Downloaded from genome.cshlp.org on June 14, 2015 - Published by Cold Spring Harbor Laboratory Press

Sequence determinants of improved CRISPR sgRNA design

Han Xu[1,4,*], Tengfei Xiao[1,2,4,*], Chen-Hao Chen[5], Wei Li[1,4], Clifford A. Meyer[1,4], Qiu Wu[1,4,6], Di Wu[7], Le Cong[8], Feng Zhang[8], Jun S. Liu[7], Myles Brown[2,3,4,§], X. Shirley Liu[1,4,§]

[1]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, MA 02115, USA

[2]Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA 02115, USA.

[3]Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02215, USA

[4]Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, MA 02215, USA

[5]Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA 02215, USA

[6]Department of Bioinformatics, School of Life Science and Technology, Tongji University, Shanghai, 200092, China

[7]Department of Statistics, Harvard University, Cambridge, MA 02138, USA

[8]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, MA 02142, USA, and McGovern Institute for Brain Research, Department of Brain and Cognitive Sciences, Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

* Equal contributions

§ Correspondence to: myles_brown@dfci.harvard.edu (M.B.); xsliu@jimmy.harvard.edu (X.S.L.)

Key words: CRISPR/Cas9 knockout, CRISPR/dCas9 inhibition and activation, single guide RNA.

Downloaded from genome.cshlp.org on June 14, 2015 - Published by Cold Spring Harbor Laboratory Press

ABSTRACT

The CRISPR/CAS9 system has revolutionized mammalian somatic cell genetics. Genome-wide functional screens employing CRISPR/Cas9-mediated knockout or dCas9 fusion-mediated inhibition/activation (CRISPRi/a) are powerful techniques for discovering phenotype-associated gene function. We systematically assessed the DNA sequence features that contribute to single guide RNA (sgRNA) efficiency in CRISPR-based screens. Leveraging the information from multiple designs, we derived a new sequence model for predicting sgRNA efficiency in CRISPR/Cas9 knockout experiments. Our model confirmed known features, and suggested new features including a preference for cytosine at the cleavage site. The model was experimentally validated for sgRNA-mediated mutation rate and protein knockout efficiency. Tested on independent datasets, the model achieved significant results in both positive and negative selection conditions, and outperformed existing models. We also found that the sequence preference for CRISPRi/a is substantially different from that for CRISPR/Cas9 knockout and propose a new model for predicting sgRNA efficiency in CRISPRi/a experiments. These results facilitate the genome-wide design of improved sgRNA for both knockout and CRISPRi/a studies.

Downloaded from genome.cshlp.org on June 14, 2015 - Published by Cold Spring Harbor Laboratory Press

INTRODUCTION

The clustered regularly interspaced short palindromic repeats (CRISPR) system allows efficient perturbation of gene regulation in somatic cells and is increasingly being used for genome-wide functional screens. The CRISPR-based screens fall into two categories: Cas9-mediated knockout, and dCas9 fusion-mediated inhibition (CRISPRi) or activation (CRISPRa). In a CRISPR/Cas9 knockout system, the Cas9 nucleases induce double-strand breaks on targeted genomic regions directed by single-guide RNAs (sgRNAs). Indels produced at these breakpoints, as a result of an error-prone DNA repair mechanism, cause loss-of-function gene knockouts at the DNA level (Jinek et al. 2012; Mali et al. 2013; Yang et al. 2013). Alternatively, in the CRISPRi/a systems, catalytically inactive Cas9 (dCas9) proteins guided by sgRNAs enable the localization of effector domains to repress or activate gene transcription without modifying DNA (Qi et al. 2013; Gilbert et al. 2013). CRISPR/Cas9 knockout screens are better than previous functional shRNA knockdown screens (Berns et al. 2004; Luo et al. 2008) for the study of phenotypes that require complete inactivation of genes (Shalem et al. 2014). CRISPRi/a are more flexible than shRNA in defining gene functions across a wider dynamic range (Gilbert et al. 2014). In light of this, several sgRNA libraries with diverse configurations have been developed for CRISPR/Cas9 knockout or CRISPRi/a screens in human and mouse (Wang et al. 2014; Shalem et al. 2014; Koike-Yusa et al. 2014; Zhou et al. 2014; Gilbert et al. 2014; Konermann et al. 2015).

The design of sgRNA is critical to the CRISPR-based screens. In a CRISPR system, an sgRNA contains a spacer sequence complementary to the targeted DNA sequence to guide the Cas9 or dCas9 proteins to genomic targets. Several studies have investigated sgRNA off-target effects based on the alignment of spacer sequences to the genome (Hsu et al. 2013), dCas9 ChIP-seq (Wu et al. 2014) and GUIDE-seq (Tsai et al. 2015). On the other hand, DNA sequence also influences the efficiency of sgRNA at the on-target genomic loci. It is well established that the Protospacer Adjacent Motif (PAM) containing NGG consensus is required for Cas9-DNA binding and cleavage in a CRISPR/Cas9 system (Sternberg et al. 2014). A recent study showed that the sgRNA efficiency depends on the purine/pyrimidine composition near the 3' end of the spacer sequence (Wang et al. 2014). Moreover, Doench et al. found the nucleotide composition of the DNA downstream of the spacer target also contributes to the sgRNA efficiency, and proposed a predictive sequence model for the design of CRISPR/Cas9 knockout experiments (Doench et al. 2014).

Despite these advances, it is unclear whether the sequence features reported in previous studies are reproducible across different sgRNA library designs and genomes. Importantly, these features were only tested on negative selections using CRISPR/Cas9 knockout systems, where loss-of-function of

Downloaded from genome.cshlp.org on June 14, 2015 - Published by Cold Spring Harbor Laboratory Press

a gene results in growth disadvantage to the cells. For model justification and broader applications, it is necessary to investigate their relevance to positive selections where gene perturbation promotes cell growth and proliferation. Moreover, no sequence model has been developed for predicting sgRNA efficiency in CRISPRi/a screens. To fill these gaps, we systematically assessed the effect of sequence context on sgRNA efficiency based on 6 published datasets (Table 1). Our analysis led to the development of two computational models for predicting sgRNA efficiency, corresponding to the CRISPR knockout and CRISPRi/a studies, respectively. To this end, our findings and models facilitate the genome-wide design of improved sgRNA for CRISPR-based screens.

RESULTS

Selection of the efficient and inefficient sgRNAs in CRISPR/Cas9 knockout screens

To construct a predictive model, we took two steps to identify efficient and inefficient sgRNAs in published datasets. First, for each dataset, we identified a list of essential genes whose deletion resulted in a growth disadvantage in genome-wide CRISPR/Cas9 knockout screens. Second, the sgRNAs targeting these essential genes were grouped into “efficient” and “inefficient” categories based on their decline in abundance in the screens. We applied our method to Wang and Koike-Yusa data (see Table 1). These datasets have large numbers of sgRNAs per gene, which help to identify reliable sets of essential genes.

The procedures of sgRNA selection and categorization are depicted in Figure 1. In Wang data, we identified 305 and 545 essential genes in the leukemia cell lines HL-60 and KBM-7, respectively (Fig. 1A). Among them, 58 ribosomal and 163 non-ribosomal genes were essential in both cell lines. The essentiality of ribosomal genes was also reported in other studies (Shalem et al. 2014; Koike-Yusa et al. 2014), which established them as good positive controls in a CRISPR/Cas9 knockout screen. In the Wang library, each ribosomal gene was targeted by more than 20 sgRNAs and each non-ribosomal gene was targeted by 10 sgRNAs. Therefore we split the sgRNAs targeting common essential genes into “ribosomal” and “non-ribosomal” sets in order to take the difference of sgRNA sample space into consideration. We examined signals of sgRNAs by comparing their abundance before and after selection. For the “ribosomal” set, 731 (35.3%) sgRNAs showed more than 2-fold depletion in both cell lines, indicating that they reproducibly knocked out their target genes in the experiments, thus were defined to be efficient. 438 (21.1%) failed to show significant differences from control sgRNAs that do not target any genomic locus, thus were defined to be inefficient. The remaining sgRNAs showed relatively weak effects in both cell lines or had significant effects in only one cell line, and were excluded from our analysis (Fig. 1B). For the “non-ribosomal” set, we

Downloaded from genome.cshlp.org on June 14, 2015 - Published by Cold Spring Harbor Laboratory Press

selected 671 efficient and 237 inefficient sgRNAs using the same criteria (Fig. 1C). In Koike-Yusa data, we identified 311 genes that showed essentiality in two biological replicates of mouse Embryonic Stem Cells (mESCs) (Fig. 1D). Upon examining the signals of sgRNAs targeting these genes, we found a bimodal distribution of the relative sgRNA abundance in cell growth (Fig. 1E and Supplementary Fig. 1). Categorization based on the bimodality resulted in 830 efficient and 234 inefficient sgRNAs.

To this end, we selected 3 training sets of sgRNAs, namely “ribosomal”, “non-ribosomal”, and “mESC”, for the construction of sequence models (Supplementary Table 1). The differences among these sets reflect variations of sgRNA design, species, and the experimental protocols.

Sequence features associated with sgRNA efficiency in CRISPR/Cas9 knocking-out

To explore sequence features that contribute to sgRNA efficiency, we computed the log odds-ratio of nucleotide frequency between DNA sequences targeted by efficient and inefficient sgRNAs (Fig. 2A-2C). These sequences are 40-bp in length, including the 19-bp or 20-bp spacer targets as well as their 3’ and 5’ flanking DNAs. All sequences were aligned at the PAM. The log odds-ratios are correlated between the ribosomal and non-ribosomal sgRNA sets (Fig. 2E), as well as between the two libraries developed independently (Fig. 2F). The correlations indicate that many sequence features are robust against the variation of sgRNA designs, species and the spacer length. Meanwhile, some features are discordant among the three different sgRNA training sets. For example, guanine is preferred at the 5’ end of the spacer in the “ribosomal” and “non-ribosomal” sets, but not in the “mESC” set. This might be ascribed to subtle differences in sgRNA design. If the 5’ end of spacer is not “G”, Wang et al. prepended a “G” for the expression of sgRNAs from a U6 promoter, in the form of G(A/C/T)X_{19}NGG; otherwise a form of GX_{19}NGG was taken without any prepended “G”. Alternatively, Koike-Yusa et al. chose the form of GX_{19}NGG regardless of nucleotide composition at the 5’ end. These different designs of sgRNAs can result in the variation of the nucleotide pairings and 5’ overhang in sgRNA-DNA binding structure, which could further impact sgRNA efficiency.

To determine the reproducible features, we first selected nucleotides that satisfy the following criteria across the 3 training sets: i) signs of the odds-ratios are concordant; ii) magnitudes of the odds-ratios are above a threshold in all three sgRNA sets, where the threshold was computed from a statistical significance analysis (see METHODS section). To identify the dominating nucleotide features, we next applied Elastic-Net, a penalized linear regression model with proven utility in feature selection (Zou and Hastie 2005). In our analysis, the predictor variable of Elastic-Net was a binary vector representing the presence or absence of the nucleotides selected based on the odds-ratio, and the

Downloaded from genome.cshlp.org on June 14, 2015 - Published by Cold Spring Harbor Laboratory Press

response variable was 1 or -1 corresponding to the efficient and inefficient sgRNAs, respectively. The union of the three sgRNA training sets was used for Elastic-Net feature selection and parameter estimation.

With this two-stage feature selection approach, we identified 28 sequence features, as shown in Figure 2D. Most features are located within the spacer region. Our result confirmed several features reported previously. First, guanines are strongly preferred at the -1 and the -2 positions proximal to the PAM sequence, which are associated with the sequence preference in Cas9 loading (Wang et al. 2014). Second, thymines are disfavored at the four positions closest to the PAM, consistent with the fact that multiple uracils in the spacer cause low sgRNA expression (Wu et al. 2014). Third, in line with recent findings, the nucleotides downstream of the PAM contribute to sgRNA efficiency, whereas the sequences upstream of the spacer have no significant effect (Doench et al. 2014). We also identified novel features that reproducibly impact the sgRNA efficiency. For example, cytosine is preferred at the -3 position, which is the DNA cleavage site by the CRISPR/Cas9 complex (Cong et al. 2013). This feature might contribute to the efficiency of cleavage or the introduction of mutation upon DNA repair. Moreover, adenines are preferred from position -5 to -12, and guanines are preferred at positions -14 to -17.

Experimental validation of the sequence model in predicting mutation rates and protein knockout efficiency

To experimentally test the model in Figure 2D, we first assessed the mutation rates mediated by a small set of sgRNAs targeting the *AAVS1* safe harbor genomic locus. We selected 10 targets with low efficiency scores and 10 with high scores in the *AAVS1* region (Supplemental Table 3), and designed CRISPR/Cas9 lentivirus based on the selected targets. Upon examining the transduced 293T cells, we found that sgRNAs with higher scores showed much higher indel mutation rates than those with lower scores (p=1.9E-5 in *t*-test, Fig. 3A). This confirmed that the sequence features we selected have significant effect on the rate of on-target mutation in a CRISPR/Cas9 system.

We next tested if the sequence model is informative to predict knockout efficiency on the protein expression level. We designed sgRNAs against *AR* and *FOXA1*, and tested them in prostate cancer LNCaP-abl cells using Western-blot to assay protein expression (Supplementary Fig. 2 and Supplementary Table 4). The predicted sgRNA efficiency score showed strong positive correlation with the protein knockout efficiency (p=9.0E-6 in Pearson correlation test, Fig. 3B), where the efficient and inefficient sgRNAs can be clearly separated. Our result suggested a cutoff of zero as a reasonable threshold for sgRNA classification, in line with the configurations in model training where the values of 1 and -1 were assigned to efficient and inefficient sgRNAs, respectively.

Downloaded from genome.cshlp.org on June 14, 2015 - Published by Cold Spring Harbor Laboratory Press

Predictive power of the sequence model in negative selection screens

With the satisfactory results in the low-throughput validation mentioned above, we next assessed the predictive power of the sequence model using published high-throughput CRISPR/Cas9 knockout screening data. We designed four schemes for this *in silico* validation. First, randomized 3-fold cross validation was carried out on the categorized sgRNAs in the ribosomal set to assess the predictive power under identical experimental settings; second, the Elastic-Net model was trained on the ribosomal set and tested on the non-ribosomal set, of which the result was expected to reflect the within-library performance under different configurations of sgRNA sampling; third, to evaluate the inter-library and inter-species performance, the model was trained on the union of ribosomal and non-ribosomal sets in human and tested on the mESC set; last, the model was trained on the union of all three sets, and tested on an independent library of sgRNAs (Shalem et al. 2014). The sgRNA list used in the fourth validation was compiled by Doench et al. to assess their model in predicting sgRNA "activity", a measure analogous to the "efficiency" defined in this paper (Doench et al. 2014). We achieved reasonable predictive power in all four validations, in which the Area Under Curve (AUC) scores were greater than 0.7 in Receiver Operating Characteristic (ROC) (Fig. 4A and 4B). Our model also outperformed the model proposed by Doench et al. in predicting sgRNA activity (Fig. 4B). With a cut-off threshold of zero, 50-60% of the inefficient sgRNAs are predictable, at the cost of 10-20% of efficient sgRNAs misclassified.

Since the sequence model was trained based on the sgRNAs targeting common essential genes, we further tested if the model is applicable to the sgRNAs targeting cell-type-specific essential genes. We examined the correlations between sequence scores and the relative sgRNA abundances for *ABL1* and *BCR*, two genes coding for an oncogenic fusion protein in KBM-7 cells(Andersson et al. 1995). For both genes, significant correlations were observed between predicted sgRNA efficiency and the changes of sgRNA abundance in the screen (Figure 4C). To evaluate whether sequence-specific sgRNA efficiency also influences sgRNAs targeting non-essential genes, we compared distributions of the correlations for essential and non-essential genes identified in KBM-7 (Fig. 4D) or HL-60 (Supplementary Fig. 3). As expected, sequence scores and relative sgRNA abundances are in strong negative correlation for essential genes. Meanwhile, there are weak but statistically significant negative correlations for non-essential genes, which might be attributed to: (i) the false negatives in identifying essential genes; (ii) decreased cell growth due to intrinsic toxicity of DNA cleavage, similar to the toxicity reported on Cre recombinase (Silver and Livingston 2001); and/or (iii) potential off-target effects of sgRNAs. Notably, no correlation was observed for those control sgRNAs lacking genomic targets (Supplementary Fig. 4), consistent with the requirement of chromatin binding for an sgRNA to exert its cell functions.

Downloaded from genome.cshlp.org on June 14, 2015 - Published by Cold Spring Harbor Laboratory Press

We next checked if a proper selection of sgRNAs based on the prediction result could refine the sgRNA pool. As shown in Figure 4E, the signals of essential and non-essential genes are not well separable with the predicted inefficient sgRNAs. In contrast, the efficient sgRNAs produced highly informative signals that distinguish essential genes from others. This indicates that the removal of inefficient sgRNAs can reduce the number of required sgRNAs without compromising the sensitivity of the CRISPR/Cas9 knockout screens.

Generality of the sequence model in positive selection screens

Our model was trained based on the sgRNAs targeting essential genes that were negatively selected in the experiments. We next asked if the model is applicable to positive selection screens. To answer this, we curated 4 lists of genes involved in the resistance to drug or toxin treatments, including the BRAF inhibitor Vemurafenib (PLX) in melanoma A375 cells (Shalem et al. 2014), alpha-toxin and 6-thioguanine (6TG) in mouse embryonic stem cells (Koike-Yusa et al. 2014), and anthrax toxin in Hela cells (Zhou et al. 2014) (Supplementary Table 5). The sgRNAs targeting these genes were categorized based on the reproducibility of positive selection, *i.e.* the increase of the sgRNA abundance in multiple biological replicates upon treatments. As shown in Figure 5A-5E, the predicted efficient sgRNAs expressed higher reproducibility of positive selection in all the experiments, indicating our model's applicability to positive selection screens. Next we assessed the predictive power of our model and the model by Doench et al. on the curated sgRNA sets, where the sgRNAs "selected in all replicates" were defined as positives and the sgRNAs "not selected in any replicates" were defined as negatives. Our model achieved an AUC score of 0.711, significantly outperformed the Doench et al. model (Fig. 5F). Notably, 42.3% of the efficient sgRNAs contain a cytosine at the cleavage site, in comparison with 25.9% for inefficient sgRNAs, suggesting this novel feature also contributes to sgRNA efficiency in positive selections.

Our prediction is therefore consistent with the phenotype in both positive and negative selections, where the abundance of efficient sgRNAs is expected to change in opposite directions. This line of evidence, together with our validation results on mutation rate and protein knockout efficiency, confirmed that the selected sequence features impact cell phenotype mainly through the loss-of-function of the target genes, but not through their intrinsic toxicity on cell growth.

Sequence-specific sgRNA efficiency in CRISPRi/a screens

Recently, Gilbert et al. proposed to use a CRISPRi/a system for genome-wide functional screens (Gilbert et al. 2014). To explore the impact of sequence context on the efficiency of sgRNAs in their

Downloaded from genome.cshlp.org on June 14, 2015 - Published by Cold Spring Harbor Laboratory Press

study, we selected 6,624 sgRNAs targeting the 500 most essential genes in the published CRIPSRi negative selection data in K562 cells (Supplementary Table 6). These sgRNAs were grouped into "efficient" and "inefficient" based on the phenotype scores that measure the relative sgRNA abundance (Fig. 6A).

Since the spacers have variable lengths in this dataset, we first examined the percentage of efficient sgRNAs with different spacer lengths. We found sgRNAs with a spacer length of 19nt have the highest efficiency, slightly better than those with 20nt spacers (Fig. 6B). Despite the previous report that truncated sgRNAs with 17-18nt spacers have less off-target effects (Fu et al. 2014), shorter spacers showed less on-target effect in comparison with 19nt or 20nt spacers. To remove the confounding effect of spacer length, we computed log odds-ratio of nucleotide frequency for the efficient and inefficient sgRNAs based on different spacer lengths (Fig. 6C). Similar to CRISPR/Cas9 knockout, the CRISPRi/a system prefers purines at most of the nucleotides in the spacer. However, there are also significant differences. In CRISPR/Cas9, a small number of nucleotides, mostly adjacent to the PAM, dominate the sequence preference (Fig. 2A-D). In contrast, many nucleotides throughout the spacer region collectively contribute to sgRNA efficiency in the CRISPR/dCas9 system. The disparity can be ascribed to several factors. First, unlike CRISPR/Cas9 knockout, CRISPRi/a does not introduce DNA cleavage, which explains the observation that no cytosine enrichment at the -3 position. Second, the structural and functional differences between Cas9 and dCas9 can lead to diverse sequence preference for protein loading and the initiation of gRNA-DNA pairing. Third, the effector domains that play crucial roles for gene perturbation in CRISPRi/a do not exist in a CRISPR/Cas9 system. Fourth, the U6 promoter and the sgRNA constant region are different between the Gilbert et al. design and other platforms (Chen et al. 2013). Last by not least, CRISPRi/a mainly targets gene promoters where the sequence contexts are substantially different from those in coding regions.

To evaluate the sequence-specific sgRNA efficiency in CRISPRi/a, we designed linear classifiers using the Elastic-Net, and validated the classifiers with four settings: i) 3-fold cross-validation on the sgRNAs targeting negatively selected essential genes in the CRISPRi experiment; ii) training on the essential genes, and testing on the positively selected genes upon CTx-DTA treatment; iii) training on the essential genes in the CRISPRi experiment, and testing on the growth-inhibiting genes in CRISPRa data (Gilbert et al. 2014); iv) repeating the training process in iii) and testing in Konermann CRISPRa data (Konermann et al. 2015). We achieved ROC-AUC scores of 0.6-0.65 in the four validations, which indicated that the sequence model is applicable to positive selections, CRISPRa experiments, as well as to the CRISPRi/a systems of independent designs (Fig. 6D). There is less predictive power when applying the CRISPR/Cas9 knockout model on CRISPRi/a data (ROC-AUC < 0.6). Notably, comparing with the higher predictive power in CRISPR/Cas9 knockout screens (ROC-

Downloaded from genome.cshlp.org on June 14, 2015 - Published by Cold Spring Harbor Laboratory Press

AUC > 0.7), we found the sgRNA efficiency in CRISPRi/a to be less dependent on the sequence context. It is possible that transcription factor binding and chromatin states at the target sites of CRISPRi/a influence knockdown efficiency, which cannot be explained by the sequence context alone.

DISCUSSION

We have systematically assessed the contributions of sequence features to sgRNA efficiencies, which are critical to the performance of genome-wide CRISPR based functional screens. To achieve sufficient statistical power, at least hundreds of sgRNA samples are needed for the analysis (see supplementary material). Therefore we sought to collect samples from published high-throughput screening datasets. In the study of CRISPR/Cas9 knockout screens, we used an integrative method to select sequence features that are reproducible among different datasets. Such a method is helpful to improve the robustness of the sequence model for the prediction in various applications. In addition to known features reported previously, our model suggested novel features to better understand the molecular mechanism of the CRISPR/Cas9 genome-editing systems. We demonstrated the predictive power of our model through multiple validations including tests on independent designs, various cell types, and different growth selections, which support the robustness of the model for customized sgRNAs designs.

Despite these findings, ~40% of inefficient sgRNAs are not predictable with our sequence model, which can be ascribed to sgRNA secondary structure, chromatin structure, exon-skipping during transcription, strand and location (e.g. 5' vs 3' end) of sgRNA target relative to the gene, possibly nonlinear interactive effects among different positions, as well as the sgRNA-specific likelihood of a frame shift (Wang et al. 2014; Doench et al. 2014). The performance of the model might be further improved by including these other features in the future. Another limitation of our study is the relatively small number of cell lines (5 for CRISPR/Cas9 KO and 2 for CRISRPi/a) and library designs for the training and testing of our computational model. It is possible that some cell-specific sequence preferences were missed in our study. With the expected growth of CRISPR datasets, more comprehensive and systematic assessment could be helpful to refine the predictive model.

It is worth noting that selection against sgRNAs targeting non-essential genes is also dependent on the sgRNA efficiency in a moderate but statistically significant manner. This can be partially explained by false negatives in identifying essential genes and the potential off-target effects of sgRNAs. On the other hand, it has been reported that the DNA cleavage induced by Cre recombinase results in growth retardation of cells that is independent of the targeted genomic loci (Silver and Livingston 2001). Therefore, it is possible that the on-target DNA cleavage introduced by

Downloaded from genome.cshlp.org on June 14, 2015 - Published by Cold Spring Harbor Laboratory Press

CRISPR/Cas9 system might have similar toxicity on cell growth. Future work is needed to validate this hypothesis, and to normalize the CRISPR/Cas9 screen data against the sgRNA toxicity and efficiency.

In this study, we found that the sequence preference in CRISPRi/a is significantly different from that in CRISPR/Cas9 knockout. This is expected since the CRISPRi/a system uses a distinct mechanism to perturb gene function without introducing DNA cleavage. We note that our analysis is based on published datasets where coding regions were chosen for CRISPR/Cas9 targeting and promoters were chosen for CRISPRi/a targeting. The difference of nucleotide composition between the coding regions and the promoters, such as codon structure and GC enrichment, might introduce bias to our analysis. It would be interesting to compare these two systems with an unbiased design in future study. Despite of this, our validation result on *AAVS1* locus suggests that the CRISPR/Cas9 knockout model is also applicable to non-coding regions.

The models proposed in this study focus on sgRNA sensitivity. Meanwhile, the specificity of sgRNAs should also be considered when designing an sgRNA pool. Wu et al. observed numerous off-target binding sites at the genomic loci with sufficient sequence similarity to the sgRNA spacer, and reported that thymine-rich sgRNA spacers have fewer off-target sites than others due to the low level of sgRNA expression and the scarcity of complementary sites in accessible chromatin (Wu et al. 2014). This finding, however, is not contradictory to our model showing disfavored thymine near the 3' end of sgRNA target, but rather reflects the requirement of an appropriate trade-off between specificity and sensitivity in sgRNA design. Surprisingly, a recent study showed that majority of off-target cleavage sites are undetectable using existing computational methods or dCas9 ChIP-seq (Tsai et al. 2015). Further studies are needed to address both sgRNA efficiency and off-target effect for optimal sgRNA designs.

METHODS

Pre-processing

Suppose there are N experiments performed on a set of M sgRNAs, and the read count of sgRNA i on experiment j is x_{ij}, $1 \leq i \leq M, 1 \leq j \leq N$. To normalize against the sequencing depth and the global trend of cell growth, we assume that majority of the sgRNA-targeted genes have no effect on cell growth. Therefore we adjusted read count $x'_{ij} = x_{ij}/s_j$, where s_j is the median read count of all sgRNAs in the jth experiment. This median normalization approach was applied to Wang, Koike-Yusa, Shalem and Konermann data. For Zhou data, the median read count was zero in the

Downloaded from genome.cshlp.org on June 14, 2015 - Published by Cold Spring Harbor Laboratory Press

experiments after positive selection since the strength of selection was very strong so that majority of the sgRNAs were completely depleted except a few that contribute to drug resistance. In this case, we set the normalization factor s_j to be the average read count of all sgRNAs. Normalization in such way might introduce bias to the estimation of relative sgRNA abundance, but will not affect the conclusion drawn from Figure 5E since the bias is a constant added to all sgRNAs.

Next, we compute the relative sgRNA abundance between two experiments. Suppose x'_{ij} and x'_{ik} are the adjusted read counts of the ith sgRNA in the jth and kth experiments, the relative sgRNA abundance is computed as the log2-transformed ratio:

$$r_{ijk} = log_2\left(\frac{x'_{ij} + \epsilon}{x'_{ik} + \epsilon}\right)$$

where ϵ is a small positive value to avoid large variation due to low read counts, and was set to be 0.05.

For negative selection in Wang data, the relative sgRNA abundance was computed by comparing the cell population after 12 cell doublings against the initial seeding states in HL-60 and KBM-7 cells respectively. For negative selection in Koike-Yusa's dataset, the transfected mESCs cultured for 6-7 days were compared against plasmid control. For positive selection in Koike-Yusa data, the mESCs treated with toxin were compared against plasmid control. For positive selection in Shalem data, A375 cells with PLX treatment were compared against the population without PLX treatment at day 7 and day 14. For positive selection in Zhou data, Hela cells after three rounds of toxin treatment were compared against plasmid control. For negative selection in Konermann CRISPRa data, A375 cells with DMSO treatment were compared between day 21 and day 3. The negative selection data by Shalem et al. and the CRISPRi/a data by Gilbert et al. were not pre-processed, as the phenotypic measure of the sgRNAs were retrieved directly from their studies.

Selecting phenotype-associated genes

For the CRISPR/Cas9 knockout datasets, we ran the beta-0.1 version of RIGER (http://www.broadinstitute.org/cancer/software/rnai/riger/) to call the negatively selected genes (Luo et al. 2008). In the configurations of RIGER, the measures of relative sgRNA abundance were used to rank the sgRNAs, and Kolmogorov-Smirnov test was chosen to determine the significant genes. The threshold of p-values was set to be 0.001. The positively selected genes in Figure 5 were collected based on experimental validations in the original studies (Shalem et al. 2014; Koike-Yusa

Downloaded from genome.cshlp.org on June 14, 2015 - Published by Cold Spring Harbor Laboratory Press

et al. 2014; Zhou et al. 2014), and a cut-off threshold of 2 folds in relative sgRNA abundance were used to determine efficient and inefficient sgRNAs. The phenotype-associated genes in the CRISPRi/a dataset were retrieved from the original paper (Gilbert et al. 2013).

Thresholds for identifying efficient and inefficient sgRNAs

As shown in Figures 1B, 1C, and 1E, we used different thresholds for identifying efficient and inefficient sgRNAs in Wang data and in Koike-Yusa data. This is because the sgRNAs targeting essential genes in Wang data and Koike-Yusa data have different distributions of log-ratios, due to the variance of cell growth rate, Cas9 activity, and the depth of selection in the experiments. In Wang data, the sgRNAs with a log-ratio smaller than -1.0 in both HL-60 and KBM-7 were identified to be efficient, and the thresholds for inefficient sgRNAs are the 5 percentile of the log-ratios of control sgRNAs in HL-60 and KBM-7. The thresholds in Koike-Yusa data were determined based on the bimodal distribution of signals (Supplementary Fig. 1). Notably, when the thresholds in Koike-Yusa data were set to be identical to those used in Wang data, we observed little variation of nucleotide frequency ($r = 0.96$). Therefore, our sequence model is insensitive to the threshold chosen.

Extracting sequence features

For Wang, Koike-Yusa and Shalem data, sequences of spacer targets and their flanking regions were extracted from hg19 or mm9 genome assembly. Since the length of sgRNA spacers varies from 18 bps to 22 bps in Zhou et al.'s platform, we mapped the spacer sequences to hg19 assembly to find the genomic loci of the targets, and extracted the 40-bp sequence (spacer + 3' + 5') based on the aligned PAM loci. The sequences in CRISPRi/a data were retrieved from the original paper (Gilbert et al. 2014; Konermann et al. 2015).

Determining reproducible features in CRISPR/Cas9 knockout screens

To measure the sequence preference, we computed the log2 odds-ratio of nucleotide frequency between efficient and inefficient sgRNAs, for the "ribosomal", "non-ribosomal" and "mESC" sets. Let s_i^1, s_i^2, and s_i^3 represents the log2 odds-ratios for the ith nucleotide computed from the 3 sgRNA sets, we define a score s_i as:

$$s_i = \begin{cases} sgn(s_i^1) \times \min(|s_i^1|, |s_i^2|, |s_i^3|) & if\ sgn(s_i^1) = sgn(s_i^2) = sgn(s_i^3) \\ 0 & else \end{cases}$$

where $sgn(s)$ is the sign of s, and $|s|$ is the magnitude of s.

Downloaded from genome.cshlp.org on June 14, 2015 - Published by Cold Spring Harbor Laboratory Press

To determine the background distribution of s_i, we randomly permuted the sgRNAs in the efficient and inefficient categories within each set, and computed s_i based on the permuted data. This results in an empirical null distribution of s_i. A threshold T were then set to be at the 95% confidence interval of the null distribution, and the nucleotides with $|s_i| > T$ were selected as reproducible features.

The Elastic-Net model

Suppose $\boldsymbol{X} = \{X_1, X_2, \ldots, X_N\}$ is the set of encoded sequence vectors and $\boldsymbol{Y} = \{y_1, y_2, \ldots, y_N\}$ is the set of outputs representing the efficiency of sgRNAs, where N is the number of sgRNA samples for training. Let M be the length of the input vectors, the Elastic-Net model computes the parameters $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_M]^T$ that minimize an objective function E:

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

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.

When training the Elastic-Nets for sequence feature selection, we found the optimal α to be 1.0 for CRISPR/Cas9 knockouts and 0.0 for CRISPRi/a (Supplementary Figures 5 and 6). We note that the Elastic-Net is equivalent to the Least Absolute Shrinkage and Selection Operator (LASSO) when $\alpha = 1.0$ (Tibshirani 1996), and is equivalent to the Ridge regression model when $\alpha = 0.0$ (Tikhonov 2014). In practice, the LASSO model is preferred to select a small number of features from a large amount of candidates, whereas the Ridge regression is more relevant in situations when many features cumulatively contribute to the response. Therefore, our result on the selection of optimal α is explainable since the sgRNA efficiency depends on fewer sequence features in Cas9 knockout in comparison with that in CRISPRi/a.

Application of Doench et al.'s model

To apply Doench et al.'s model (Doench et al. 2014), we downloaded the python script at http://tefor.net/crispor/doenchScore.py. For fair comparison, the tested sgRNA targets and our regression model (Fig. 2D) were truncated into 30-bp sequence harboring the spacer target, following the requirement in Doench et al.'s model.

Cell lines and cell culture

Downloaded from genome.cshlp.org on June 14, 2015 - Published by Cold Spring Harbor Laboratory Press

LNCaP-abl (abl) cell line was provided by Zoran Culig (Innsbruck Medical University, Austria). The abl cells were cultured in the RPMI 1640 phenol red-free medium supplemented with 10% charcoal/dextran- treated fetal bovine serum, 2mM glutamine, 100 ug/ml penicillin and 100units/ml streptomycin for the experiments. The 293T cells obtained from the American Type Culture Collection were maintained in DMEM media supplemented with 10% fetal bovine serum, glutamine and penicillin-streptomycin.

Plasmid construction, lentivirus production and transduction into human cells

20 sgRNAs targeting *AAVS1* locus, 9 sgRNAs targeting *AR* gene, and 6 sgRNAs targeting *FOXA1* gene were designed and selected according to our prediction of sgRNA efficient scores. The sgRNA oligos were cloned into lentiCRISPRv2 plasmid (Addgene) as previously described (Shalem et al. 2014). Each plasmid containing inserted sgRNA sequence was verified using Sanger sequencing. To make lentivirus, the constructed lentiCRIPSR v2 plasmids were co-transfected into 293T cells with packaging plasmids pMD2.G and psPAX2 (Addgene) using X-tremeGENE HP DNA Transfection Reagent (Roche) in 12 well plates as manufacturer's instruction. These 35 types of packaged lentivirus were then transduced into 293T cell and LNCaP-abl cell respectively using 24 well plates, followed by puromycin selection for three days. The survived cells were maintained for another three days before isolation of the DNAs and proteins.

SURVEYOR nuclease assay and Western Blot

The DNA fragments of *AASV1* locus were amplified from extracted genomic DNA by PCR using the Q5 high-fidelity DNA polymerases (NEB). SURVEYOR nuclease assays were subsequently performed as manufacturer's instruction (Trans genomic). The indel mutation rates were calculated as previously described(Ran et al. 2013). Western Blot for *AR* and *FOXA1* knockout in LNCaP-abl cells by different sgRNAs carried out using whole cell lysates as described previously(Xu et al. 2012), The antibody used for immunoprecipitation was anti-AR (N-20, Santa Cruz Biotechnology) and anti-FOXA1 (ab23738, Abcam), or anti-GAPDH (sc-25778, Santa Cruz Biotechnology). The images were analyzed using ImageJ (Schneider et al. 2012).

Software availability

A software package named Spacer Scoring for CRISPR (SSC) was developed for predicting sgRNA efficiency from genomic sequences. The source codes of SSC are downloadable at

Downloaded from genome.cshlp.org on June 14, 2015 - Published by Cold Spring Harbor Laboratory Press

http://sourceforge.net/projects/spacerscoringcrispr/. A web-based application is available at http://crispr.dfci.harvard.edu/SSC

ACKNOWLEDGEMENTS

The authors would like to thank Drs. Carl Novina, Tim Wang, David Sabatini, Eric Lander, Ophir Shalem and Neville Sanjana for providing the datasets and helpful discussions. The project was supported by the NIH grant U01 CA180980 (to X.S.L.), R01 GM113242-01 (to J.S.L.), NSF grant DMS-1120368 (to J.S.L.), and the Claudia Adams Barr Award in Innovative Basic Cancer Research from the Dana-Farber Cancer Institute.

AUTHOR CONTRIBUTIONS

H.X. and X.S.L. designed the study. C.A.M., J.S.L, and M.B. provided conceptual advice. H.X., C.H.C, and D.W. developed the algorithm and performed the analyses. T.X. conducted experimental validations. W.L., T.X., L.C., F.Z. and M.B. helped with technical clarifications and result interpretations. H.X., T.X., and X.S.L. wrote the manuscript. All the authors participated in discussions and manuscript revision.

DISCLOSURE DECLARATION

The authors declare no competing financial interests.

Downloaded from genome.cshlp.org on June 14, 2015 - Published by Cold Spring Harbor Laboratory Press

FIGURE LEGENDS

Figure 1: A schematic view of procedures for sgRNA selection and categorization. (A,B) Venn diagram showing the overlap of essential genes between (A) human HL-60 and KBM-7 cells, and (B) two biological replicates in mouse ESC JM8 cells. **(C-E)** Scatter plots showing the log2 fold-change of sgRNA abundance in negative selection upon cell growth: (C) sgRNAs targeting essential ribosomal genes in Wang data, (D) sgRNAs targeting essential non-ribosomal genes in Wang data, and (E) sgRNAs targeting essential genes in Koike-Yusa data. The dashed lines represent the threshold chosen to determine efficient and inefficient sgRNAs.

Figure 2: Preference of nucleotide sequences that impact sgRNA efficiency. (A,B,C) Logos showing the sequence preference of the three sgRNA sets defined in Fig. 1. The height of the nucleotides represents the log odds-ratio of nucleotide frequency between efficient and inefficient sgRNAs. **(D)** A logo showing the selected features that reproducibly impact sgRNA efficiency in the three sgRNA sets. The height of the nucleotides represents the coefficients computed from the Elastic-Net. **(E,F)** Scatter plots showing the correlation of sequence preference for: (E) sgRNAs targeting ribosomal vs. non-ribosomal genes in Wang data, and (F) sgRNAs in Wang et al.'s platform vs. Koike-Yusa data. Each dot represents a nucleotide in a 40bp region centered by the spacer. The sequence preference is measured as the $\log_2$ odd-ratio of nucleotide frequency between efficient and inefficient sgRNAs.

Figure 3: Experimental validation of the sequence model in predicting sgRNA efficiency. (A) A SURVERYOR gel picture (top) and a bar-chart (bottom) showing the indel rates of the sgRNAs predicted to be inefficient (low sequence score) or efficient (high sequence score). The sgRNAs were selected to target the *AAVS1* locus. The experiment was conducted in 293T cells. **(B)** A scatter plot showing the correlation of the predicted sequence scores and the protein knockout efficiency for sgRNAs targeting *AR* and *FOXA1* in LNCaP-abl cells. The knockout efficiency is measured as the percentage of reduction in protein level upon sgRNA infection.

Figure 4: Predicting sgRNA efficiency from sequence context in CRISPR/Cas9 knockout screens. (A) ROC curves showing the predictive power of the proposed model. Red: 3-fold cross-validation on the sgRNAs targeting ribosomal genes in Wang data; Blue: trained on ribosomal genes, and tested on non-ribosomal genes in Wang data; Green: trained on Wang data, and tested on Koike-Yusa data. The black error bars on the red curve represent standard deviations computed from 10 iterations of random sampling in cross-validation. **(B)** ROC curves comparing the performance of the proposed model and Doench et al.'s model in predicting sgRNA efficiency in Shalem data. **(C)** Scatter plot showing the correlation between the predicted sequence score and the

Downloaded from genome.cshlp.org on June 14, 2015 - Published by Cold Spring Harbor Laboratory Press

relative sgRNA abundance for *ABL1* and *BCR* in KBM-7 cells. The p-values were computed based on Pearson correlation test. **(D)** Box-plot showing the distributions of correlations between sequence scores and relative sgRNA abundances, for essential and non-essential genes in KBM-7. The distribution of random background was computed by permuting the sequence scores within each gene in the dataset. **(E)** Distributions of relative sgRNA abundances in KBM-7 cells, where the sgRNAs were categorized based on the predicted efficiency and the essentiality of their targeted genes.

Figure 5: Assessment of the sequence models in predicting sgRNA efficiency in positive selection experiments. (A-E) Bar charts showing the capability of selection and the experimental reproducibility for predicted efficient and inefficient sgRNAs. The tested sgRNAs target the genes known to be involved in the resistance to different drug treatment or external stimulus. **(F)** ROC curves comparing the performance of the proposed model and Doench et al.'s model in predicting sgRNA efficiency in positive selection experiments. In the evaluation, the positive test set consists of the sgRNAs selected in all replicates in B-E; and the negative test set consists of those not selected in B-E.

Figure 6: Preference of the length and sequence context of spacers in CRISPR/dCas9 inhibition (CRISPRi) and activation (CRISPRa) screens. (A) Distribution of phenotype scores (Gilbert et al. 2014) for sgRNAs targeting the top 500 essential genes and the control sgRNAs in CRISPRi experiments. The dash line represents the threshold chosen to determine efficient and inefficient sgRNAs; **(B)** A bar-chart showing the effect of spacer length on sgRNA efficiency; **(C)** Logos showing the sequence preference of spacers. The height of the nucleotides represents the log odds-ratio of nucleotide frequency between efficient and inefficient sgRNAs. The nucleotide at the 5' end of the spacers are fixed to be guanines in the library design, and is excluded from the logos; **(D)** Bar-charts comparing the performance of CRISPRi model and CRISPR/Cas9 KO model in predicting sgRNA efficiency in CRISPRi negative selection, CRISPRi positive selection upon CTx-DTA treatment, and CRISPRa negative selections in Gilbert data and Konermann data. The length of spacers is 20nt. Cross-validation were used to assess the performance of CRISPRi model in CRISPRi negative selection experiment. The error bars represent the standard deviations in 10 iterations of 3-fold cross-validation. The p-value was computed using paired t-test.

Downloaded from genome.cshlp.org on June 14, 2015 - Published by Cold Spring Harbor Laboratory Press

TABLES

Table 1: A collection of CRISPR screening datasets used in the study.

Name used in this paper	Technique	Length of pacers	# of sgRNAs per gene	Type of selection	Reference
Wang data	CRISPR/Cas9 knockout	20	>20 for ribosomal genes, 10 for non-ribosomal genes	negative	Wang et al. 2014
Koike-Yusa data	CRISPR/Cas9 knockout	19	6	positive / negative	Koike-Yusa et al. 2014
Shalem data	CRISPR/Cas9 knockout	20	3-5	positive / negative	Shalem et al. 2014
Zhou data	CRISPR/Cas9 knockout	19-22	3-5	positive	Zhou et al. 2014
Gilbert data	CRISPR/dCas9 inhibition and activation	18-25	10	positive / negative	Gilbert et al. 2014
Konermann data	CRISPR/dCas9 activation	20	3	positive / negative	Konermann et al. 2015

Downloaded from genome.cshlp.org on June 14, 2015 - Published by Cold Spring Harbor Laboratory Press

REFERNCES

Andersson BS, Collins VP, Kurzrock R, Larkin DW, Childs C, Ost A, Cork A, Trujillo JM, Freireich EJ, Siciliano MJ. 1995. KBM-7, a human myeloid leukemia cell line with double Philadelphia chromosomes lacking normal c-ABL and BCR transcripts. *Leukemia* **9**: 2100–2108.

Berns K, Hijmans EM, Mullenders J, Brummelkamp TR, Velds A, Heimerikx M, Kerkhoven RM, Madiredjo M, Nijkamp W, Weigelt B, et al. 2004. A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature* **428**: 431–437.

Chen B, Gilbert LA, Cimini BA, Schnitzbauer J, Zhang W, Li GW, Park J, Blackburn EH, Weissman JS, Qi LS et al. 2013. Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas9 system. *Cell* **155**:1479-1491.

Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, et al. 2013. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**: 819–823.

Doench JG, Hartenian E, Graham DB, Tothova Z, Hegde M, Smith I, Sullender M, Ebert BL, Xavier RJ, Root DE. 2014. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nature Biotechnology* **32**: 1262–1267.

Fu Y, Sander JD, Reyon D, Cascio VM, Joung JK. 2014. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nature Biotechnology* **32**: 279–284.

Gilbert LA, Larson MH, Morsut L, Liu Z, Brar GA, Torres SE, Stern-Ginossar N, Brandman O, Whitehead EH, Doudna JA, et al. 2013. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**: 442–451.

Gilbert LA, Horlbeck MA, Adamson B, Villalta JE, Chen Y, Whitehead EH, Guimaraes C, Panning B, Ploegh HL, Bassik MC, et al. 2014. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* **159**:647-661.

Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, Li Y, Fine EJ, Wu X, Shalem O, et al. 2013. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature Biotechnology* **31**: 827–832.

Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. 2012. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**: 816–821.

Koike-Yusa H, Li Y, Tan E-P, Velasco-Herrera MDC, Yusa K. 2014. Genome-wide recessive genetic

Downloaded from genome.cshlp.org on June 14, 2015 - Published by Cold Spring Harbor Laboratory Press

screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nature Biotechnology* **32**: 267–273.

Konermann S, Brigham MD, Trevino AE, Joung J, Abudayyeh OO, Barcena C, Hsu PD, Habib N, Gootenberg JS, Nishimasu H, et al. 2015. Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature* **517**: 583–588.

Luo B, Cheung HW, Subramanian A, Sharifnia T, Okamoto M, Yang X, Hinkle G, Boehm JS, Beroukhim R, Weir BA, et al. 2008. Highly parallel identification of essential genes in cancer cells. *Proceedings of the National Academy of Sciences* **105**: 20380–20385.

Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM. 2013. RNA-guided human genome engineering via Cas9. *Science* **339**: 823–826.

Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP, Lim WA. 2013. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**: 1173–1183.

Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA, Zhang F. 2013. Genome engineering using the CRISPR-Cas9 system. *Nat Protoc* **8**: 2281–2308.

Schneider CA, Rasband WS, Eliceiri KW. 2012. NIH Image to ImageJ: 25 years of image analysis. *Nature Methods* **9**: 671–675.

Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelsen TS, Heckl D, Ebert BL, Root DE, Doench JG, et al. 2014. Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**: 84–87.

Silver DP, Livingston DM. 2001. Self-excising retroviral vectors encoding the Cre recombinase overcome Cre-mediated cellular toxicity. *Mol Cell* **8**: 233–243.

Sternberg SH, Redding S, Jinek M, Greene EC, Doudna JA. 2014. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507**: 62–67.

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B* **58**, 267–288.

Tikhonov AN. 2014. *Nonlinear Ill-Posed Problems*. Springer.

Tsai SQ, Zheng Z, Nguyen NT, Liebers M, Topkar VV, Thapar V, Wyvekens N, Khayter C, Iafrate AJ, Le LP, et al. 2015. GUIDE-seq enables genome-wide profiling of off-target cleavage by

Downloaded from genome.cshlp.org on June 14, 2015 - Published by Cold Spring Harbor Laboratory Press

CRISPR-Cas nucleases. *Nature Biotechnology* **33**: 187–197.

Wang T, Wei JJ, Sabatini DM, Lander ES. 2014. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**: 80–84.

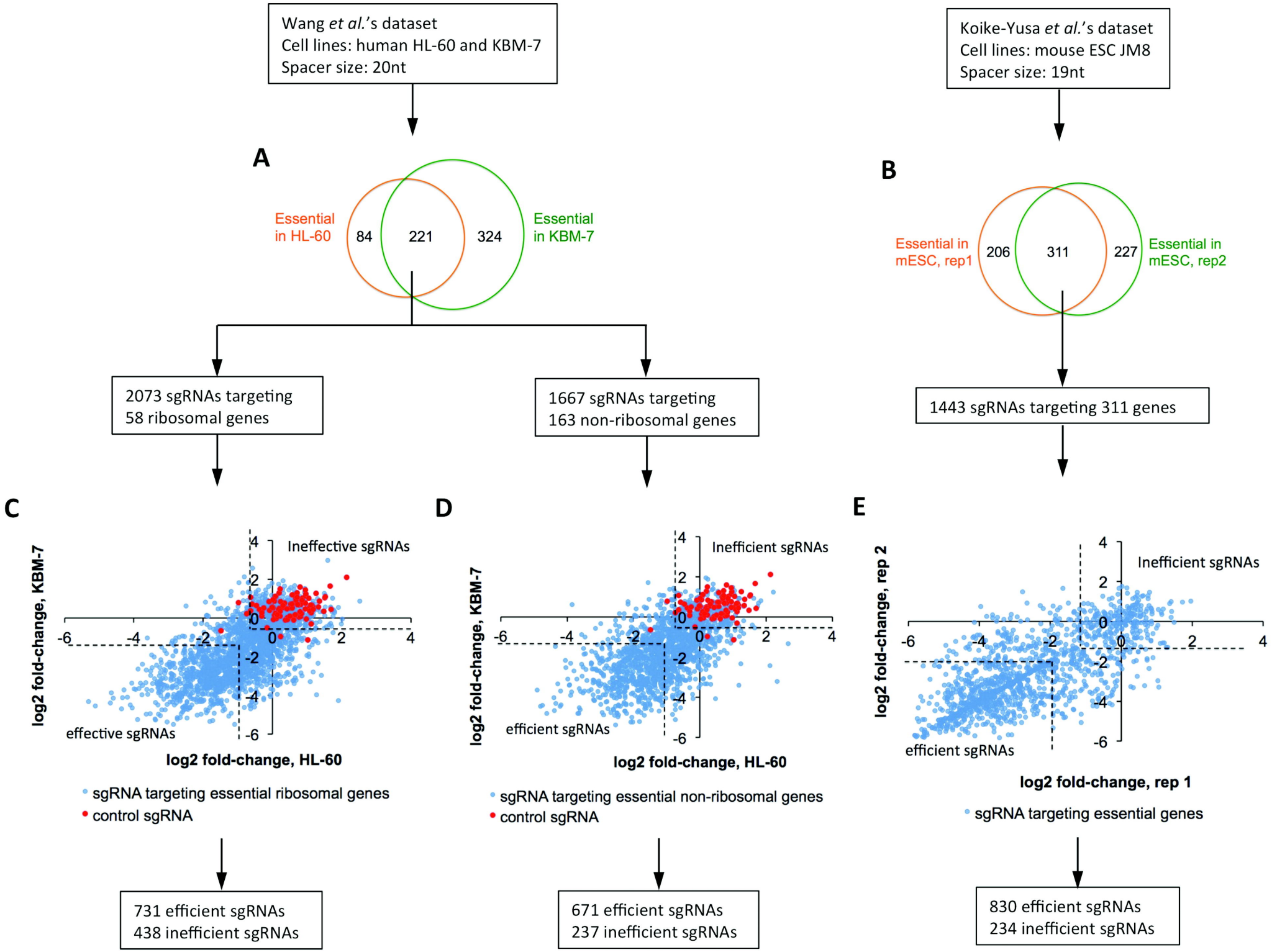
Wu X, Scott DA, Kriz AJ, Chiu AC, Hsu PD, Dadon DB, Cheng AW, Trevino AE, Konermann S, Chen S, et al. 2014. Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nature Biotechnology* **32**: 670–676.

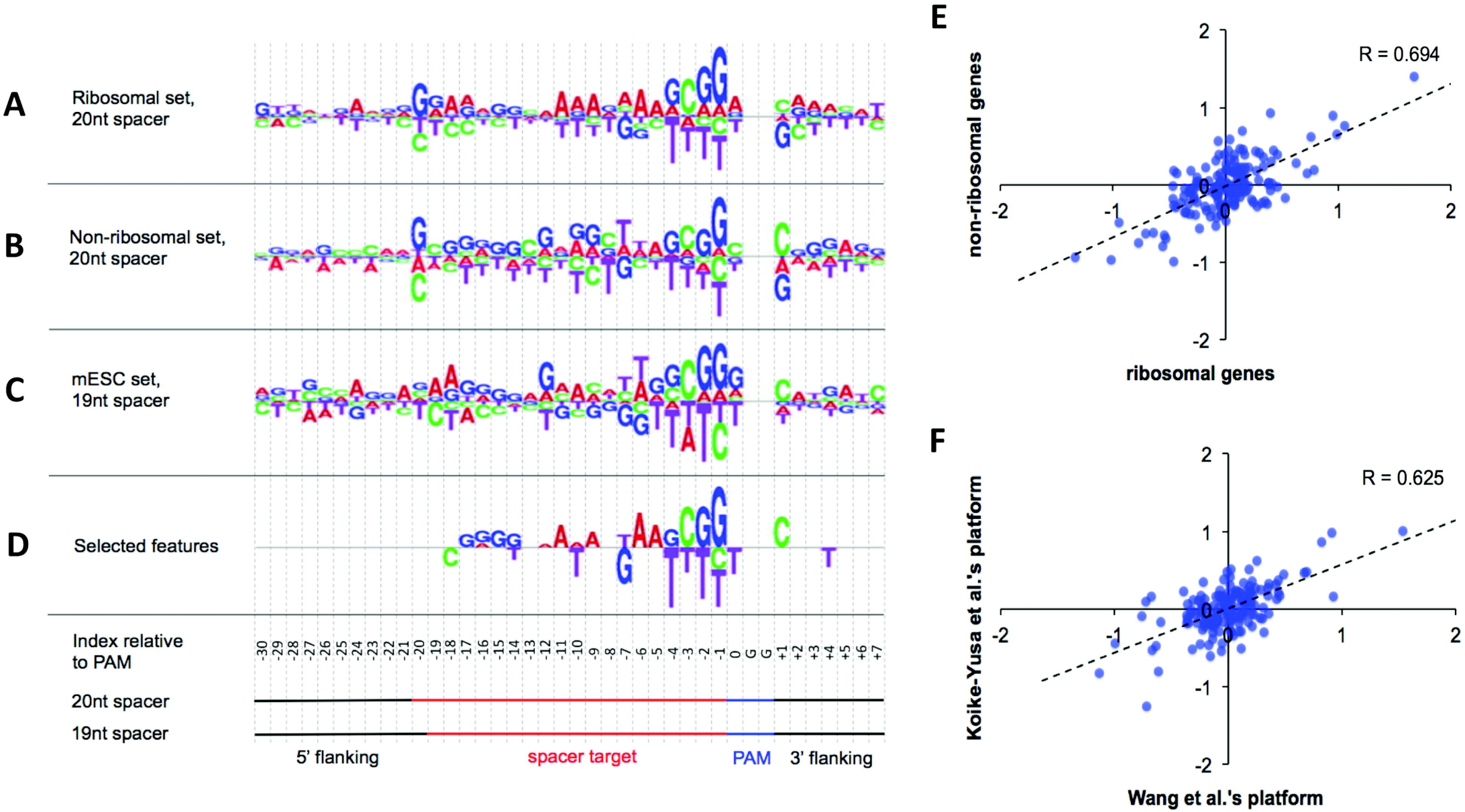
Xu K, Wu ZJ, Groner AC, He HH, Cai C, Lis RT, Wu X, Stack EC, Loda M, Liu T, et al. 2012. EZH2 oncogenic activity in castration-resistant prostate cancer cells is Polycomb-independent. *Science* **338**: 1465–1469.

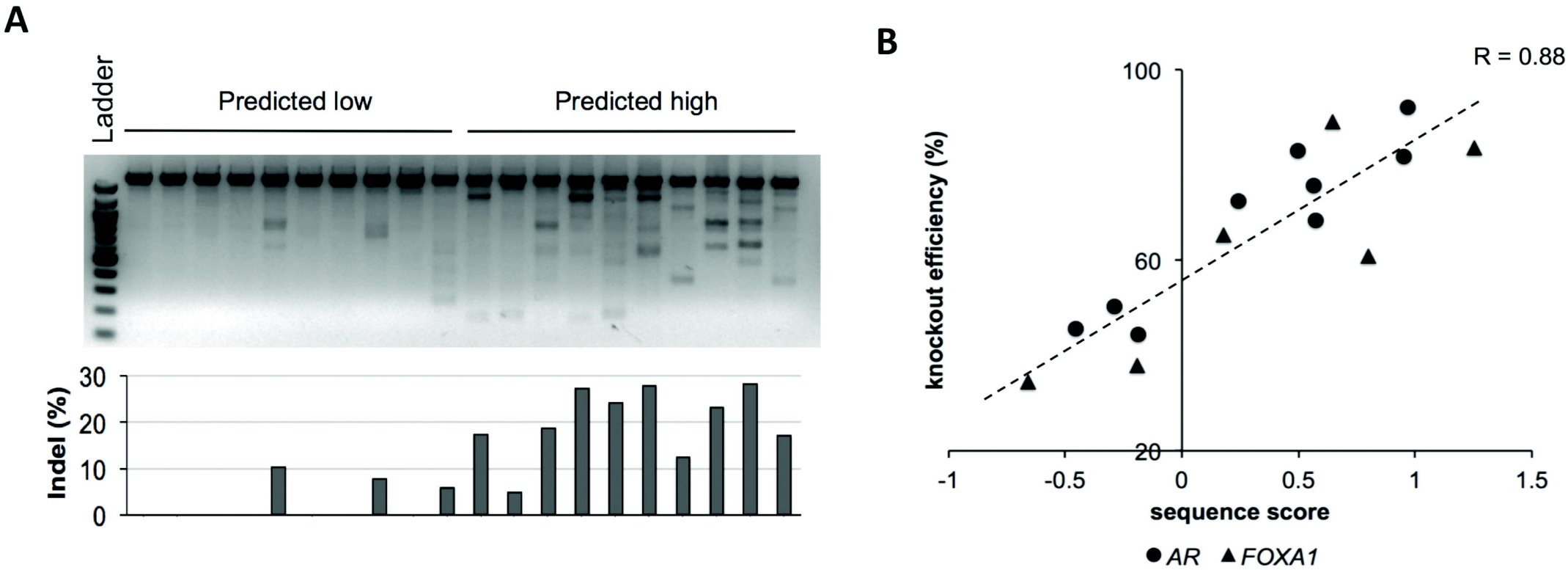
Yang H, Wang H, Shivalila CS, Cheng AW, Shi L, Jaenisch R. 2013. One-step generation of mice carrying reporter and conditional alleles by CRISPR/Cas-mediated genome engineering. *Cell* **154**: 1370–1379.

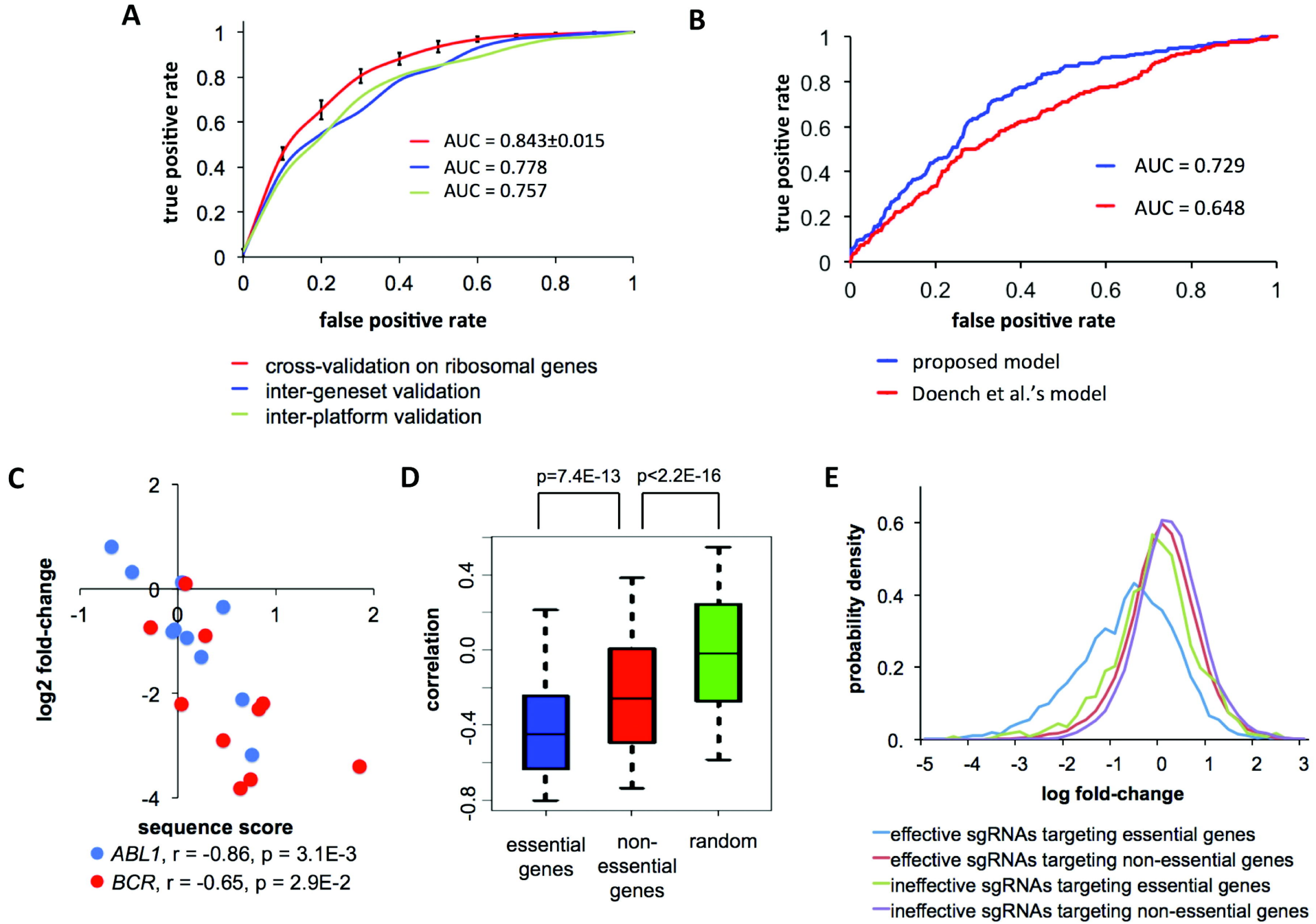
Zhou Y, Zhu S, Cai C, Yuan P, Li C, Huang Y, Wei W. 2014. High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. *Nature* **509**: 487–491.

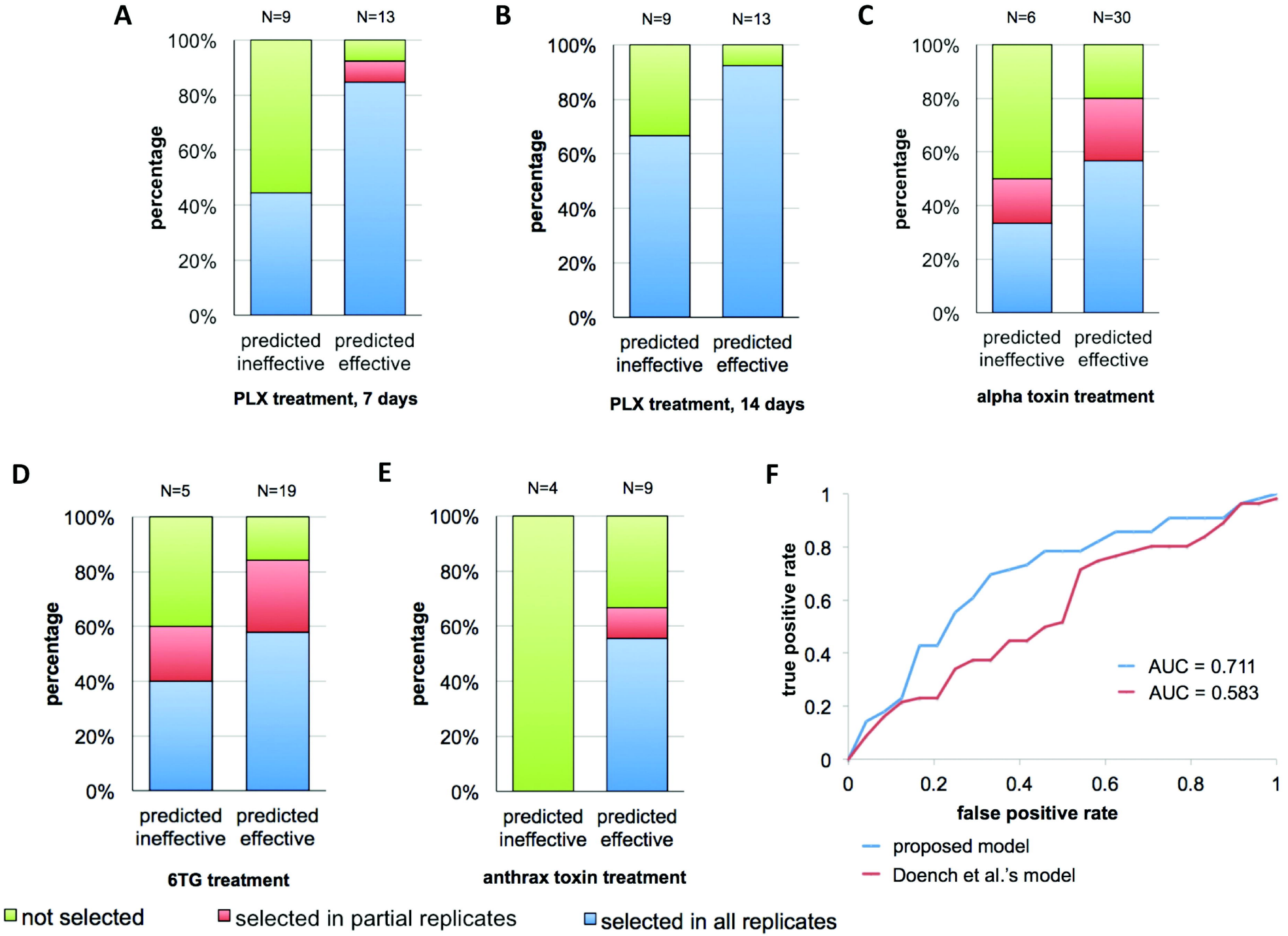
Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. *J Royal Statistical Soc B* **67**: 301–320.

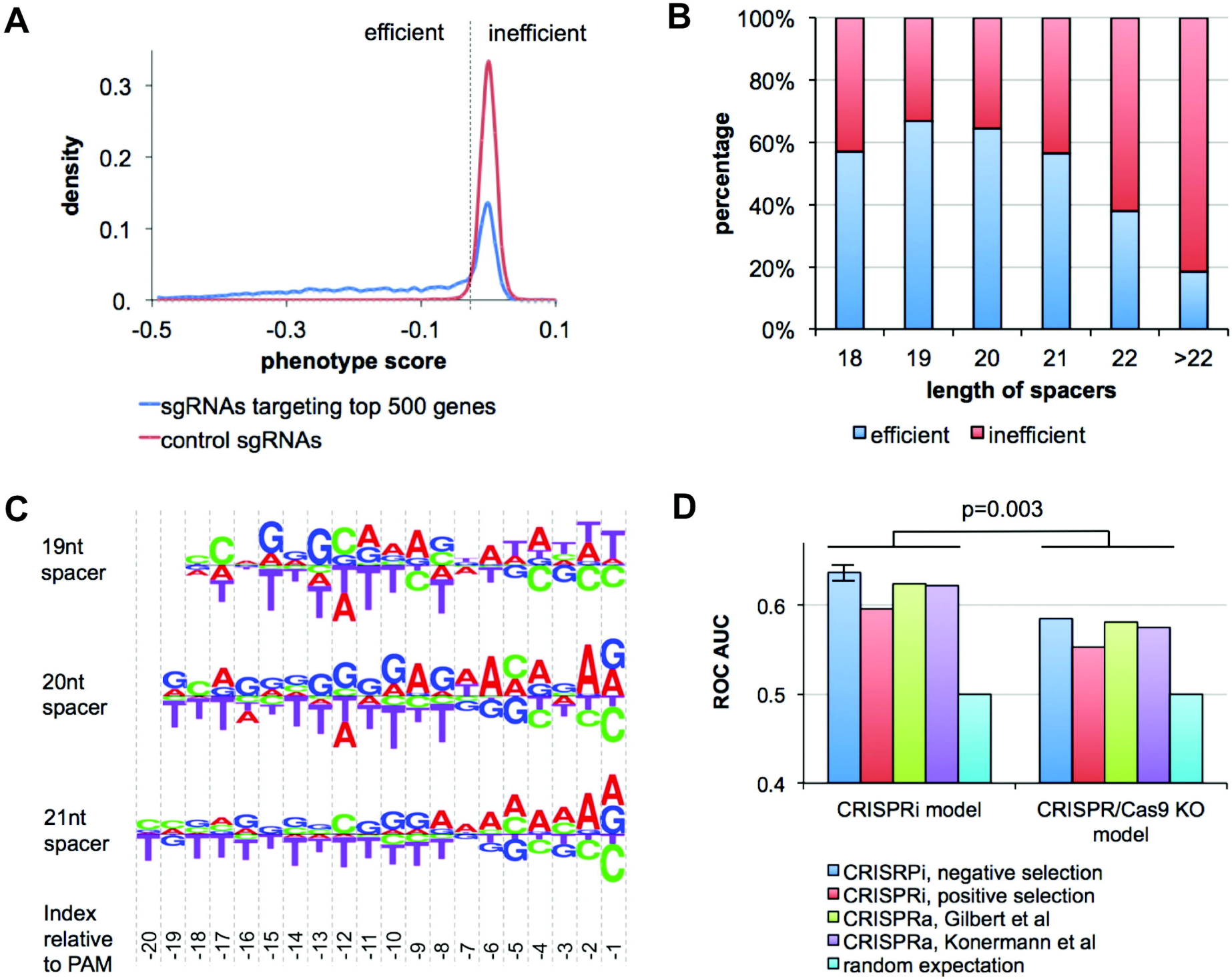

Wang et al.'s dataset
Cell lines: human HL-60 and KBM-7
Spacer size: 20nt
A
Essential in HL-60
84
221
324
Essential in KBM-7
2073 sgRNAs targeting 58 ribosomal genes
1667 sgRNAs targeting 163 non-ribosomal genes
Koike-Yusa et al.'s dataset
Cell lines: mouse ESC JM8
Spacer size: 19nt
B
Essential in mESC, rep1
206
311
227
Essential in mESC, rep2
1443 sgRNAs targeting 311 genes
C
Ineffective sgRNAs
effective sgRNAs
log2 fold-change, KBM-7
log2 fold-change, HL-60
sgRNA targeting essential ribosomal genes
control sgRNA
731 efficient sgRNAs
438 inefficient sgRNAs
D
Inefficient sgRNAs
efficient sgRNAs
log2 fold-change, KBM-7
log2 fold-change, HL-60
sgRNA targeting essential non-ribosomal genes
control sgRNA
671 efficient sgRNAs
237 inefficient sgRNAs
E
Inefficient sgRNAs
efficient sgRNAs
log2 fold-change, rep 2
log2 fold-change, rep 1
sgRNA targeting essential genes
830 efficient sgRNAs
234 inefficient sgRNAs


A
Ribosomal set,
20nt spacer
B
Non-ribosomal set,
20nt spacer
C
mESC set,
19nt spacer
D
Selected features
Index relative
to PAM
-30 -29 -28 -27 -26 -25 -24 -23 -22 -21 -20 -19 -18 -17 -16 -15 -14 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 G G +1 +2 +3 +4 +5 +6 +7
20nt spacer
19nt spacer
5' flanking
spacer target
PAM
3' flanking
E
R = 0.694
non-ribosomal genes
ribosomal genes
2
1
0
-1
-2
F
R = 0.625
Koike-Yusa et al.'s platform
Wang et al.'s platform


A
Ladder
Predicted low
Predicted high
Indel (%)
30
20
10
0
B
R = 0.88
knockout efficiency (%)
100
60
20
-1
-0.5
0
0.5
1
1.5
sequence score
AR
FOXA1


A
true positive rate
false positive rate
AUC = 0.843±0.015
AUC = 0.778
AUC = 0.757
cross-validation on ribosomal genes
inter-geneset validation
inter-platform validation
B
true positive rate
false positive rate
AUC = 0.729
AUC = 0.648
proposed model
Doench et al.'s model
C
log2 fold-change
sequence score
ABL1, r = -0.86, p = 3.1E-3
BCR, r = -0.65, p = 2.9E-2
D
p=7.4E-13
p<2.2E-16
correlation
essential genes
non-essential genes
random
E
probability density
log fold-change
effective sgRNAs targeting essential genes
effective sgRNAs targeting non-essential genes
ineffective sgRNAs targeting essential genes
ineffective sgRNAs targeting non-essential genes


A
N=9
N=13
100%
80%
60%
40%
20%
0%
percentage
predicted ineffective
predicted effective
PLX treatment, 7 days
B
N=9
N=13
PLX treatment, 14 days
C
N=6
N=30
alpha toxin treatment
D
N=5
N=19
6TG treatment
E
N=4
N=9
anthrax toxin treatment
F
true positive rate
false positive rate
AUC = 0.711
AUC = 0.583
proposed model
Doench et al.'s model
not selected
selected in partial replicates
selected in all replicates


A
efficient
inefficient
density
0.3
0.2
0.1
0.
-0.5
-0.3
-0.1
0.1
phenotype score
sgRNAs targeting top 500 genes
control sgRNAs
B
100%
80%
60%
40%
20%
0%
percentage
18
19
20
21
22
>22
length of spacers
efficient
inefficient
C
19nt spacer
20nt spacer
21nt spacer
Index relative to PAM
-20
-19
-18
-17
-16
-15
-14
-13
-12
-11
-10
-9
-8
-7
-6
-5
-4
-3
-2
-1
D
p=0.003
ROC AUC
0.6
0.5
0.4
CRISPRi model
CRISPR/Cas9 KO model
CRISRPi, negative selection
CRISPRi, positive selection
CRISPRa, Gilbert et al
CRISPRa, Konermann et al
random expectation

Downloaded from genome.cshlp.org on June 14, 2015 - Published by Cold Spring Harbor Laboratory Press



Sequence determinants of improved CRISPR sgRNA design

Han Xu, Tengfei Xiao, Chen-Hao Chen, et al.

Genome Res. published online June 10, 2015
Access the most recent version at doi:10.1101/gr.191452.115

Supplemental Material http://genome.cshlp.org/content/suppl/2015/06/12/gr.191452.115.DC1.html

P<P Published online June 10, 2015 in advance of the print journal.

Accepted Manuscript Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.

Creative Commons License This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**



To subscribe to *Genome Research* go to:
http://genome.cshlp.org/subscriptions

Published by Cold Spring Harbor Laboratory Press