Supplementary Materials for

Sequence determinants of improved CRISPR sgRNA design

Xu et al.

Table of Contents

The sample size in validations of sequence features

The sample size is important to achieve sufficient statistical power in testing a hypothesis. The power of a statistical test refers to the probability of rejecting a null hypothesis H_0 when it is not correct, i.e.:

$$\text{Power} = \text{Prob}(\text{reject } \text{H}_0 | \text{H}_0 \text{ is false})$$

The rejection of H_0 is subject to the Type I error, which is usually set to be 0.05, 0.01 or below, corresponding to a confidence interval of 95%, 99% or above.

To explore the relationship between sample size and statistical power in our study, we performed computational analysis in two scenarios:

i) How many samples are required to validate the model for predicting sgRNA efficiency in CRISPR/Cas9 knockout experiment?
ii) How many samples are required to validate individual sequence features discovered by the computational model?

Sample size for the validation of predictive model

In Figure 3B, we showed that the predictive sequence score is correlated with protein knockout efficiency (r=0.88). To check if we have sufficient samples for the validation, we plotted the statistical power as a function of sample size in a Pearson Correlation Test, where the expected correlation of two variables is 0.88 (Supplementary Fig. 7). Our result showed that a statistical power of 0.9 requires 7 samples in the validation when the Type I error is set to be 0.05. Even with a more stringent Type I error of 0.001, 13 samples are sufficient for the validation. As we included 15 sgRNAs in the experiment in Fig. 3B, we have achieved enough statistical power to validate our predictive model.

Sample size for the validation of individual sequence features

The sample size for validating a feature depends on the significance of the feature in the comparison between efficient and inefficient sgRNAs. More samples are needed if the feature is relatively weak. We first choose the feature at the -3 position (see Fig. 2D for reference of index), where the cytosine is preferred. In Wang data, “C” is two-fold more enriched at the -3 position in efficient sgRNAs compared to that in inefficient sgRNAs (36% vs. 18%, Supplementary Fig. 8A).

Given n samples that include $n/2$ efficient sgRNAs and $n/2$ inefficient ones, the occurrence of “C” at the -3 position among efficient sgRNAs, denoted x_1, follows a binomial distribution $binom(\frac{n}{2}, 0.36)$. Similarly, the number of inefficient sgRNAs containing “C” at the -3 position, denoted x_2, follows a distribution of $binom(\frac{n}{2}, 0.18)$. To determine the statistical power when sample size is n, we randomly generated 100,000 pairs of (x_1, x_2), and calculated a p-value of Fisher Exact Test based on each pair of (x_1, x_2) and n. The statistical power with a Type I error α and a sample size n was computed to be the fraction of p-values smaller than α. As shown in Supplementary Figure 8B, approximately 220 samples are needed to achieve a statistical power >0.9 with $\alpha = 0.05$, and more than 300 samples are needed when $\alpha = 0.01$.

Next we took an example of relatively weaker features. The feature is at the -1 position, where “C” is approximately 1.5-fold depleted in efficient sgRNAs compared to inefficient ones (25% vs. 37%, Supplementary Fig. 9A). We repeated the above simulation based on binomial distributions $x_1 \sim binom(\frac{n}{2}, 0.25)$ and $x_1 \sim binom(\frac{n}{2}, 0.37)$, and estimated statistical power as a function of sample size given certain Type I error. As the result, 540 and 800 samples are need to achieve a statistical power > 0.9, corresponding to Type I errors of 0.05 and 0.01, respectively (Supplementary Fig. 9B).

List of Supplementary Tables

Supplementary Table 1 (Supplementary_table_1.xlsx)

Lists of efficient and inefficient sgRNAs in the “ribosomal”, “non-ribosomal” and “mESC” training sets.

Supplementary Table 2 (Supplementary_table_2.xlsx)

The matrix of coefficients learnt by Elastic-Net to represent the nucleotide preference in CRISPR/Cas9 knockouts (Fig. 2D).

Supplementary Table 3 (Supplementary_table_3.xlsx)

A list of sgRNAs that target *AAVS1* locus for the validation of mutation rates (Fig. 3A).

Supplementary Table 4 (Supplementary_table_4.xlsx)

A list of sgRNAs that target *AR* and *FOXA1* for the validation of protein knockout efficiency (Fig. 3B).

Supplementary Table 5 (Supplementary_table_5.xlsx)

Lists of sgRNAs that target known genes involved in drug or toxin resistance, for the analysis of sgRNA efficiency in positive selection (Fig. 5).

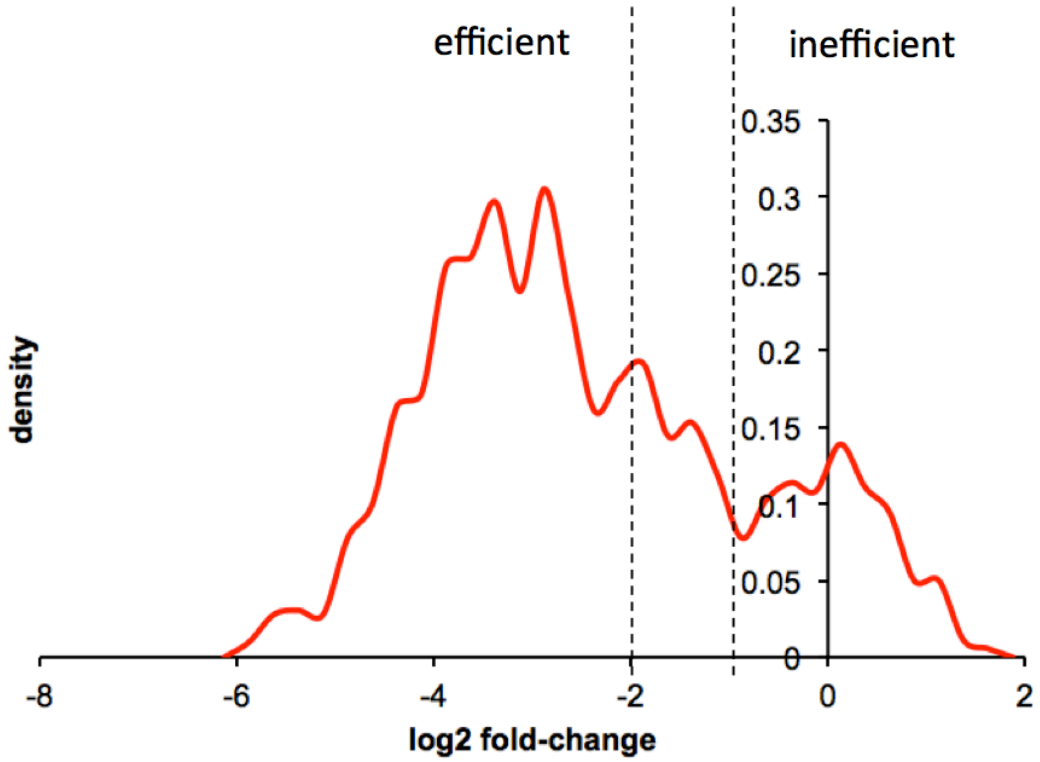
Supplementary Table 6 (Supplementary_table_6.xlsx)

Lists of efficient and inefficient sgRNAs used in the analysis on Gilbert et al.’s CRISPRi/a data.

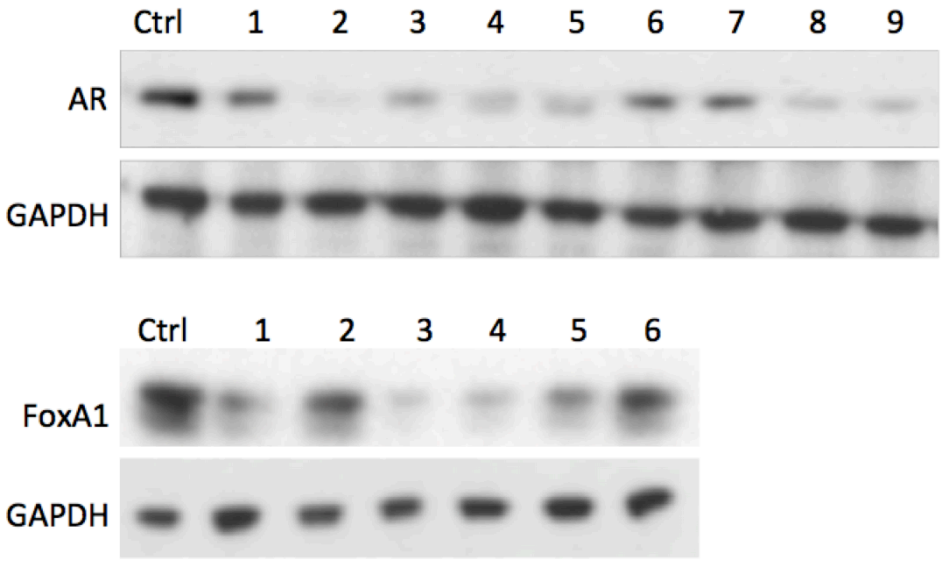
Supplementary Table 7 (Supplementary_table_7.xlsx)

The matrices of coefficients learnt by Elastic-Net to represent the nucleotide preference in CRISPRi experiments.

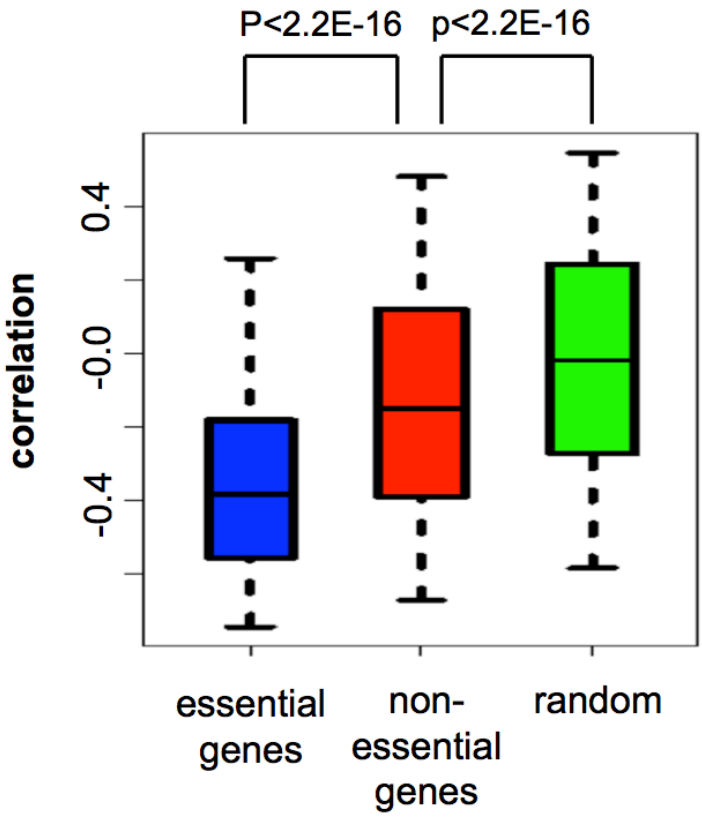
Supplementary Figures



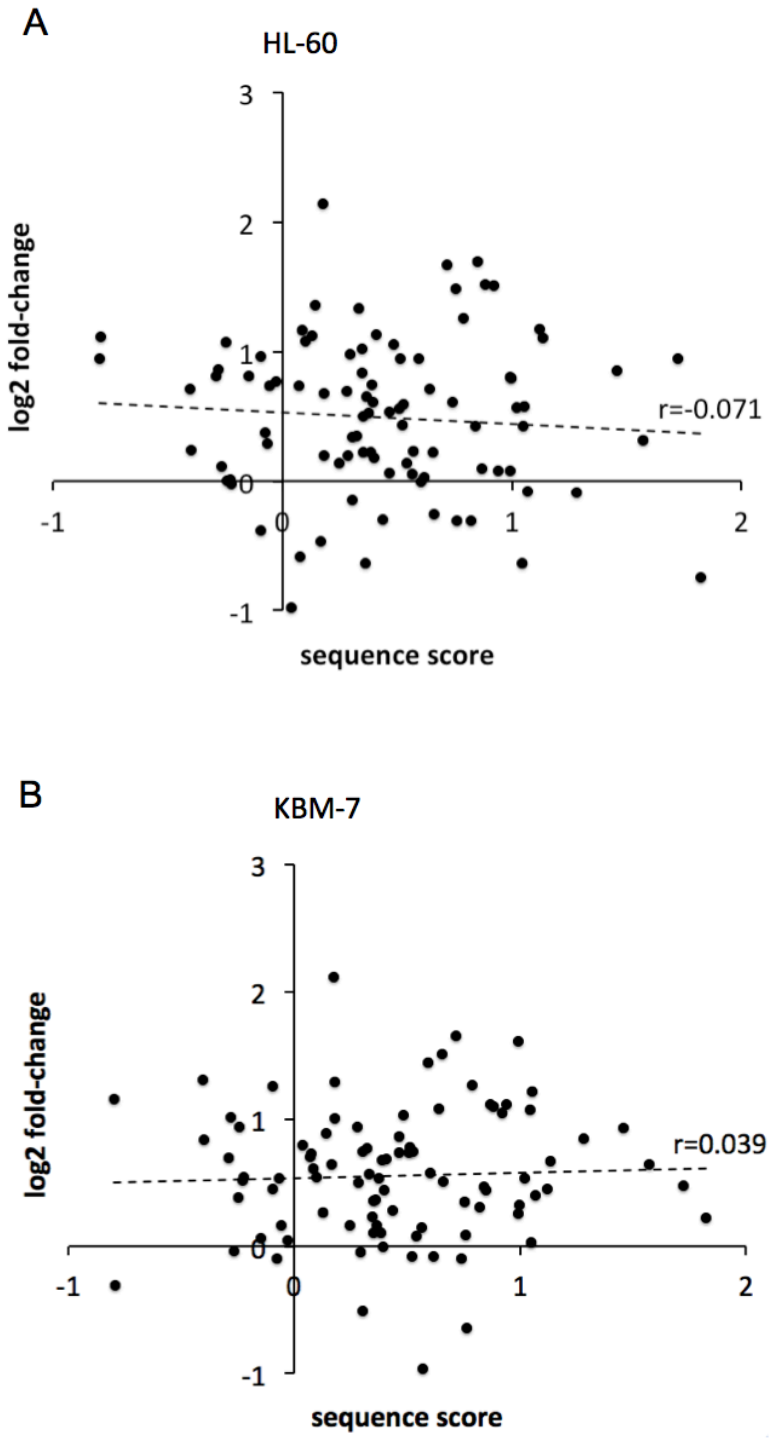
Supplementary Figure 1: Distribution of log2 fold-change showing the bimodality of relative sgRNA abundance for sgRNAs targeting essential genes in mESC. The log2 fold-changes were averaged on two biological replicates. The dashed lines represent the threshold chosen for determining efficient and inefficient sgRNAs.



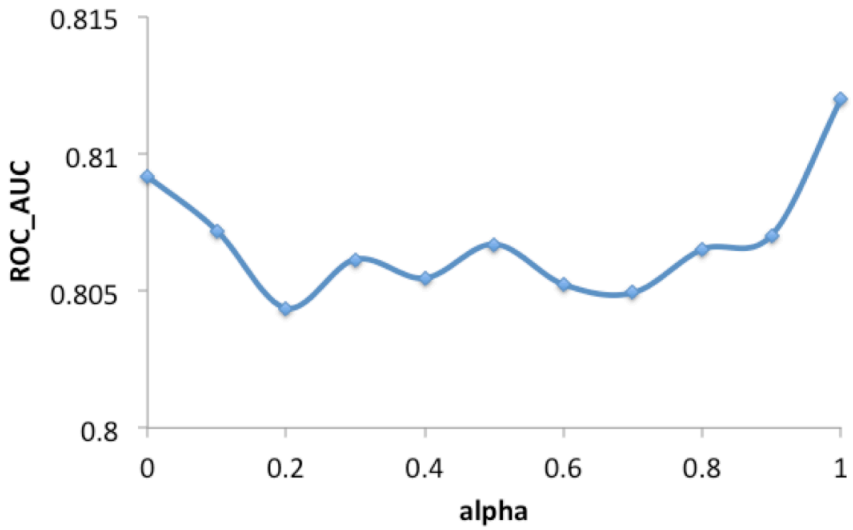
Supplementary Figure 2: Western blot showing the protein knockout efficiency mediated by sgRNAs in LNCaP-abl cells. See Supplementary Table 4 for the information of sgRNAs in the experiment.



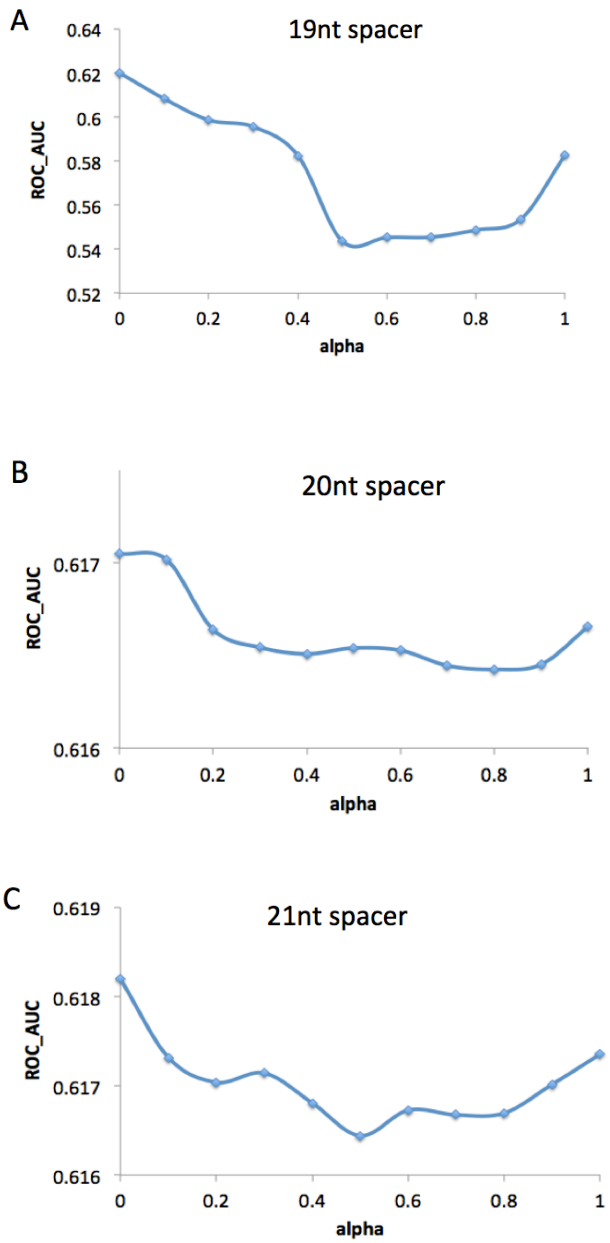
Supplementary Figure 3: Box-plot showing the distributions of correlations between sequence scores and relative sgRNA abundances, for essential and non-essential genes in HL-60. The distribution of random background was computed by permuting the sequence scores within each gene in the dataset.



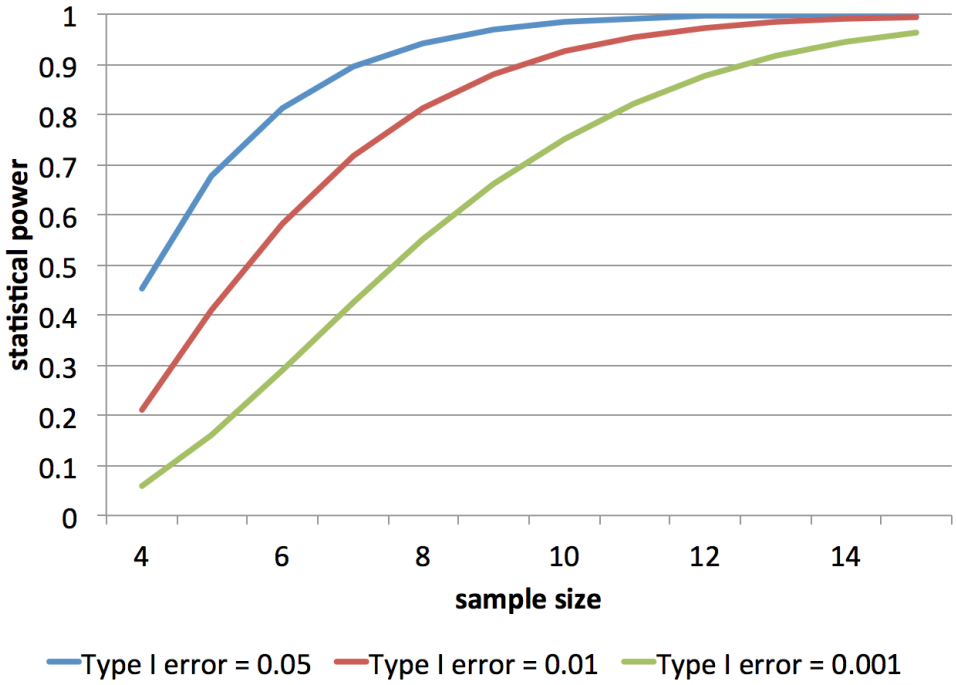
Supplementary Figure 4: Scatter plot showing the correlation between the sequence score and relative sgRNA abundance for control sgRNAs in (A) HL-60 and (B) KBM-7.



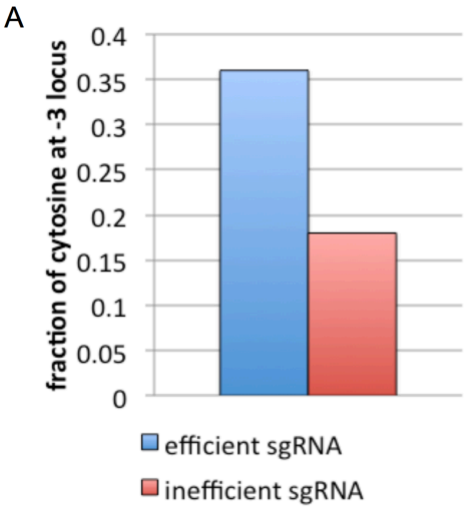
Supplementary Figure 5: The predictive power of the Elastic-Net as a function of the parameter alpha when the model is applied to CRISPR/Cas9 knockout data. The predictive power was measured to be the average ROC_AUC score upon 10 cross-validations.

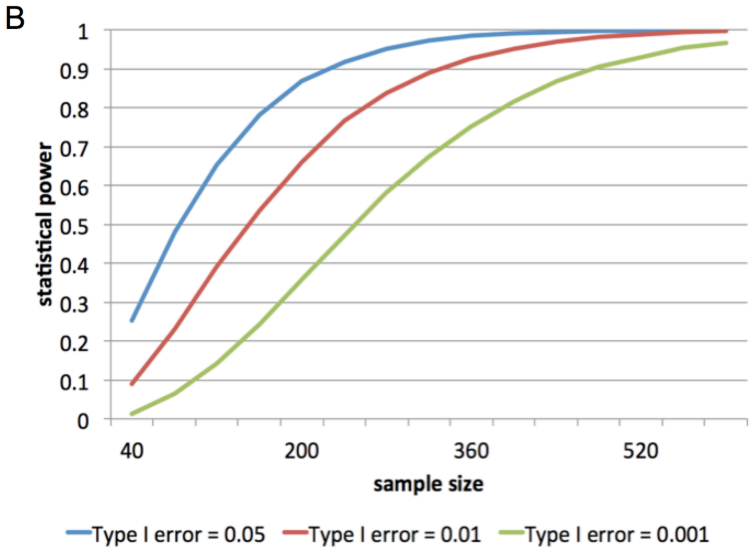


Supplementary Figure 6: The predictive power (AUC-ROC score) of the Elastic-Net as a function of the parameter alpha when the model is applied to CRISPRi data, for sgRNAs with different lengths of spacers. The predictive power was measured to be the average ROC-AUC score upon 10 cross-validations.

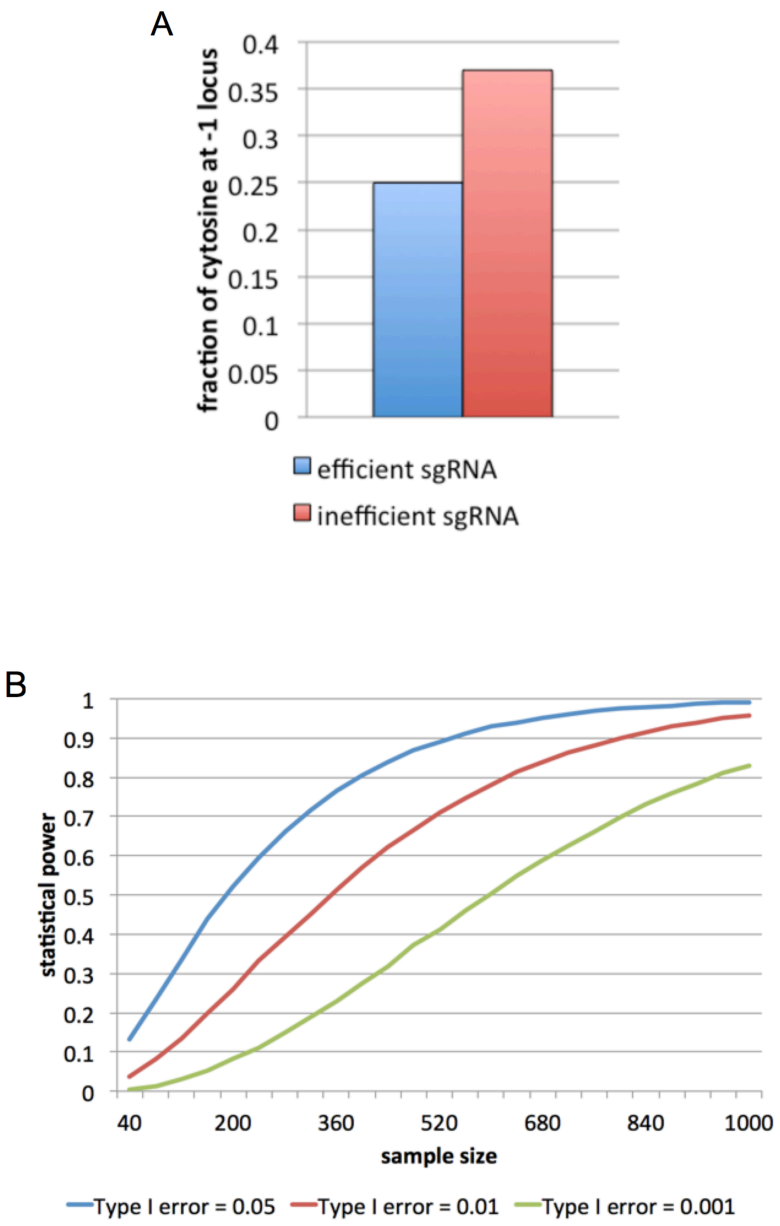


Supplementary Figure 7: Relationship between sample size and statistical power in the validation experiment shown in Figure. 3B.





Supplementary Figure 8: Relationship between sample size and statistical power for validating the preference of cytosine at the -3 position. (A) The fraction of cytosine at the -3 position in efficient and inefficient sgRNAs, based on 2,077 samples in Wang data; (B) The statistical power as a function of sample size in a Fisher Exact test to validate the sequence feature of cytosine at -3 position.



Supplementary Figure 9: Relationship between sample size and statistical power for validating the depletion of cytosine at the -1 position. (a) The fraction of cytosine at the -1 position in efficient and inefficient sgRNAs, based on 2,077 samples in Wang data; (b) The statistical power as a function of sample size in a Fisher Exact test to validate the sequence feature of cytosine at -1 position.