar before and after correction. Yet, a distinct subset of gRNAs had z-scores that substantially changed with CSC (**Figure 5d**). The majority of these gRNAs targeted essential genes identified as hits only before CSC (**Figure 5d**, left panel). In addition, this uncorrelated population was comprised exclusively of gRNAs with more than one perfect target site in the genome (**Figure 5d**, right panel). This suggests that although CSC is able to retrieve depletion information from highly unspecific gRNAs, it is generally unable to assign function to guides that target multiple identical sites. Indeed, about 80% of all gRNAs targeting essential genes that did not score after CSC had more than one perfect target site in the human genome (**Figure 5e**). It is worth noting however, that a small fraction of gRNAs with multiple perfect targets still ranked highly even after CSC implementation (**not shown**). Thus, for genes whose dis-

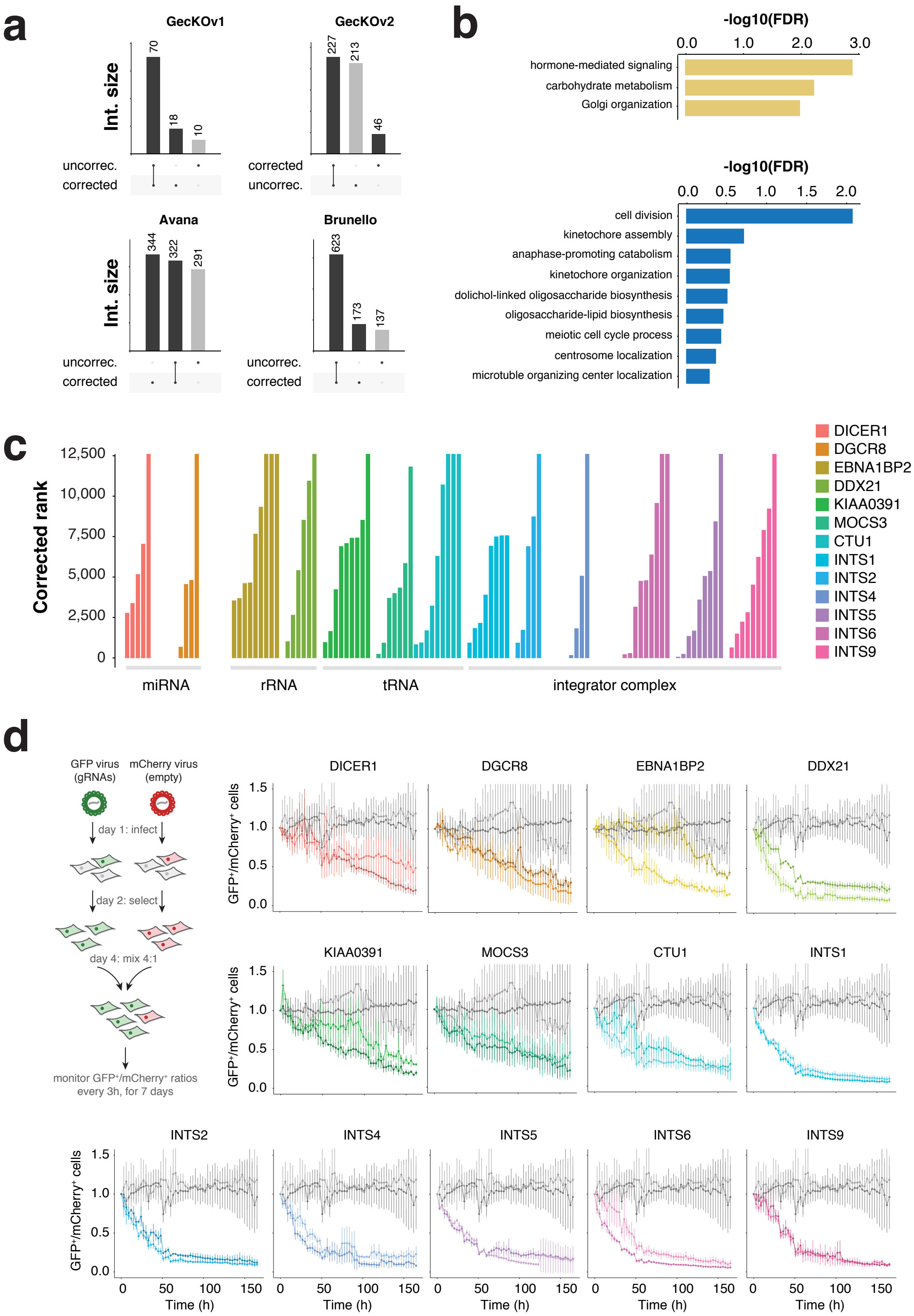


Figure 6. CSC uncovers new genetic dependencies in melanoma cells. (a) Overlaps between uncharacterized genes identified as hits by STARS (FDR<10%) before or after correcting the log2 fold-changes with CSC for each individual library. The subset of genes identified only before CSC implementation is highlighted in grey. **(b)** Gene Ontology (GO) analysis focusing on uncharacterized genes that were identified as essential hits only before (top, yellow) or only after (bottom, blue) CSC correction. **(c)** Guide RNA ranks for subset of genes involved in RNA metabolic pathways that are identified as essential after correction of off-targeting with CSC. **(d)** Growth competition assays in A375 cells for selected genes. A schematic representation of the experiment is shown on the left. Plots show ratios of GFP+/mCherry+ cells over the course of a week, normalized to the first time point. Each gene was targeted by two independent gRNAs. Control gRNAs are shown in grey and plotted in each graph for comparison. Curves show mean and standard deviation between three replicate experiments.

ruption is very deleterious to the cell, CSC may still be able to decouple off-targeting effects from gene-knockout effects even from gRNAs with multiple identical target sites in the genome.

We draw two main conclusions from this analysis. First, that even when essential genes are correctly identified as hits in high-throughput CRISPR screens that may be driven in large part by off-targeting (**Figure 5d**). This may help explain why efficiency scores learned on the depletion of essential genes perform worse than others (**Figure 2a**, **2b**). Second, because unspecific gRNAs are preferentially depleted from dropout screens and tend to occupy the top ranks of hit calling algorithms (**Figure 4**) they often mask true positive hits. By adjusting for the contribution of off-targeting to the total depletion of the gRNA, CSC allows known essential genes that are targeted by specific gRNAs (and as a consequence generally at a disadvantage compared to unspecific gRNAs) to be identified as hits (**Figure 5b**). This suggests that the same approach may be able to uncover unknown genetics dependencies from high-throughput CRISPR data.

CSC uncovers genetic dependencies in melanoma cells

We next turned our attention to the set of uncharacterized genes and asked how CSC impacted the identification of hits by STARS. Because the analysis above suggested that CSC is able to uncover true positive hits while minimizing the occurrence of false negatives, we first checked if this also led to better reproducibility of hits between libraries. Indeed, CSC led to a higher hit overlap among hits identified by all four libraries for this gene set (compare **Supplementary Figure 3b** with Figure 3e), with a significantly higher number of genes scoring in multiple libraries (**Supplementary Figure 3c**).

As before, a subset of genes scored only after CSC implementation (**Figure 6a**). To understand if these reflected real genetic requirements for cell proliferation and/or viability we performed a Gene Ontology analysis with two independent tools[26-28]. In both cases, we found that genes identified after $\log_2$ fold-change correction by CSC were enriched in terms related to cell division and chromosome segregation (**Figure 6b**, **Supplementary Table 11**), which are expected to score in dropout assays. In contrast, genes identified as hits only before correction were enriched for terms related to hormone signaling, carbohydrate metabolism, and Golgi organization (**Figure 6b**), and were driven by highly promiscuous gRNAs targeting multiple members of the same protein family (**Supplementary Figure 3d**). Although we cannot exclude that these genes play essential roles in A375 cells, this data strongly suggests that in the context of current genome-wide libraries their depletion is largely driven by off-targeting.

Aside from terms related to cell division and chromosome segregation, the majority of Gene Ontology terms enriched amongst new hits were related to metabolism and processing of RNA (**Supplementary Table 11**). We found numerous hits involved in small non-coding RNA processing including genes required for the biogenesis of miRNAs (DICER, DGCR8) rRNAs (EBNA1BP2, DDX21) and tRNAs (KIAA0391, MOCS3, CTU1) (**Figure 6c**, **Supplementary Table 11**). In addition, multiple components of the Integrator complex (INTS1, INTS2, INTS4, INTS5, INTS6, INTS9) scored as essential after CSC correction (**Figure 6c**). Of these, DICER, DGCR8, and CTU1 have been previously implicated in tumor cell growth, highlighting CSC's ability to uncover true essential genes as hits. To test if these and the remaining candidate hits represent genetic dependencies in A375 melanoma cells, we selected two gRNAs for each gene and measured the impact of these guides on cellular growth using competition assays (**Figure 6d**; left). As expected, control gRNAs targeting "safe-targeting" regions[20] (grey lines) showed no evidence of depletion (**Figure 6d**; right). In contrast, cells expressing gRNAs against our positive controls or our candidate hits (colored lines) were outcompeted by wild-type cells over the course of our experiment, suggesting that disruption of the genes they target impairs cellular proliferation or viability.

Together these observations highlight the essentiality of RNA metabolic pathways to melanoma cell growth and demonstrate that true genetic dependencies such as these can be systematically uncovered by CSC. To facilitate the use of CSC as a component of the analysis workflow of CRISPR screens, we are making the software available to the community via our Bitbucket repository, along with all the data produced in this study.

DISCUSSION

CRISPR high-throughput functional assays rely on the principle that, within a population of cells infected with a lentiviral library, the abundance of individual gRNA sequences reflects the importance of their targets to the biological process being studied. One of the most common modalities are negative selection screens, where gRNAs targeting genes essential for cell growth are expected to specifically dropout from the population over the course of the experiment. These assays are extremely powerful at defining genetic dependencies.

Yet, it is well appreciated that the abundance of gRNAs in these assays does not depend solely on the function of the gene they target. Guide RNA efficiencies are a known confounder in CRISPR screens since gRNAs with low activity are unlikely to score even if they target an essential gene. Similarly, unintended cleavage can contribute to the erroneous depletion of gRNA, particularly through toxicity caused by the generation of multiple double stranded DNA breaks at amplified genomic regions[15, 29, 30]. Both gRNA efficiency and toxicity therefore influence the effectiveness of libraries as gene discovery tools. The extent to which off-targeting outside highly amplified genomic regions affects CRISPR library performance is less well characterized[20, 31, 32]. While toxicity caused by off-target cleavage has been documented in large-scale assays[20], studies suggest it may be generally small, with CRISPR technology producing few systematic off-targets effects[31, 32] and gRNA abundance reflecting predominantly on-target activity[32].

Because our previous work[11] demonstrated that short-read aligners do not accurately enumerate potential off-target sites for gRNAs leading to an underestimation of their promiscuity we set out to revisit the impact of off-target effects on large-scale CRISPR screens. We show that both first- and second-generation libraries are affected by extensive off-targeting which decreases their performance in negative selection screens. Decreased performance stems from at least two phenomena. First, genes can score as hits solely by virtue of being targeted by multiple promiscuous gRNAs. Second, promiscuous guides can be ranked highly by hit-calling algorithms effectively outcompeting specific gRNAs against essential genes. Thus, gRNA off-targeting contributes to both the occurrence of false-positive and false-negative hits in loss-of-function negative selection screens. Both have clear implications to biomedical research as they increase the efforts required for secondary validation of identified hits and contribute to low replication of data when performing identical screens with independent tools.

We also show that gRNA promiscuity can be quantified—and gene-independent depletion predicted—using GuideScan's specificity scores[11], an aggregate metric that takes into consideration the number and type of off-targets for each guide. Using this score, we define a specificity threshold (0.16) above which depletion of gRNAs due to off-target toxicity is no longer detected. The incorporation of this rule in newly designed CRISPR libraries should minimize the noise generated by off-target cleavage and yield libraries with increased sensitivity. Yet, we foresee that implementation of this rule will not always be possible. First, large discovery efforts have already been deployed using current tools and it is unlikely that they will be replicated with improved libraries in the near future. Second, there is an increasing interest in using CRISPR to identify essential noncoding regulatory sequences in large scale. While GuideScan scores also predict gene-independent depletion of gRNAs in this setting[21], many of these elements—such as transcription factor binding sites or RNA Binding Protein motifs—are so small in size that only a limited number of gRNAs that can potentially disrupt them, making further filtering unachievable.

To deal with these constraints we developed CSC to adjust for the contribution of off-targeting in gRNA depletion. We use CSC in four dropout screens performed under identical conditions and show that this correction can remove hits whose identification is driven by gRNA promiscuity. CSC also uncovers gold-standard essential genes targeted by specific guides and improves the concordance of hits between all four independent libraries. Finally, CSC uncovers numerous genes involved in RNA metabolism as genetic vulnerabilities in melanoma cells, which we validate experimentally using CRISPR-Cas9–based cell competition assays.

DICER and DGCR8 have been previously implicated in tumorigenesis where they have been shown to act as haplo-insufficient tumor suppressors[33-35]. Our results support the notion that even if compromised gene regulation by the miRNA pathway may be advantageous to tumor cells, its complete disruption is detrimental to tumor cell growth. EBNA1BP2 is a conserved protein required for pre-rRNA processing and ribosome assembly in yeast3[6, 37], whose depletion in Saccharomyces cerevisiae, leads to an arrest in cell division under restrictive conditions[36]. The same requirement seems to also exist in melanoma cells. CTU1 is a subunit of the cytosolic thiouridylases complex, involved in wobble position post-transcriptional modifications[38] which optimizes codon usage during gene-specific translation[38]. This activity has recently been shown to be critical for cells carrying oncogenic BRAF^{V600E} [39]. Finally, Integrator was initially described as important for the processing of snRNA 3'-ends[40, 41], but has since been implicated in the biogenesis of other RNA molecules including enhancer RNAs[42, 43] and messenger RNAs[43-45]. At mRNAs, Integrator subunits seem to play various roles including the processing of replication-dependent histones (INTS3, INTS9)[45], the initiation of transcription downstream of MAPK signaling (INTS1, INTS11)[43, 44] and the release of paused polymerase II (pol II) from the promoters of growth factor responsive genes (INTS1, INTS11)[44]. More recent reports suggest that integrator may also destabilize the association of pol II with promoters preventing productive elongation at a subset of genes (INTS1, INTS4, INTS9, INTS11, INTS12)[46, 47]. The scoring of multiple Integrator subunits in our analysis suggests that one or several of these functions are essential for the viability/proliferation of melanoma cells. Compromised snRNA biogenesis is perhaps one of the obvious explanations for impaired cell growth following loss of Integrator since it affects essential processes such as splicing. However, we find that subunits not required for snRNA processing in Drosophila[41, 46] also score well following CSC implementation, suggesting that Integrator functions at protein-coding genes may also be essential. Our data lends further supports to the idea that stimulus-dependent recruitment of Integrator to MAPK-responsive genes is required for the growth of cells with activating mutations in BRAF[43].

Together, these data suggest that CSC is an effective strategy to maximize data recovery from essentiality screens performed with published high-throughput libraries. We predict that CSC implementation will also further enable the use of high-throughput CRISPR screens against small regulatory sequences, by allowing the use of gRNAs with low specificity in cases filtering for gRNA specificity is not possible. Finally, while our current study is limited to Cas9-based screens, CRISPRi and CRISPRa assays can also be confounded by off-targets[21]. Therefore, we predict CSC will also be useful to uncover genetic dependencies in those contexts.

In summary, we characterize off-targeting in high-throughput CRISPR screens and develop a computational strategy to minimize the noise it creates. We expect this method will prove useful to maximize data recovery from screens targeting both the coding and non-coding genomes.

MATERIALS & METHODS

Enumeration of Targets

We constructed retrieval trees (tries) consisting of all possible 20mer Cas9 gRNA target sites in the mouse and human genomes as previously published GuideScan[11]. Unlike the original tries, these were constructed without alternative chromosome data and thus produce a more accurate description of the off-target space of individual guide RNAs. To determine the mismatch neighborhood for each gRNA in the library, we traversed each of their sequences through the trie to exhaustively determine all neighbors up to and including Hamming and Levenshtein distances of 3. Specificity scores for each gRNA was computed using Hamming distance neighbors as previously described[11].

Model Development for Predicting log2 fold-change

Data

A total of 95,344 gRNAs from the Avana library cloned into the lentiGuide vector[7, 8] were selected for model development. Guide RNAs raw read counts were converted into logarithmic normalized (lognorm) counts and technical replicates were combined and averaged. Prior to computing $\log_2$ fold-change, counts between averaged control and averaged gRNA knock down were also logarithmically normalized. The resulting lognorm counts between control and knock down conditions were then subtracted to get $\log_2$ fold-change. This $\log_2$ fold-change was the predicted feature during model development.

Feature Engineering

A total of 426 features were used in the learning of the model to predict Avana gRNA $\log_2$ fold-change. All features from Rule Set 2 were generated and incorporated into the model. Pertinently, strings were represented in the model by one-hot encoding of position-dependent 1mers and 2mers as well as position-independent 1mers and 2mers. GuideScan gRNA specificity features (specificity score, and Hamming distances at 0,1,2,3) were represented as numeric values and were taken directly from trie-based mismatch neighborhood enumerations and specificity computations. Self-complementarity computations of the gRNA complementary sequence with itself and with the tracrRNA scaffold were done by concatenating the complementary sequence with the scaffold sequence to generate an aggregate string. The aggregate string was then divided into substrings of length k (where k is 3, 4, or 5). The k-string and the reverse complement of the k-string were then placed into a hash function to determine a hash slot for two hash tables. If the value of the forward k-string was greater than the value of the reverse k-string then the k-string was assigned to the first hash table; otherwise the k-string was assigned to the second hash table. The value of the k-string in each hash slot equaled the occurrence of the k-string in the aggregate string. An inner product between the two hash tables was computed for each string at each k value to determine self-complementarity between the complementary sequence and the tracrRNA scaffold. Self-complementarity of the complementary sequence with itself was done in the same manner except the aggregate string is simply the complementary sequence.

Model Selection and Training

Regression models were selected to predict the value of $\log_2$ fold-change from gRNA features. Specifically, regression models where the importance of each feature's contribution to the prediction could be trivially extracted and interpreted were utilized. Linear regression, Random Forest Regression, and Boosted Gradient Regression Tree models were selected for this regression task. Hyperparameters were tuned through grid search with fivefold cross validation to optimize the models prior to prediction on test data. Training was done on 90% of total data and testing was done on 10% of held out total data. The data was randomly divided into training and testing sets. Models never encountered test data during training. Accuracy of the models were assessed by mean squared error between predicted values and test values. Error residuals were computed for each model to assess for systemic learned error and no pattern was appreciated. The gradient boosted regression tree model had the lowest mean squared error and its features importance were extracted for use in constructing a depletion screen correction.

CRISPR Specificity Correction (CSC)

CSC Derivation

A mathematical correction to account for the influence of non-specificity on screen log2 fold-change data was derived using the formula below.

$$L - c(x_0(L - sL) + x_1\left(L - \frac{L}{h_0}\right) + \sum_{n=1}^{z} x_{n+1}\left(L - \frac{L}{h_n+1}\right))$$

Where:

L = log2 fold-change
c = dynamic coefficient
s = GuideScan specificity score
x_0 = learned value coefficient for specificity score
x_1 = learned value coefficient for Hamming distance 0
x_n = learned value coefficient for Hamming distance n+1
h_0 = number of occurrences at Hamming distance 0
h_n = number of occurrences at Hamming distance n

Where the $\log_2$ fold-change is the unmodified $\log_2$ fold-change from any depletion screen data. The occurrence of target sites at arbitrary Hamming distances is their enumerated occurrences in a genome as determined by traversal through a trie of target sites. The learned value coefficients were determined by extracting the feature importance values from the gradient boosted regression tree model. The aggregate of all features importance in this model add to one and the feature values used as coefficients in the correction represent confounders for on target $\log_2$ fold-change prediction. Explicitly, the value coefficients were learned from the gradient boosted regression tree, while the values of a gRNA's specificity score and Hamming neighborhood were directly computed. The specificity score was computed using the GuideScan specificity score detailed below:

$$GuideScan\ Specificity\ Score = s = \frac{1}{\sum_{i=1}^{n} CFD_i * q_i}$$

Where:

CFD = likelihood of a gRNA cutting at the ith neighbor
qi = number of times the ith neighbor occurs in the genome

Notably for a unique target site up to z mismatches, the GuideScan Specificity Score would be 1 since CFD = i = n = qi = 1. The specificity score applies only to Cas9 gRNAs. In its special form the CSC only corrects for off-targeting of Cas9 NGG PAM gRNAs.

Dynamic Coefficient
The coefficient that scales the correction is computed for each screen and ranges from 1 to 10. For each value of the coefficient, a two-sided Kolmogorov-Smirnov test is computed between a set of a priori defined ground-truth gene $\log_2$ fold-changes and all other gene $\log_2$ fold-changes in a screen. The p-value for each iteration is computed and the value of the coefficient that produces the largest p-value is taken as the optimal coefficient value for the screen. The assumption behind the dynamic coefficient is that the aggregate distribution of $\log_2$ fold-changes in a screen should not deviate significantly from the distribution of ground truth gene $\log_2$ fold-changes.

General Form of CSC
CSC in its special form is specific to Cas9 NGG PAM gRNAs. However, it is written such that it can be generalized to any CRISPR system if one sets the value of x_0 to 0. Additionally, if a user does not have ground truth gene $\log_2$-fold changes to compare against they may set the value of c to 1. The most general form of the correction is therefore:

$$L - (x_1\left(L - \frac{L}{h_0}\right) + \sum_{n=1}^{z} x_{n+1}\left(L - \frac{L}{h_n+1}\right))$$

Cutting Efficiency Computations

Cutting efficiency for gRNAs were computed through command line versions of Rule Set 1, Rule Set 2, and SSC. All cutting efficiency metrics required 30mer sequences for gRNAs. The expanded 30mer sequence for each gRNA was determined through GuideScan derived gRNA target sequence to coordinate listed hash table lookup. In this manner all coordinates tied to a gRNA occurrence could be accessed in constant time. Coordinates for screen gRNAs were determined by hash table lookup and the coordinate was expanded to a 30mer length. The resulting coordinates were then used to determine the 30mer sequence from an indexed fasta file. The resulting 30mer sequence was used to compute cutting efficiency scores using Rule Set 1, Rule Set 2, and the SSC scorers.

Rule Set 1 was accessed at https://portals.broadinstitute.org/gpp/public/analysis-tools/sgrna-design-v1 on March 23, 2019. Rule Set 2 was (Azimuth 2.0) was accessed at https://github.com/maximilianh/crisporWebsite/tree/master/bin/Azimuth-2.0 on March 24, 2019. SSC was accessed at https://sourceforge.net/projects/spacerscoringcrispr/ on April 1, 2019

Code availability

Scripts for off-target enumeration and CSC implementation are freely available at our bitbucket repository (URL).

STARS

The STARS software (v1.3)[7] was used to predict gene essentiality based on raw and corrected $\log_2$-fold changes. The STARS software was run on default parameters with following explicit parameters specified: threshold percentage set to 10, directionality set to N. STARS was accessed at https://portals.broadinstitute.org/gpp/public/software/stars on June 20, 2019.

Gene Ontology Analysis

Enrichment of Gene Ontology in the Biological Processes category was calculate using two different web-based tools. Gorilla[26] was accessed at http://cbl-gorilla.cs.technion.ac.il/ on July 31st, 2019. These results retrieved from this analysis are shown in Figure 6b. The Gene Ontology Resource[27] was accessed at http://geneontology.org/ on August 1st, 2019. The results from this analysis are shown as Supplementary Table 11.

Cell Culture and generation of Cas9-expressing cells

Cells were cultured at 37 °C (5% CO2) in DME-HG supplemented with 10% FCS, L-glutamine (2 mM), penicillin (100 U ml^{-1}) and streptomycin (100 µg ml^{-1}). For infections, 293T cells (ATCC; # CRL-3216) were transfected with lentiviral constructs and packaging plasmids using Lipofectamine 2000 (Invitrogen) according to manufacturer's instructions. Viral media was collected 48h after transfection, concentrated using Lenti-X Concentrator (Takara) and used to infect A375 cells (ATCC; # CRL-1619). To generate a Cas9-expressing cell line, A375 cells were infected with lentiCas9-Blast (Addgene #52962) and selected for 7 days with 10 µg ml^{-1} of Blasticidin.

Competition Assays

We selected two gRNAs against each of our candidate genes and cloned them into LentiGuide-NLS–GFP[48], using standard oligo cloning protocols. Sequences of all gRNAs along with their target coordinates are show in Supplementary Table 12. Cas9-expressing cells were infected with these constructs or with and LentiGuide-NLS–mCherry virus[48] as a control. One day after infection, cells were selected with 1 µg ml^{-1} of puromycin for 48h after which point, gRNA-expressing cells were mixed in a 4:1 ratio with mCherry-labeled cells and plated in a well of a 24-well-plate (total 5000 cells per well). The total number of cells expressing GFP or mCherry fluorescence was determined every 3h for the course of seven days with IncuCyte Live Cell Analysis Systems (Sartorius).

ACKNOWLEDGEMENTS

We would like to thank all members of the Vidigal lab for critical comments on the manuscript. We thank Ralph Garippa and Michael Lichten for suggesting the development of

CSC and Mathew Hall for discussions on the implications of gRNA specificity to data reproducibility. We also thank John Doench for sharing data from his lab ahead of peer-review publication. Karen Wolcott and the CCR Flow Cytometry Core helped with the flow cytometry data analysis and Eros Denchi and Benura Riveri provided critical help and advice on performing the competition assays. We also thank Raphеal Pellosof for discussion on quantifying gRNA-scaffold complementarity. Eros Denchi, Luke E Dow, Lee Zamparo, and Ram Kannan provided critical feedback on the manuscript. Finally, ARP thanks Olaf Andersen for mentorship and support throughout his professional training. This work utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov) and the UCSF Wynton cluster (https://wynton.ucsf.edu), and was supported by the Intramural Research Program of the National Institutes of Health (NIH) and a FLEX grant from the Center for Cancer Research (JAV).

AUTHOR CONTRIBUTIONS

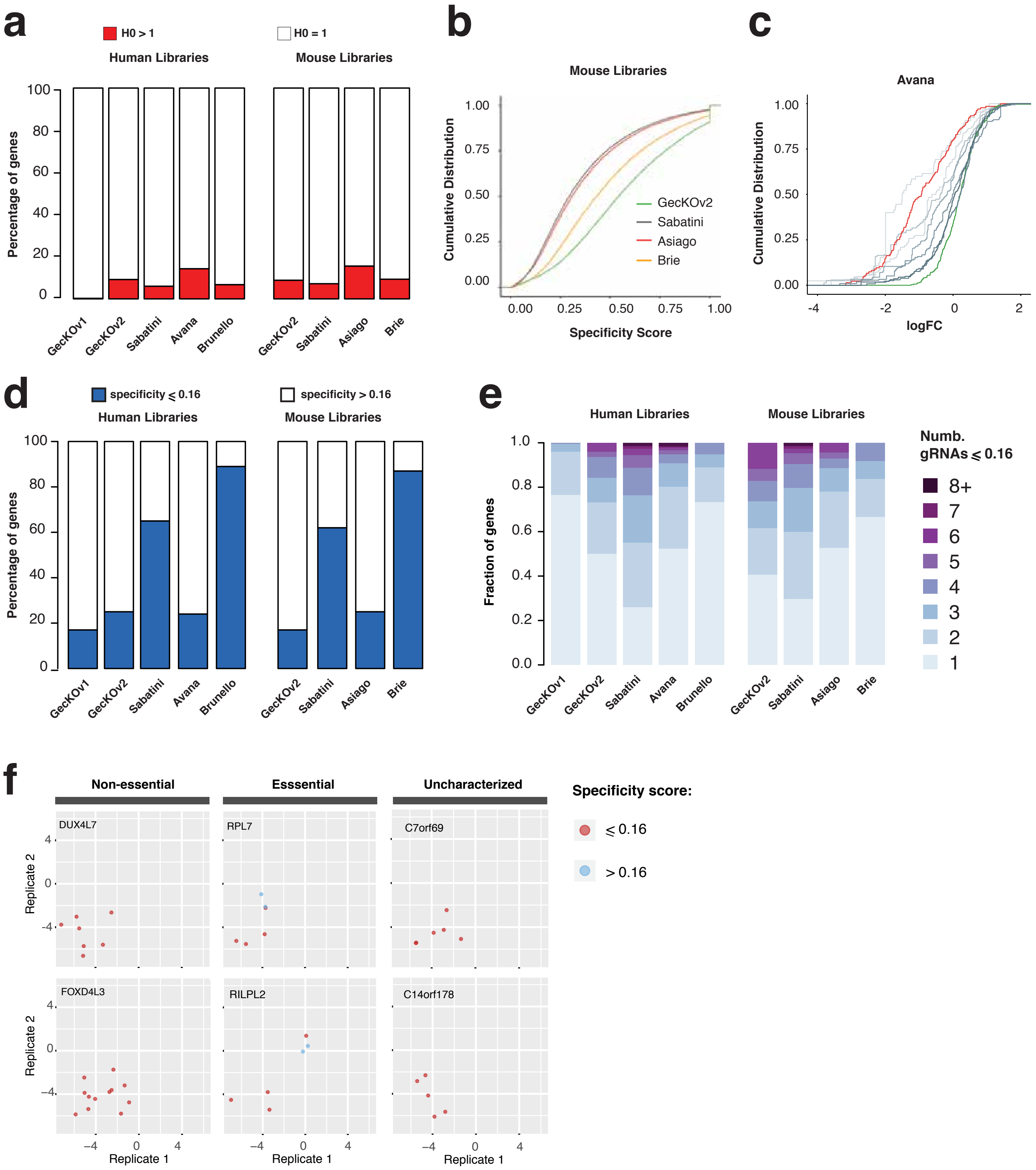
A.R.P. and J.A.V. conceived and designed the study. A.R.P. and J.A.V. performed computational analysis and interpretation of results. A.R.P. wrote and implemented all software. A.R.P., and J.A.V. processed and managed data. R.K.P. assisted with computational analysis. L.S. designed and performed CRISPR experiments. L.S. and J.A.V. analyzed and interpreted data from CRISPR experiments. J.A.V. provided project management. J.A.V, A.R.P., and L.S. wrote and/or revised the manuscript with assistance from R.K.P.. J.A.V. supervised the study.

REFERENCES

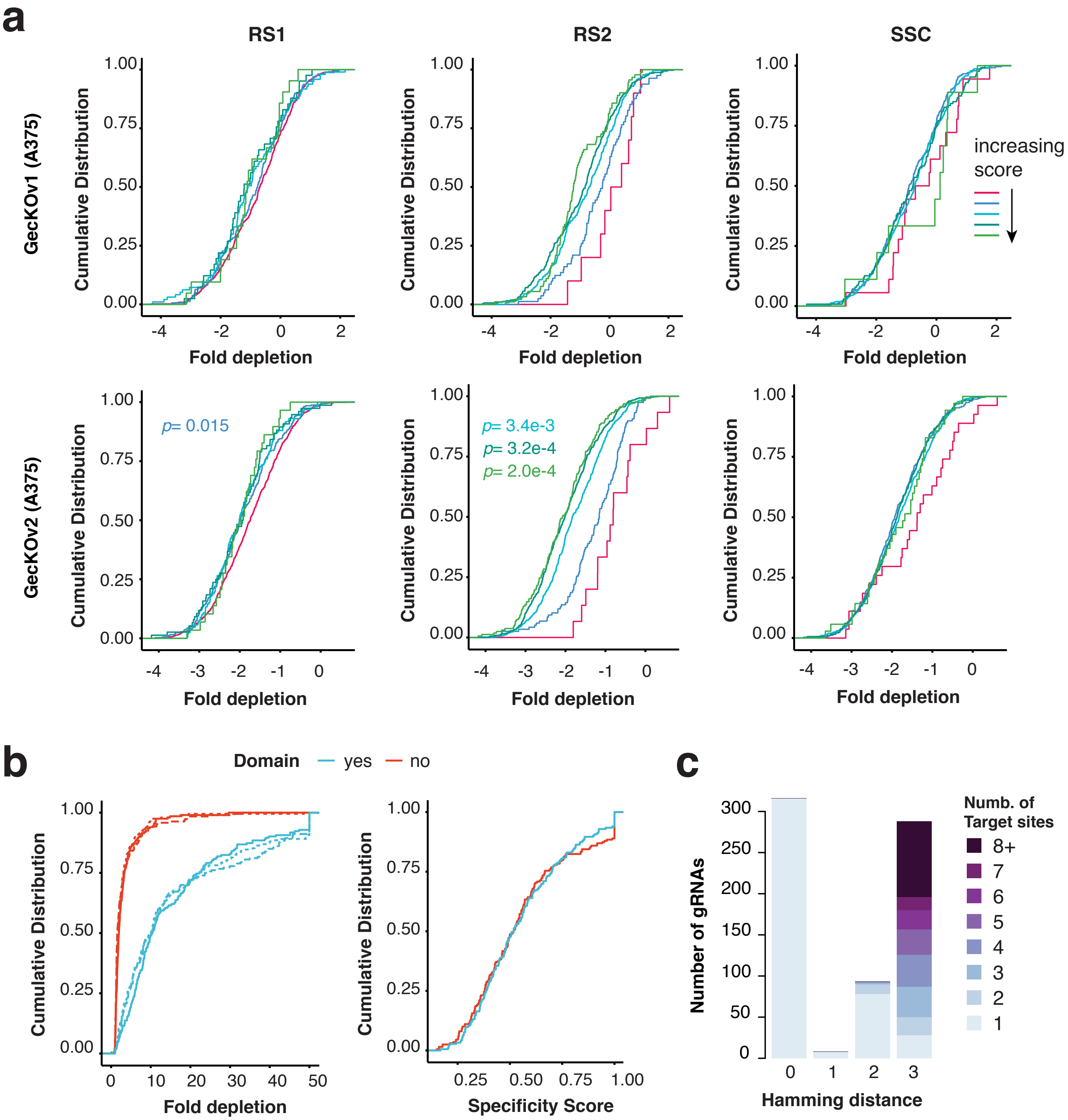
1. Shalem, O., Sanjana, N.E. & Zhang, F. High-throughput functional genomics using CRISPR-Cas9. *Nat Rev Genet* **16**, 299-311 (2015).
2. Sanson, K.R. et al. Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. *Nat Commun* **9**, 5416 (2018).
3. Shalem, O. et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84-87 (2014).
4. Shi, J. et al. Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. *Nat Biotechnol* **33**, 661-667 (2015).
5. Wang, E. et al. Targeting an RNA-Binding Protein Network in Acute Myeloid Leukemia. *Cancer Cell* **35**, 369-384 e367 (2019).
6. Rogers, Z.N. et al. A quantitative and multiplexed approach to uncover the fitness landscape of tumor suppression in vivo. *Nat Methods* **14**, 737-742 (2017).
7. Doench, J.G. et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat Biotechnol* **34**, 184-191 (2016).
8. Sanjana, N.E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat Methods* **11**, 783-784 (2014).
9. Doench, J.G. et al. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat Biotechnol* **32**, 1262-1267 (2014).
10. Vidigal, J.A. & Ventura, A. Rapid and efficient one-step generation of paired gRNA CRISPR-Cas9 libraries. *Nat Commun* **6**, 8083 (2015).
11. Perez, A.R. et al. GuideScan software for improved single and paired CRISPR guide RNA design. *Nat Biotechnol* **35**, 347-349 (2017).
12. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359 (2012).
13. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
14. Tsai, S.Q. et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat Biotechnol* **33**, 187-197 (2015).
15. Aguirre, A.J. et al. Genomic Copy Number Dictates a Gene-Independent Cell Response to CRISPR/Cas9 Targeting. *Cancer Discov* **6**, 914-929 (2016).
16. Wang, T. et al. Identification and characterization of essential genes in the human genome. *Science* **350**, 1096-1101 (2015).
17. Wang, T. et al. Gene Essentiality Profiling Reveals Gene Networks and Synthetic Lethal Interactions with Oncogenic Ras. *Cell* **168**, 890-903 e815 (2017).
18. Fortin, J.P. et al. Multiple-gene targeting and mismatch tolerance can confound analysis of genome-wide pooled CRISPR screens. *Genome Biol* **20**, 21 (2019).
19. Hart, T., Brown, K.R., Sircoulomb, F., Rottapel, R. & Moffat, J. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Mol Syst Biol* **10**, 733 (2014).
20. Morgens, D.W. et al. Genome-scale measurement of off-target activity using Cas9 toxicity in high-throughput screens. *Nat Commun* **8**, 15178 (2017).
21. Tycko, J. et al. Mitigation of off-target toxicity in CRISPR-Cas9 screens for essential non-coding elements. *Nat Commun* **10**, 4063 (2019).
22. Xu, H. et al. Sequence determinants of improved CRISPR sgRNA design. *Genome Res* **25**, 1147-1157 (2015).
23. Wang, T., Wei, J.J., Sabatini, D.M. & Lander, E.S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80-84 (2014).
24. Luo, B. et al. Highly parallel identification of essential genes in cancer cells. *Proc Natl Acad Sci U S A* **105**, 20380-20385 (2008).
25. Li, W. et al. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol* **15**, 554 (2014).
26. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009).
27. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29 (2000).
28. The Gene Ontology, C. The Gene Ontology Resource: 20 years and still GOing strong.

Nucleic Acids Res **47**, D330-D338 (2019).

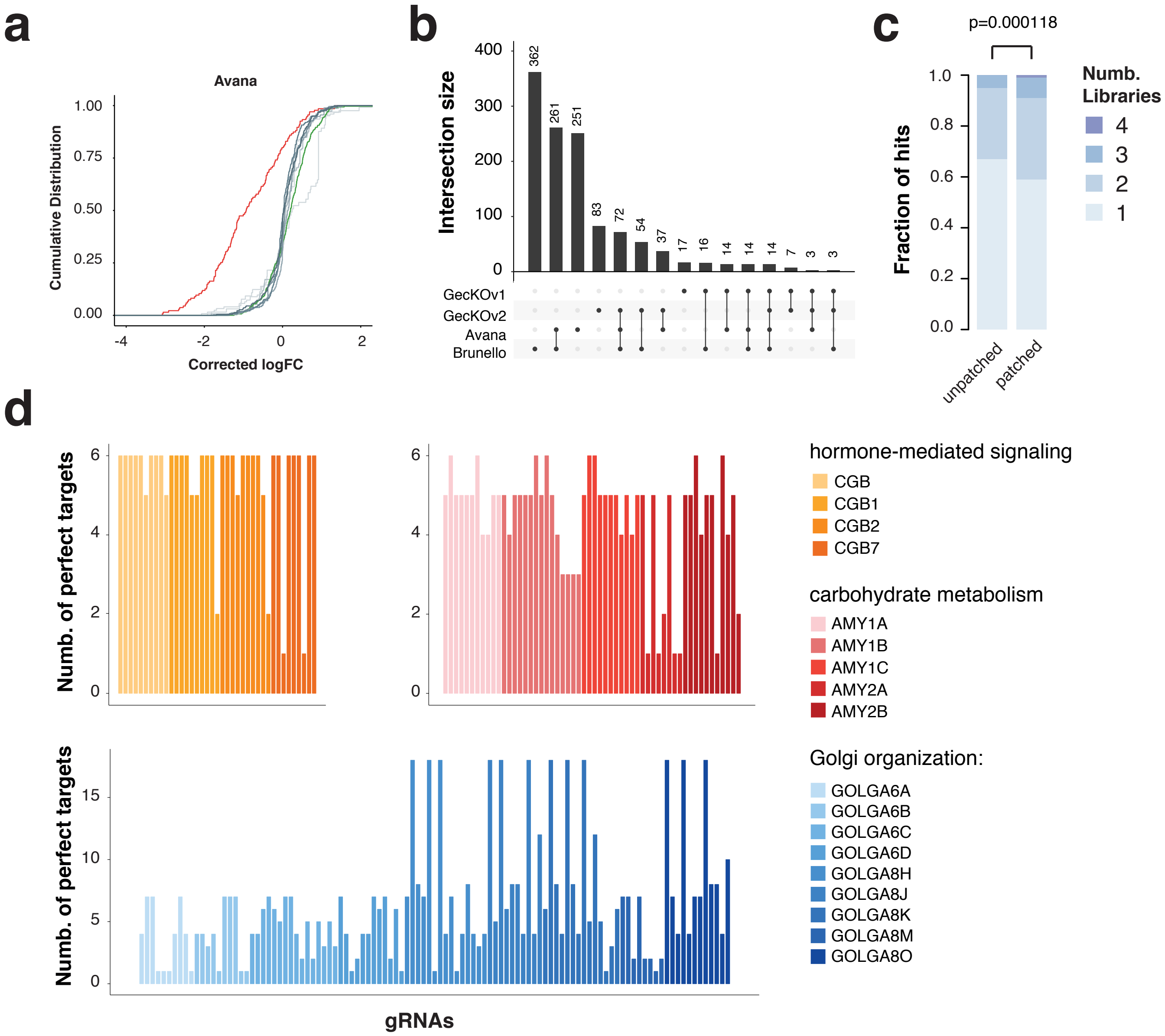
29. Munoz, D.M. et al. CRISPR Screens Provide a Comprehensive Assessment of Cancer Vulnerabilities but Generate False-Positive Hits for Highly Amplified Genomic Regions. *Cancer Discov* **6**, 900-913 (2016).
30. Meyers, R.M. et al. Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat Genet* **49**, 1779-1784 (2017).
31. Evers, B. et al. CRISPR knockout screening outperforms shRNA and CRISPRi in identifying essential genes. *Nat Biotechnol* **34**, 631-633 (2016).
32. Smith, I. et al. Evaluation of RNAi and CRISPR technologies by large-scale gene expression profiling in the Connectivity Map. *PLoS Biol* **15**, e2003213 (2017).
33. Lu, J. et al. MicroRNA expression profiles classify human cancers. *Nature* **435**, 834-838 (2005).
34. Kumar, M.S. et al. Dicer1 functions as a haploinsufficient tumor suppressor. *Genes Dev* **23**, 2700-2704 (2009).
35. Kumar, M.S., Lu, J., Mercer, K.L., Golub, T.R. & Jacks, T. Impaired microRNA processing enhances cellular transformation and tumorigenesis. *Nat Genet* **39**, 673-677 (2007).
36. Huber, M.D., Dworet, J.H., Shire, K., Frappier, L. & McAlear, M.A. The budding yeast homolog of the human EBNA1-binding protein 2 (Ebp2p) is an essential nucleolar protein required for pre-rRNA processing. *J Biol Chem* **275**, 28764-28773 (2000).
37. Tsujii, R. et al. Ebp2p, yeast homologue of a human protein that interacts with Epstein-Barr virus nuclear antigen 1, is required for pre-rRNA processing and ribosomal subunit assembly. *Genes Cells* **5**, 543-553 (2000).
38. Delaunay, S. & Frye, M. RNA modifications regulating cell fate in cancer. *Nat Cell Biol* **21**, 552-559 (2019).
39. Rapino, F. et al. Codon-specific translation reprogramming promotes resistance to targeted therapy. *Nature* **558**, 605-609 (2018).
40. Baillat, D. et al. Integrator, a multiprotein mediator of small nuclear RNA processing, associates with the C-terminal repeat of RNA polymerase II. *Cell* **123**, 265-276 (2005).
41. Ezzeddine, N. et al. A subset of Drosophila integrator proteins is essential for efficient U7 snRNA and spliceosomal snRNA 3'-end formation. *Mol Cell Biol* **31**, 328-341 (2011).
42. Lai, F., Gardini, A., Zhang, A. & Shiekhattar, R. Integrator mediates the biogenesis of enhancer RNAs. *Nature* **525**, 399-403 (2015).
43. Yue, J. et al. Integrator orchestrates RAS/ERK1/2 signaling transcriptional programs. *Genes Dev* **31**, 1809-1820 (2017).
44. Gardini, A. et al. Integrator regulates transcriptional initiation and pause release following activation. *Mol Cell* **56**, 128-139 (2014).
45. Skaar, J.R. et al. The Integrator complex controls the termination of transcription at diverse classes of gene targets. *Cell Res* **25**, 288-305 (2015).
46. Tatomer, D.C. et al. The Integrator complex cleaves nascent mRNAs to attenuate transcription. *Genes Dev* (2019).
47. Nathan D. Elrod, T.H., Kai-Lieh Huang, Deirdre C. Tatomer, Jeremy E. Wilusz, Eric J. Wagner, Karen Adelman The Integrator complex terminates promoter-proximal transcription at protein-coding genes. *bioRxiv* (2019).
48. Noordermeer, S.M. et al. The shieldin complex mediates 53BP1-dependent DNA repair. *Nature* **560**, 117-121 (2018).



Supplementary Figure 1. Characterizing Off-targets in Genome-wide libraries. (a) Percentage of genes targeted by gRNAs with a single (H0=1) or more than one (H0>1) perfect target in the genome. **(b)** Cumulative distributions of specificity scores for the gRNAs in mouse libraries. **(c)** Cumulative distributions of log2 fold-change (logFC) of gRNAs from the Avana libraries during viability screens in A375 cells. Guides targeting non-essential genes were binned based on decreasing specificity scores (grey curves). Distributions of gRNAs targeting essential (red) or non-essential (green) genes and with specificity of 1 are plotted for comparison. **(d)** Percentage of genes targeted by gRNAs with a specificity score equal or below 0.16 in each of the libraries. **(e)** Fraction of genes with unspecific gRNAs that are targeted by only one (light blue) or by more (2-8+, graded colors) unspecific guides. **(f)** Examples of gRNAs from the Avana library targeting genes with known (non-essential, essential) or unknown (uncharacterized) requirements for cell viability/proliferation. Log2 fold-changes for two replicates experiments in A375 cells are plotted. Guide RNAs are color-coded based on their specificity values.



Supplementary Figure 2. Efficiency rules and depletion of gRNAs targeting essential genes. (a) Cumulative distributions showing fold depletion of specific gRNAs (specificity score > 0.16) targeting known essential genes in viability screens performed in A375 cells using the GecKOv1 (top) or GecKOv2 (bottom) library. Guides are binned based on increased RS1, RS2, and SSC scores. A statistically significant segregation of the curves is observed for increasing RS2 scores in the GecKOv2 dataset (Kolmogorov–Smirnov test, Bonferroni correction), and to a lesser extent for RS1 scores. **(b)** Preferential depletion of gRNAs targeting protein domains (left) is not driven by higher promiscuity of gRNAs (right). **(c)** Characterization of the off-target space of gRNAs plotted in (b), showing fraction of guides that have only one (light blue) or by more (2-8+, graded colors) potential target sites with zero (hamming distance 0) or up to three (hamming distance 3) mismatches to the gRNA.



Supplementary Figure 3. CSC improves hit calling amongst CRISPR libraries. (a) Cumulative distributions of corrected log2 fold-change of gRNAs from the Avana libraries during viability screens in A375 cells. Guides targeting non-essential genes were binned based on decreasing specificity scores (grey curves). Distributions of gRNAs targeting essential (red) or non-essential (green) genes and with specificity of 1 are plotted for comparison. **(b)** Upset plot showing the size of the intersections between uncharacterized genes identified as hits by each of the libraries after CSC implementation (FDR<10%). **(c)** Fraction of hits identified by 1, 2, 3, or all 4 libraries before and after correcting gRNA log2 fold-changes with CSC. P-values were calculated using the Chi-square test. **(d)** Number of perfect target sites for gRNAs contributing to hit calling of genes driving GO terms in Figure 6b.