Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Subject Section

Improved design and analysis of CRISPR knockout screens

Chen-Hao Chen[1,2,3]*, Tengfei Xiao[1,4]*, Han Xu[1,2,7], Peng Jiang[1,2], Clifford A. Meyer[1,2], Wei Li[1,2,5,6]§, Myles Brown[1,4]§, X. Shirley Liu[1,2]§

[1]Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, MA 02115, USA. [2]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, and Harvard School of Public Health, Boston, MA 02115, USA. [3]Biological and Biomedical Science Program, Harvard Medical School, Boston, MA 02115, USA. [4]Department of Medical Oncology, Dana-Farber Cancer Institute, and Harvard Medical School, Boston, MA 02115, USA. [5]Center for Genetic Medicine Research, Children's National Health System, Washington, DC 20010, USA. [6]Department of Genomics and Precision Medicine, The George Washington School of Medicine and Health Sciences, Washington, DC 20010, USA. [7]Department of Epigenetics and Molecular Carcinogenesis, The University of Texas MD Anderson Cancer Center, Smithville, TX 78957, USA.

§To whom correspondence should be addressed.

*These authors contributed equally to this work.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Genome-wide CRISPR-Cas9 screen has been widely used to interrogate gene functions. However, the rules to design better libraries beg further refinement.

Results: We found sgRNA outliers are characterized by higher G-nucleotide counts, especially in regions distal from the PAM motif, and are associated with stronger off-target activities. Furthermore, using non-targeting sgRNAs as negative controls lead to strong bias, which can be mitigated by using sgRNAs targeting multiple "safe harbor" regions. Custom-designed screens confirmed our findings and further revealed that 19nt sgRNAs consistently gave the best signal-to-noise ratio. Collectively, our analysis motivated the design of a new genome-wide CRISPR/Cas9 screen library and uncovered some intriguing properties of the CRISPR-Cas9 system.

Availability: The MAGeCK workflow is available open source at https://bitbucket.org/liulab/mageck_nest under the MIT license.

Contact: xsliu@jimmy.harvard.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The clustered, regularly interspaced, short palindromic repeat (CRISPR)-Cas9 system is a new genome editing technology that becomes prominent in many biomedical research areas. In this system, single guide RNAs (sgRNAs) direct Cas9 nucleases to induce double-strand breaks at targeted genomic regions(Cong, et al., 2013; Jinek, et al., 2012; Mali, et al., 2013). Based on this system, CRISPR-Cas9 loss-of-function screens can interrogate the functions of coding genes(Koike-Yusa, et al., 2014; Shalem, et al., 2014; Wang, et al., 2014; Zhou, et al., 2014) and non-coding elements(Canver, et al., 2015; Korkmaz, et al., 2016; Zhu, et al., 2016), and generate hypotheses on cell dependency, drug response, and gene regulation in a high-throughput and unbiased manner(Diao, et al., 2016; Hart, et al., 2015; Parnas, et al., 2015; Wang, et al., 2015). From a computational biology perspective, several algorithms have been developed to characterize sgRNAs with high specificity and efficiency(Doench, et al., 2016; Doench, et al., 2014; Hsu, et al., 2013; Xu, et al., 2015) that can be used in designing CRISPR screen libraries. Despite these efforts, methods for designing CRISPR screens are still

being refined from different aspects. First, sgRNA outliers, or sgRNAs with discrepant behaviors from other sgRNAs targeting the same gene, are common in screen data, but their features and mechanisms remain poorly characterized. Second, it's known that spacer length may vary in the CRISPR-Cas9 system (Fu, et al., 2014; Morgens, et al., 2017), but the optimal length was only studied in single guides and single targets. Furthermore, it remains unclear how spacer lengths affect signal-to-noise ratio (the extent of the fold changes of guides compared to their variances) in the screening settings.

We studied both issues based on the MAGeCK-VISPR model we previously developed (Jiang, et al., 2015; Li, et al., 2015). By examining published screens(Wang, et al., 2015; Wang, et al., 2014), we identified outlier sgRNAs and uncovered their sequence features to inform future library design. We further showed stronger off-target cleavages contribute to the outlier behaviors. We also found a strong bias in CRISPR screen when normalizing read counts with commonly used non-targeting sgRNAs and proposed an alternative normalization to mitigate such bias. We performed custom-designed screens to validate these findings, and further explored sgRNA design rules that can improve the screening results, including the optimal spacer length for higher cutting efficiencies and better signal-to-noise ratios. Finally, we designed a genome-wide CRISPR/Cas9 screening library based on these new rules and demonstrated its performance in identifying known essential genes in different cell types.

2 Methods

2.1 The MAGeCK and MAGeCK-VISPR model

Our laboratory has previously developed algorithms MAGeCK and MAGeCK-VISPR for identifying CRISPR screen hits in different scenarios(Li, et al., 2015; Li, et al., 2014). In two-condition comparisons, MAGeCK uses a negative binomial model to assess the degree of selections of individual sgRNAs and adopts robust rank aggregation (RRA) algorithm(Kolde, et al., 2012) to aggregate multiple sgRNAs on a gene to evaluate gene selection. MAGeCK-VISPR(Li, et al., 2015) further quantitatively estimates gene selections by optimizing a joint likelihood function of observing the read counts of different sgRNAs with varying behaviors in multiple conditions. The output of MAGeCK-VISPR is a "beta score" for gene g in condition r, β_{gr}, analogous to the "log fold change" in differential gene expression analysis. More specifically, the read count of sgRNA i in sample j, or K_{ij}, is modeled as:

$$K_{ij} \sim NB(\mu_{ij}, \alpha_i)$$

Where μ_{ij} and α_i are the mean and over-dispersion factor of the negative binomial (NB) distribution, respectively. The mean value μ_{ij} is further modeled as:

$$\mu_{ij}(\vec{\beta}) = s_j \exp(\sum_r d_{jr}\beta_{gr})$$

Where s_j is the size factor of sample j for adjusting sequencing depths of the samples, and $\vec{\beta}$ is the vector of all beta scores for gene g. To deal with complex experimental settings, we included design matrix (D). With J samples affected by R conditions, D is a binary matrix with its element $d_{jr} = 1$ if sample j is affected by condition r and 0 otherwise. The objective function is a form of regularization:

$$\hat{\beta}_{gr} = argmax\left(\sum_{ij} log f_{NB}\left(K_{ij}; \mu_{ij}(\vec{\beta}), \alpha_i\right) + \Lambda(\vec{\beta})\right) - (1)$$

Where f_{NB} is the probabilistic density function (PDF) of the Negative Binomial distribution, and

$$\Lambda(\vec{\beta}) = \sum_r \frac{-{\beta_{gr}}^2}{2\sigma_r^2}$$

The estimated standard deviation, σ_r, was calculated using the naive estimators of β_{gr}.

2.2 Identifying sgRNA outliers

sgRNA outliers are those that have different behaviors compared with other sgRNAs targeting the same gene. A single outlier that does not fit the assumed distributions can overly influence the estimations of the beta score. Therefore, we tried to identify these outliers using 3-step approach: candidate outlier prediction, candidate outlier validation, and outlier detection.

Step-1: Candidate outlier prediction

A sgRNA is likely to be an outlier if its log fold is extremely different from other sgRNAs. Therefore, in the first step, candidate outlier prediction, we identified the potential sgRNAs outliers by considering their log fold changes (LFCs). For each paired condition, we calculated the median and standard deviation of the LFCs and defined the candidate outliers if their LFCs fall beyond median $\pm$ 1.5 standard deviation estimation (σ). Specifically, we followed the "quantile matching" approach in DESeq2(Love, et al., 2014) : σ is chosen such that the (1-p) empirical quantile of the absolute values of LFC ($Q_{|LFC|}$) matches the (1-p/2) theoretical quantile of $N(0, \sigma^2)$ (Q_N), where p is set as 0.32:

$$\sigma = \frac{Q_{|LFC|}(1-p)}{Q_N(1-\frac{p}{2})}$$

Note that for a distribution with a long tail, the traditional estimation of standard deviation will be distorted. Assuming that samples with beta scores close to 0 follows normal distribution, we set a value of p=0.32 to calculate standard deviation using only the 68% of samples (samples within 1 standard derivation) closes to zero. In this way, the samples with beta scores far from zero will not distort the estimation of standard deviation.

Step-2: Candidate outlier *in silico* validation

Noticing that a sgRNA outlier may significantly influence the beta score estimation, a candidate outlier is validated if there is a significant change of beta score, β_{gr}, after removing the candidate outlier. Therefore, in the second step, the candidate outlier *in silico* validation, we calculated the beta score with and without the candidate outlier respectively using Equation (1). Define:

$\beta^{raw} = \beta_{gr}$, when all sgRNAs are used;

$\beta^{i} = \beta_{gr}$, when sgRNA i is excluded.

Then candidate outlier i is *in silico* validated if:

$$\log(abs(\beta^{raw})/abs(\beta^{i})) > (5 - 0.2 * number\ of\ sgRNAs)$$

With outlier removal, we could prevent the beta score estimation from distortion by strong outliers.

Step-3: Outlier detection

With previous two steps, we could estimate the beta scores robustly. However, some moderate outliers cannot be identified if sufficient sgRNAs prevent the beta score from distortion by a single outlier. Therefore, with robust estimators of beta scores, in the final step, we re-defined a sgRNA as an outlier if the probability of observing its count conditioned on pre-calculated beta score falls below a certain threshold. In other words, sgRNA i is an outlier if:

$$\sum_j log f_{NB}\left(K_{ij}; \mu_{ij}(\vec{\beta}), \alpha_i\right) < T$$

where f_{NB} is the probabilistic density function (PDF) of the Negative Binomial distribution. The threshold T was determined such that that 90% of the validated outliers defined in step 2 can be removed.

2.3 Extracting sequence features using Elastic-Net regression

To identify the sequence features that associate with stronger sgRNA outliers, we applied Elastic-Net regression to extract the sequence features as our previous work(Xu, et al., 2015). Suppose $X = \{X_1, X_2, \ldots, X_n\}$ is the set of encoded sequence vectors and $Y = \{Y_1, Y_2, \ldots, Y_n\}$ is the set of outputs representing whether the sgRNAs are stronger outliers, where n is the number of sgRNAs samples for training. If sgRNA i is an outlier, the corresponding $Y_i = 1$ and 0 otherwise. Let M be the length of the input vectors; the Elastic-Net regression computes the parameters $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_M]^T$ that minimize an object function E:

$$E = ||Y - \boldsymbol{\beta}^T X||^2 + \lambda(\alpha||\boldsymbol{\beta}||^1 + (1-\alpha)||\boldsymbol{\beta}||^2)$$

Where α and λ are parameters estimated using cross validation, $||\boldsymbol{\beta}||^1 = \sum_i |\beta_i|$ and $||\boldsymbol{\beta}||^2 = \sum_i \beta_i^2$. We used glmnet in R package to implement the Elastic-Net regression(Friedman, et al., 2010).

2.4 CRISPR screening design and experimental procedure

We designed and performed a CRISPR screening experiment to study the effects of different normalization methods and different sgRNA lengths. The screening library has four types of sgRNAs: sgRNAs targeting AAVS1 (a region whose disruption does not have any lethal phenotype), non-targeting sgRNAs, sgRNAs targeting 51 ribosomal genes and 503 cancer-related genes that are considered to be lethal. The details of the library design and the experiment are in Supplementary Data.

3 Results

3.1 sgRNAs outlier identification and characterization

Different sgRNAs targeting the same gene can lead to varying phenotypes or selection levels in the screen due to different cleavage and repair efficiencies, local chromatin structure, protein domains, and potential off-target effects, *etc.* (Hsu, et al., 2013; Knight, et al., 2015; Shi, et al., 2015). Some sgRNAs with outlier phenotypes compared with other sgRNAs on the same gene, regardless of the causes, behave consistently in multiple screen conditions(Wang, et al., 2015) (Fig. 1a, b), suggesting that the discrepant phenotypes could arise from intrinsic features of the sgRNA in addition to random variances in the experiments. We are especially interested in 'strong negative outliers' (as Fig. 1a), which are defined as having much larger negative LFCs compared with other sgRNAs targeting the same gene and are more likely caused by off-target cleavages.

Based on the MAGeCK-VISPR model, we implemented an approach to identify such outliers, which tests whether one sgRNA has big effects on the gene-level beta score estimates or the probability of observing the sgRNA conditioned on the gene-level beta score is low (see Methods). This outlier detection and removal approach did identify sgRNAs with aberrant LFC on a gene (Fig. 1c). In published screens on four leukemia cell lines (Wang, et al., 2015), nine thousand out of 182K sgRNAs on average were identified as outliers Among them, 911 sgRNAs are outliers that are consistent in all four screens (Supplementary Fig. 1), and 80% of these outliers (729/911) are "strong negative outliers" with stronger negative selection as other gRNAs on the same gene (as Fig. 1a). To rule out the possibility that these sgRNAs knockout their intended targets with extremely high efficiencies, we further limited our analysis to 564 outliers (Supplementary Table 1) that target known non-essential genes (Hart and Moffat, 2016), as inactivating these genes is unlikely to affect cell growth.

Comparing the sequence features of these 564 'strong negative outliers' with all 18,000 sgRNAs in the library, we found that they have higher G-nucleotide but lower C-nucleotide counts in the target DNA sequence (Fig. 1d, Supplementary Fig. 1b-d). To identify potential sequence features that can distinguish outliers and non-outliers, we trained an elastic net model(Friedman, et al., 2010), a regularized regression method that considers both the L1 and L2 penalties of the lasso and ridge methods. In the training dataset, the predictor variable is a binary vector representing the presence or absence of the nucleotides, and the response variable is a binary variable indicating whether the gRNA is an outlier. Our model showed that outliers tend to contain more G-nucleotides in the 10-nucleotide non-seed region distal from the PAM motif (Fig. 1e). To exclude possible biases of a single library, we confirmed our finding using another screen dataset (Meyers, et al., 2017) (Supplementary Fig. 2a-b). We further tested our predictive model on other CRISPR-Cas9 knockout (Wang, et al., 2014) or CRISPR-dCas9 inhibition screening (Horlbeck, et al., 2016) datasets. The output of the model is an "outlier score", indicating how likely the input sgRNA is an outlier. We found that 'strong negative outliers' in both datasets have significantly higher outlier scores than non-outliers (Supplementary Fig. 2d, e), suggesting outlier features we found are consistent across different datasets. These findings also suggest that a better CRISPR sgRNA design should at least avoid extreme G content in the non-seed region in case of potential off-target effects.

Considering that strong off-target activities can lead to 'strong negative outliers', we reanalyzed a previous study that measured the off-target activities between mismatched sgRNA:DNA pairs, defined as the decrease of CD33 protein level by sgRNAs with 1 nucleotide mismatch compared to the target DNA in CD33 locus (Doench, et al., 2016). Instead of modeling off-target activities as functions of mismatched nucleotide pair and position as in (Doench, et al., 2016), we tested how the nucleotide compositions in the Non-seed region affect the off-target activities. SgRNAs with more "G"s (≥5) in Non-seed region have significantly higher off-target activities than those with fewer "G"s (<5) (Fig. 1f). In contrast, there is no difference in off-target activities between sgRNAs with more (≥5) and fewer (<5) "C"s (Supplementary Fig. 2f). These findings suggest sgRNAs targeting sequences with high G-content in the non-seed region have stronger off-target activities, which can lead to strong outlier phenotypes.

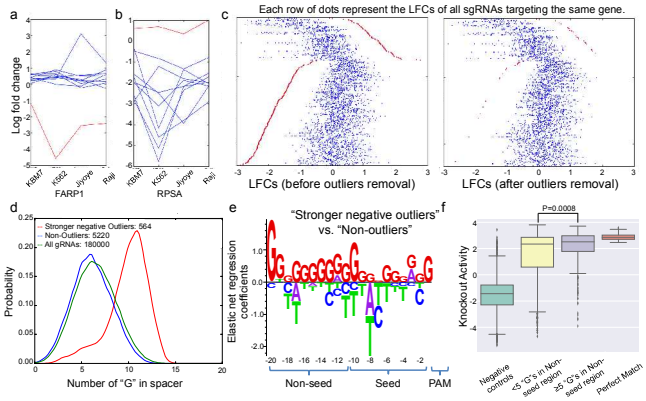

Figure 1. Identifying and characterizing stronger negative sgRNAs outliers.

(a, b) The log fold changes of 10 sgRNAs targeting FARP1 and RPSA in 4 screens (KBM7, K562, Jiyoye, and Raji). The red lines represent sgRNAs outliers, and the blue lines represent other sgRNAs.
(c) Identifying and removing aberrantly stronger negative outliers (red dots). Each row of dots represents the log fold changes (LFCs) of sgRNAs targeting the same gene. .
(d) The G-nucleotide counts of sgRNAs in three groups: stronger negative outliers (red), non-outliers (blue), and all sgRNAs (green).
(e) The sequence features of stronger negative outliers versus non-outliers derived by elastic-net regression. The "Seed" and "Non-seed" regions are defined as a 10-nucleotide window proximal to and distal from the PAM motif, respectively. The data for Fig. 1a-e is from a public screening dataset(Wang, et al., 2015).
(f) The knockout of CD33 expression with different groups of sgRNAs. The 'Perfect Match' are 65 perfect-match sgRNAs with an NGG PAM that produced effective CD33 knockout defined in (Doench, et al., 2016). The "Negative Controls' are the same set of sgRNAs with non-NGG PAM. Those in '≥5 "G"s in Non-seed region' and '<5 "G" in Non-seed region' are sgRNAs with an NGG PAM but 1-nt mismatch compared to the 'Perfect Match' sgRNAs.

3.2 SgRNAs targeting multiple non-essential genes as negative controls reduce false positives in the screen

Correct interpretations of genome-wide screens require proper read count normalization. Since most sgRNAs should generate knockouts without causing phenotype, a straightforward approach is to normalize based on the total read counts of all sgRNAs(Love, et al., 2014) ('total normalization'). Alternatively, many screen libraries include 'non-targeting' negative control sgRNAs, which match nowhere in the genome, for normalization ('non-targeting sgRNA normalization'). In public datasets(Wang, et al., 2015; Wang, et al., 2014), 'total normalization' resulted in a beta-score distribution centered on zero (Supplementary Fig. 3a), while 'non-targeting sgRNA normalization' led to a skewed distribution of beta scores where most of the genes appear as negatively selected (Fig 2a). The bias of 'non-targeting sgRNA normalization' is introduced when sgRNAs targeting non-essential genes impede cell growth from genome cleavage toxicity(Aguirre, et al., 2016; Munoz, et al., 2016), regardless of the gene knockout effects. Therefore, a more appropriate choice of negative controls is a set of sgRNAs targeting non-essential DNA regions. These sgRNAs have already been included in recent library design (Wang, et al., 2017). Indeed, when normalizing read counts using sgRNAs targeting the 'gold standard' 927 non-essential genes previously derived from pooled shRNA screens(Hart, et al., 2014), the beta score distribution is centered on zero (Fig. 2b).

In genome-wide screens, normalizations using either sgRNAs targeting non-essential genes or all genes lead to similar results (Fig. 2b, Supplementary Fig. 3a), as the majority of the genes are assumed to be non-essential. Such assumption may fail in focused (or custom) screens where many targeted genes may be under selection, which necessitates the selection of better negative control sgRNAs. AAVS1 (adeno-associated virus integration site 1) is a "safe harbor" site preferred for gene knock-ins(DeKelver, et al., 2010; Sadelain, et al., 2012). This region appears to be epigenetically open for efficient cleavage, yet cutting or modification at this site results in no phenotypic changes(Ogata, et al., 2003). To test whether sgRNAs targeting AAVS1 could serve as good negative controls, we first designed a genome-wide screen library containing 134 AAVS1-targeting sgRNAs, 349 non-targeting sgRNAs, as well as five sgRNAs per gene in the human genome, and performed screening in a prostate cancer LNCaP-abl cell line. SgRNAs targeting AAVS1 or non-essential genes induced similar LFCs that are stronger than non-targeting sgRNAs, confirming the existence of cleavage toxicity in non-essential regions (Fig. 2c). Also, by comparing normalization methods using different sets of sgRNAs (all, non-targeting, AAVS1-targeting, and non-essential-gene-targeting sgRNAs, respectively), we found normalization using the AAVS1- and non-essential-genes targeting sgRNAs result in almost identical distributions of beta scores (Fig. 2d). Moreover, both 'all sgRNA normalization' and 'non-targeting sgRNA normalization' lead to biases, though to different degrees (Fig. 2d). Since normalization using control guides is an essential step in many computational methods including MAGeCK-VISPR and CRISPR Score (CS)(Wang, et al., 2014), the results of these methods will also be affected by the choice of negative controls (Supplementary Fig. 3b). While methods that only rely on gRNA ranks such as MAGeCK-RRA(Li, et al., 2014) will not be affected, the rankings could not clearly distinguish genes that are negatively, positively, or not selected, which are important when comparing screens over multiple conditions.

To evaluate the normalization methods in a focused screen, we also designed a small screening library that targets ~600 genes, including ribosomal genes and well-known cancer-related genes (see Methods, Supplementary Tables 2, 3). The library also includes the same set of AAVS1-targeting and non-targeting sgRNAs. Similar to genome-wide screens, AAVS1-targeting sgRNAs induced stronger negative selections compared with non-targeting sgRNAs (Supplementary Fig. 3c). Furthermore, using AAVS1-targeting sgRNAs as negative controls in our MAGeCK algorithm substantially increases the sensitivity of the screen, while keeping the same level of false positives (Supplementary Fig. 3d). These results validated the applicability of including AAVS1-targeting sgRNAs in genome-wide, and more importantly in focused screen libraries.

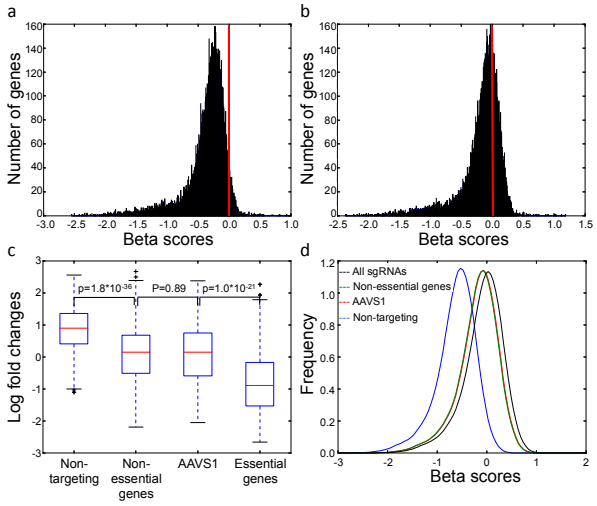

Figure 2. Normalizing read counts using sgRNAs targeting non-essential genes or AAVS1.
(a-b) The distribution of beta scores in public dataset(Wang, et al., 2015) using non-targeting sgRNAs (a) and sgRNAs targeting non-essential genes (b) for normalization.
(c) The log fold change distribution of 349 non-targeting sgRNAs, 467 non-essential genes-targeting sgRNAs, 133 AAVS1-targeting sgRNAs, and 725 essential genes-targeting sgRNAs. P values were calculated using two-sided Student's t-test.
(d) The distribution of beta score using all sgRNAs (black), non-essential genes-targeting sgRNAs (green), AAVS1-targeting sgRNAs (red), and non-targeting sgRNAs (blue) for normalizing read counts, respectively.

3.3 19nt spacers give rise to higher cutting efficiencies and better signal-to-noise ratio

In spCas9 gene editing systems, truncated sgRNAs have been reported to have a better cleavage specificity compared with full-length sgRNAs(Fu, et al., 2014). However, the performances of truncated sgRNAs in screens compared with full-length sgRNAs, as well as the optimal length of truncated sgRNAs, have yet to been fully determined. Therefore, in our small screening library, we designed sgRNAs with 20nt spacers for each ribosomal gene and AAVS1-targeting sgRNAs and then truncated them to 19nt, 18nt, and 17nt (see Methods). We found that 19nt sgRNAs give significantly stronger LFCs in ribosomal genes, reflecting higher cleavage efficiencies (Fig. 3a). If we use the difference between positive-control sgRNAs (sgRNAs targeting ribosomal genes) and negative-control sgRNA (AAVS1-targeting sgRNAs) as a metric for signal-to-noise, 19nt spacers on average give the best performance (Supplementary Fig. 4) in 11 of 12 screens. Moreover, for each ribosomal gene, 19nt sgRNAs gave lower relative standard deviation (*i.e.,* standard deviation divided by mean; see Supplementary Methods.) of LFCs, indicating a more stable behavior (and potentially less off-target cleavages) of gene knockout effects (Fig. 3b).

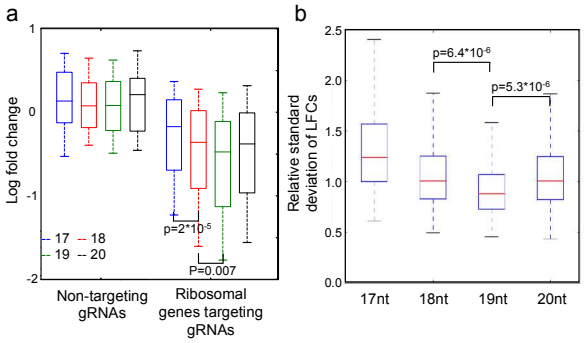

Figure 3. Comparing cleavage efficiencies and signal-to-noise ratios between different lengths of sgRNA spacers.
(a) The log fold changes of sgRNAs with spacer lengths ranging from 17- to 20-nts, including non-targeting sgRNAs and sgRNAs targeting ribosomal genes. For each spacer length, there are 100 non-targeting sgRNAs and 1020 ribosomal genes-targeting sgRNAs. P values were calculated using two-sided Student's t-test.
(b) The relative standard deviation of log fold changes of sgRNAs targeting ribosomal genes with spacer lengths ranging from 17- to 20-nts. There are 612 data points (51 ribosomes genes repeated in 12 screens) for each spacer length. P values were calculated using two-sided Student's t-test.

3.4 A new genome-wide library Improved screen performance

Using the rules we uncovered in this study and our previous work(Xu, et al., 2015), we designed two sub-libraries that target 18,493 human coding genes (named "H1" and "H2"; Supplementary Tables 4, 5). Each sub-library includes sgRNAs with 19nt-long spacers and contains 134 AAVS1-targeting sgRNAs, 349 non-targeting sgRNAs, as well as five sgRNAs targeting each gene in the human genome. After removing sgRNAs that are enriched in G-nucleotide (>40%) and have perfect matches to other coding regions, we prioritized the remaining sgRNAs based on their predicted cleavage efficiencies(Xu, et al., 2015) and the number of perfect matches in the whole genome (see Methods). We conducted screens in LNCaP, abl and T47D cell lines using the H1/H2 library and compared to other genome-wide screen datasets, including Brunello library(Doench, et al., 2016), TKO library(Hart, et al., 2015), and Ong library(Ong, et al., 2017). We found H1/H2 is among the libraries with fewest outlier sgRNA rates (Supplementary Fig. 4b). Assuming that a good library should be able to rank known essential genes as most negatively selected ones, we found that H1/H2, Brunello, and Ong libraries outperformed GeCKOv2 and TKO in identifying known essential genes (Supplementary Fig. 4c-d). These results provide support for our refined CRISPR screen library design rules.

4 Discussions

The CRISPR-cas9 knockout screen has been used to interrogate the functions of coding genes and non-coding elements systemically, but library design is still in their early stage. We first applied MAGeCK-VISPR to public genome-wide screen data and identified a set of 'strong negative outlier' sgRNAs and their sequence characteristics: higher G-nucleotide counts especially in regions distal from PAM motif. Unexpectedly, the effect of the outliers is independent of the count of C-nucleotide, different from previous studies that suggest the role of 'GC' content in determining cleavage efficiencies (Haeussler, et al., 2016; Doench, et al., 2014; Wang, et al., 2014). Since G-C hybridization strengths in DNA-RNA and RNA-DNA hybrids are similar, the distinct effect of G- and C-nucleotides suggests a more crucial role of DNA-endonuclease rather than DNA-RNA interaction in determining outlier effects. Moreover, sgRNAs with higher G-contents in regions distal from PAM motif have stronger off-target activities. It is worth noting that the off-target activity of each sgRNA in Fig. 1e was measured between one sgRNA-DNA pair, and the seemly minor difference between sgRNAs with high and low G-contents will be multiplied by the enormous mismatched sgRNA-DNA pairs in the genome and lead to sgRNA outliers in screens.

Although toxicity from CRISPR cutting has been reported, using non-targeting control for normalization is still a common practice in published literature (Aguirre, et al., 2016; Wang, et al., 2014). We found that normalization using non-targeting sgRNAs, as compared to using all sgRNAs or sgRNAs targeting non-essential genes, could lead to higher false positives (Supplementary Fig. 3d) in calling essential genes. The reason might be because cleavages in non-essential regions can still induce toxicity in cell growth, in consistency with two recent studies showing false positive hits from highly amplified regions in cancer genomes(Aguirre, et al., 2016; Munoz, et al., 2016). Through CRISPR screening experiments, we confirmed that sgRNAs targeting non-essential genes or safe-harbor region could serve as better negative controls and result in fewer false positives compared with non-targeting sgRNAs. Since a single chromatin region may be subject to copy number variations in different cell types, sgRNAs targeting multiple non-essential regions will serve as more robust negative controls. For instance, only 5% (57/1,043) CCLE cell lines have copy number gains in AAVS1 locus, such as HCC1937 and MDAMB157, suggesting that though chance is low, caution should be used when using single region as negative controls. Including correct negative controls is also necessary for custom-designed screens where genes are pre-selected and normalization using total read counts is inappropriate. We proposed a solution to reduce the biases by using either multiple non-essential genes or AAVS1-targeting guides.

Finally, sgRNAs with shorter lengths have been shown to be potent in efficiency and specificity (Fu, et al., 2014), but the optimal performance of truncated sgRNAs with different lengths has not been systematically

investigated in screen setting. We discovered that 19nt sgRNAs consistently provide better cleavage efficiencies and signal-to-noise separations compared with other lengths (17, 18, 20nt). Therefore, using 19nt sgRNAs in either low-throughput experiments or high-throughput screens may give rise to a more accurate inference of gene knockout effects.

We demonstrated that H1/H2 libraries have improved performance in identifying known essential genes with less outlier sgRNAs. However, the fact that comparisons were not performed in the same cellular context might contribute to the observed differences. Also, since different libraries used distinct approaches to improve screen performance, integrating their respective advantages might further improve the next generation library design.

Although we characterized multiple features of CRISPR screens using computational approaches, the exact mechanisms behind these findings remain unknown. First, it is unclear how sgRNAs with higher G-nucleotide content are associated with stronger outliers. We suspected that outlier gRNAs with high G-nucleotides have promiscuous off-target binding and cutting at many CpG islands in the genome. Existing experimental approaches to detect off-target cleavages (Kim, et al., 2015; Tsai, et al., 2015) may be limited to study these gRNAs, as the cleavages in each binding site may be low. Second, although we have shown the advantages of using 19bp sgRNA spacers from statistical perspectives, how different lengths of sgRNA spacers give rise to various cleavage strengths and off-targets remain to be determined. Last but not least, all the above findings are derived in the SpCas9 system, and the rules in different RNA-guided DNA endonuclease systems require further investigations.

Collectively, our study provided novel insights into the properties of CRISPR and the design of both high- and low throughput CRISPR experiments. We designed two genome-wide libraries and showed the improved performance using the rules we uncovered. The characterized features and design rules, as well as the libraries, will benefit and expedite the application of CRISPR techniques.

Funding

This work was supported by the National Institutes of Health [HG008727, HG008927]. Funding for open access charge: National Institutes of Health. WL is supported in part by funds of the Center for Genetic Medicine Research and the Gilbert Family Neurofibromatosis Institute at Children's National Health System.

Conflict of Interest: none declared.

References

Aguirre, A.J., *et al.* Genomic copy number dictates a gene-independent cell response to CRISPR-Cas9 targeting. *Cancer Discov* 2016.

Canver, M.C., *et al.* BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* 2015;527(7577):192-197.

Cong, L., *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* 2013;339(6121):819-823.

DeKelver, R.C., *et al.* Functional genomics, proteomics, and regulatory DNA analysis in isogenic settings using zinc finger nuclease-driven transgenesis into a safe harbor locus in the human genome. *Genome Res* 2010;20(8):1133-1142.

Diao, Y., *et al.* A new class of temporarily phenotypic enhancers identified by CRISPR/Cas9-mediated genetic screening. *Genome Res* 2016;26(3):397-405.

Doench, J.G., *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat Biotechnol* 2016;34(2):184-191.

Doench, J.G., *et al.* Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat Biotechnol* 2014;32(12):1262-1267.

Friedman, J., Hastie, T. and Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* 2010;33(1):1-22.

Fu, Y., *et al.* Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat Biotechnol* 2014;32(3):279-284.

Haeussler, M., *et al.* Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol* 2016;17(1):148.

Hart, T., *et al.* Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Mol Syst Biol* 2014;10:733.

Hart, T., *et al.* High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell* 2015;163(6):1515-1526.

Hart, T. and Moffat, J. BAGEL: a computational framework for identifying essential genes from pooled library screens. *BMC Bioinformatics* 2016;17:164.

Horlbeck, M.A., *et al.* Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *Elife* 2016;5.

Hsu, P.D., *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol* 2013;31(9):827-832.

Jiang, P., *et al.* Network analysis of gene essentiality in functional genomics experiments. *Genome Biol* 2015;16:239.

Jinek, M., *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 2012;337(6096):816-821.

Knight, S.C., *et al.* Dynamics of CRISPR-Cas9 genome interrogation in living cells. *Science* 2015;350(6262):823-826.

Koike-Yusa, H., *et al.* Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nat Biotechnol* 2014;32(3):267-273.

Kolde, R., *et al.* Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* 2012;28(4):573-580.

Korkmaz, G., *et al.* Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nat Biotechnol* 2016;34(2):192-198.

Li, W., *et al.* Quality control, modeling, and visualization of CRISPR screens with MAGeCK-VISPR. *Genome Biol* 2015;16:281.

Li, W., *et al.* MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol* 2014;15(12):554.

Love, M.I., Huber, W. and Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;15(12):550.

Mali, P., *et al.* CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat Biotechnol* 2013;31(9):833-838.

Meyers, R.M., *et al.* Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat Genet* 2017;49(12):1779-1784.

Morgens, D.W., *et al.* Genome-scale measurement of off-target activity using Cas9 toxicity in high-throughput screens. *Nat Commun* 2017;8:15178.

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* 2016.

Ogata, T., Kozuka, T. and Kanda, T. Identification of an insulator in AAVS1, a preferred region for integration of adeno-associated virus DNA. *J Virol* 2003;77(16):9000-9007.

Ong, S.H.*, et al.* Optimised metrics for CRISPR-KO screens with second-generation gRNA libraries. *Sci Rep* 2017;7(1):7384.

Parnas, O.*, et al.* A Genome-wide CRISPR Screen in Primary Immune Cells to Dissect Regulatory Networks. *Cell* 2015;162(3):675-686.

Sadelain, M., Papapetrou, E.P. and Bushman, F.D. Safe harbours for the integration of new DNA in the human genome. *Nat Rev Cancer* 2012;12(1):51-58.

Shalem, O.*, et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* 2014;343(6166):84-87.

Shi, J.*, et al.* Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. *Nat Biotechnol* 2015;33(6):661-667.

Wang, T.*, et al.* Identification and characterization of essential genes in the human genome. *Science* 2015;350(6264):1096-1101.

Wang, T.*, et al.* Genetic screens in human cells using the CRISPR-Cas9 system. *Science* 2014;343(6166):80-84.

Wang, T.*, et al.* Gene Essentiality Profiling Reveals Gene Networks and Synthetic Lethal Interactions with Oncogenic Ras. *Cell* 2017;168(5):890-903.e815.

Xu, H.*, et al.* Sequence determinants of improved CRISPR sgRNA design. *Genome Res* 2015;25(8):1147-1157.

Zhou, Y.*, et al.* High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. *Nature* 2014;509(7501):487-491.

Zhu, S.*, et al.* Genome-scale deletion screening of human long non-coding RNAs using a paired-guide RNA CRISPR-Cas9 library. *Nat Biotechnol* 2016.