Supplementary Information

CRISPR-DO for genome-wide CRISPR design and optimization

Online application availability

The CRISPR sgRNA Design and Optimization (CRISPR-DO) web server is publicly available at http://cistrome.org/crispr/ . The input of CRISPR-DO includes a genome build and the guide length (19 or 20nt). Users can design sgRNA based on a gene symbol, a refseq ID or a genomic coordinate region not limited to annotated genes. Users have the options to configure sgRNA specificity and efficiency scores, and to choose whether or not design sgRNA on exons or the whole region of a gene. The execution time of CRISPR-DO is estimated in Figure S5 and Table S5, depending on genome length and the number of sgRNAs available in the input region. A link will be sent to user by email when the job is finished. The result page of the application displays the designed targets in two views. The table view shows one target per row with its location, efficiency, specificity, average conservation score, as well as its overlap with genomic features like exons, DHSs and SNPs (Fig. S6). Users can sort or filter the table, and download the results into a local file for further refinement. The WashU EpiGenome browser (Zhou, et al., 2011) view allows a more comprehensive investigation of sgRNA target locations on the genome (Fig. S7) and also allows other data tracks such as ChIP-seq. These can be incorporated for further characterization of the targets.

Specificity score estimation

Here we use the method from Zhang Lab (Hsu, et al., 2013) for off-target effect evaluation. First, we collect all the potential off-target sites with a PAM of NGG or NAG and allow maximum 3 mismatches. We then calculate a score for each potential off-target site (S_{hit}) based on

$$S_{hit} = \Pi_{e \in M}\left(1 - W[e]\right) \times \frac{1}{(1-d/l)\times 4 + 1} \times \frac{1}{n_{mm}^2}$$

This formula is the multiplication of three factors. 1) The first factor quantifies the contribution of mismatch in different position. For each e (index of mismatched nucleotide position) in M (a set of mismatched nucleotide positions for the off-target site), we calculate 1-W[e], where W represents the experimentally determined effect of mismatch position on targeting, and W = [0, 0, 0.014, 0, 0, 0.395, 0.317, 0, 0.389, 0.079, 0.445, 0.508, 0.613, 0.851, 0.731, 0.828, 0.615, 0.804, 0.685, 0.583] (Hsu, et al., 2013). 2) The second term factorizes in the effect of mean pairwise distance between mismatches. d is the average distance for each two adjacent mismatched positions. l is the length of spacer minus 1. 3) The third factor quantifies the impact of the number of mismatches, denoted n_{mm}. is the number of mismatched nucleotides.

With more mismatches, the site is less likely to be an off-target.

Here is an example to further clarify the formula. If the target sequence (20nt+PAM) is CCGTGGCCATTCAGGCGCCT**GGG**, and one of the off-target sequence is CCGTGGCGGTTCAGGCGCCA**GAG**. The mismatched nucleotide position indexes are 7, 8, 19 (index starts from 0). Thus

$$d = \frac{((8-7)+(19-8))}{n_{mm}-1} = 6$$

$$S_{hit} = (1-W[7])\times(1-W[8])\times(1-W[19])\times\frac{1}{(1-6/19)\times 4+1}\times\frac{1}{3^2} = 0.0076$$

To summarize the scores of all potential off-targets, we calculated the specificity score for an sgRNA by the formula

$$S_{guide} = \frac{100}{1+\sum_{i=1}^{n_{off-target}} S_{hit}}$$

Where ΣS_{hit} is the sum of the scores of all potential off-targets. This specificity score ranges from 0 to 100, and the sgRNAs with lower off-target effect will have a higher specificity score. The distribution of specificity score for each genome is showed in Fig. S4. The average score for fly and worm are much higher, with most sgRNAs having specificity score larger than 60 (78.2% of total sgRNAs in worm and 82.9% of total sgRNAs in fly). This is likely ascribed to the short genome sequence of these two species.

Efficiency score estimation

The efficiency score calculation is based on our previous model (Xu, et al., 2015). For model construction, we selected a list of "efficient" and "inefficient" sgRNAs targeting a list of essential genes by comparing their relative abundance in public datasets (Koike-Yusa, et al., 2014; Wang, et al., 2014). We next extended the target sequence to 40nt (10nt extension for both side of the spacer) for each selected sgRNA, and encoded target sequence into a binary vector. Since each position takes one of the nucleotides {A, C, G, T}, the length of the vector is 4 times the sequence length. For example, a sequence "AGTC" is encoded into [1,0,0,0,0,0,1,0,0,0,0,1,0,1,0,0].

We then applied the Elastic-Net (Zou and Hastie, 2005) model to train a linear model for prediction. Suppose $X = [X_1, X_2, ..., X_N]^T$ is the set of encoded sequence vectors and $Y = [y_1, y_2, ..., y_N]^T$ 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 $\beta = [\beta_1, \beta_2, ..., \beta_M]^T$ that minimize an objective function *E*:

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

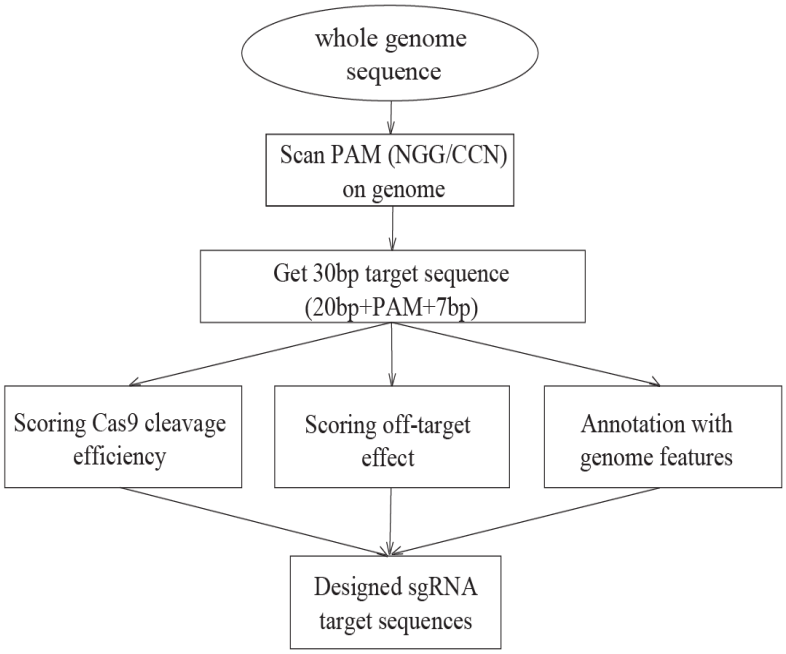
Where α and λ are parameters estimated using cross-validation, $\|\beta\|^1 = \sum_i |\beta_i|$, and

$\|\beta\|^2 = \sum_i \beta_i^2$.

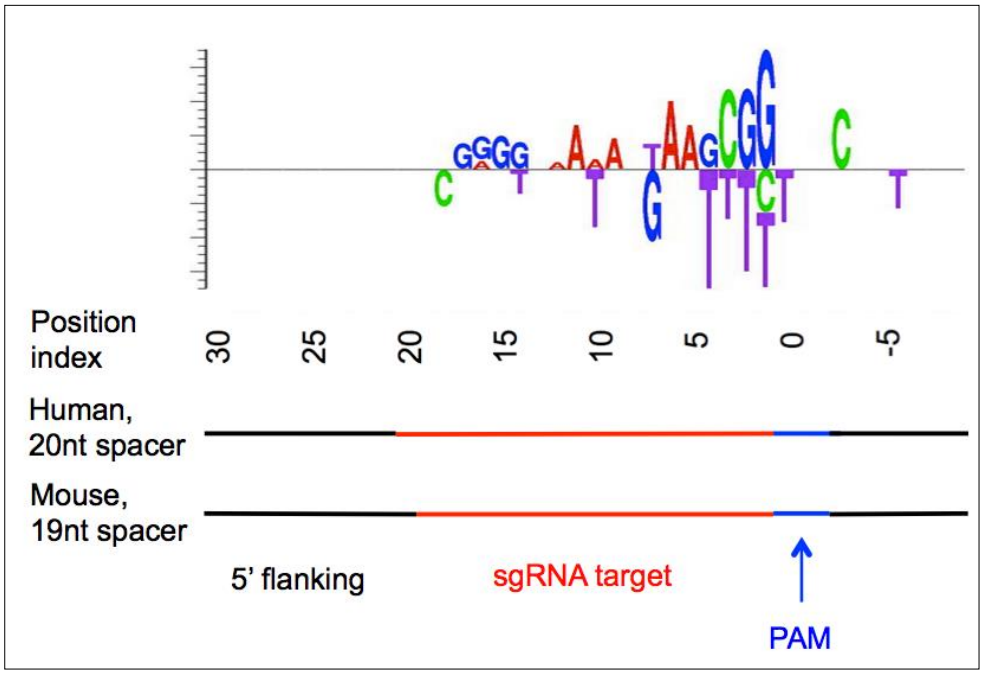
Given an encoded sequence vector *X*, the efficiency score *S* is computed to be $S = \beta^T X$.

Xu et al. have shown that an ROC-AUC score of 0.7-0.8 can be achieved when using this model for sgRNA efficiency prediction.We found the sgRNAs with scores > 1.4 (0.048% of total sgRNAs in human, and 0.070% of total sgRNAs in mouse) are highly GC-enriched. Xu et al.'s study was focused on coding exons, where such high GC-content is rare. In contrast, CRISPR-DO is designed for genome-wide annotation, where high GC-content can be more frequent near regulatory elements such as CpG islands. We reason that high GC-content might result in PCR artifacts and mild yet promiscuous off-target cutting in regulatory regions, therefore suggest a safe range of 0.2-1.4 in practice.The distribution of efficiency score for each genome is showed in Fig. S3.

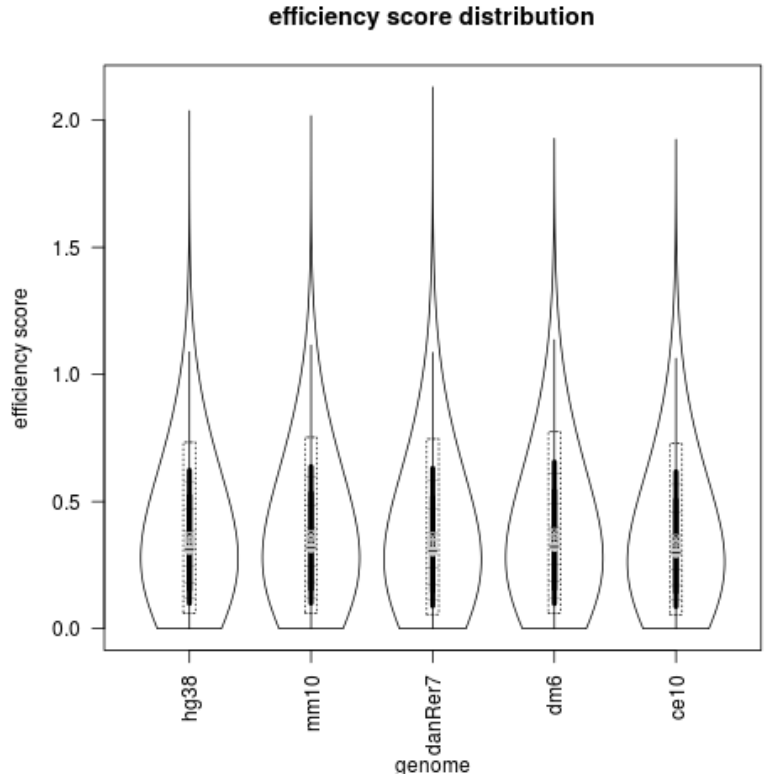
Supplementary Figures



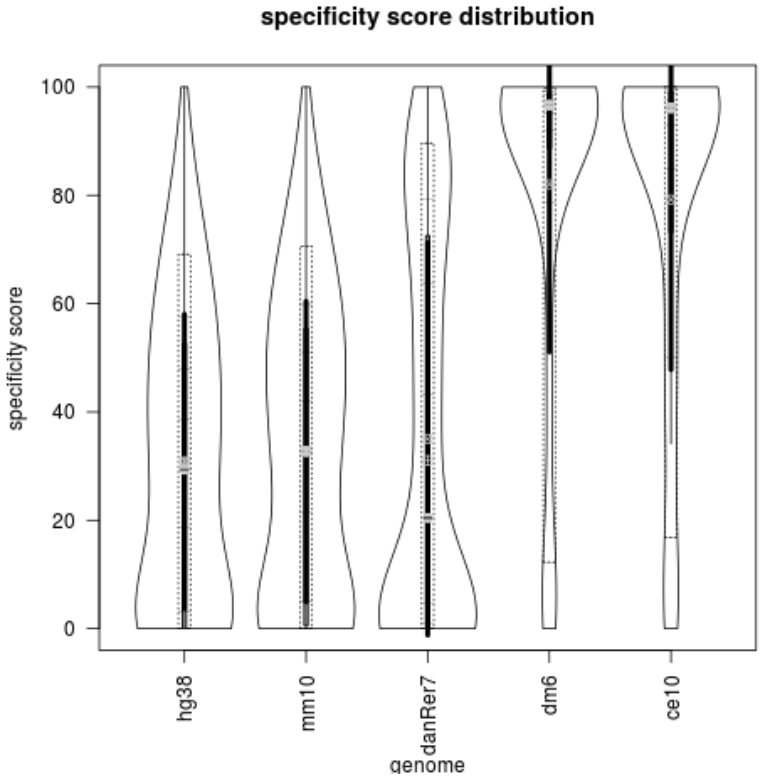
Supplementary Figure S1. Workflow to generate the core database of genome-wide guide sequences.



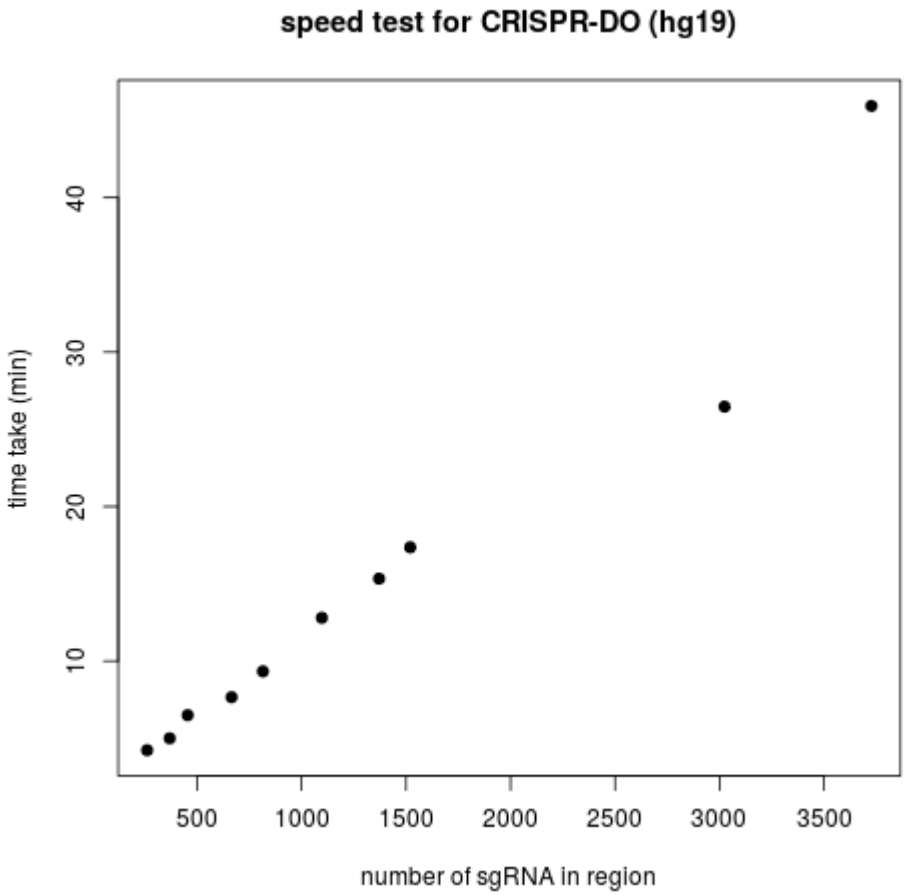
Supplementary Figure S2. Sequence logo for the efficiency model prediction. The height of the nucleotides represents the coefficients calculated from the Elastic-Net. Alphabets under zero-line indicate that the corresponding nucleotide at that position has negative contribution to the sgRNA efficiency.



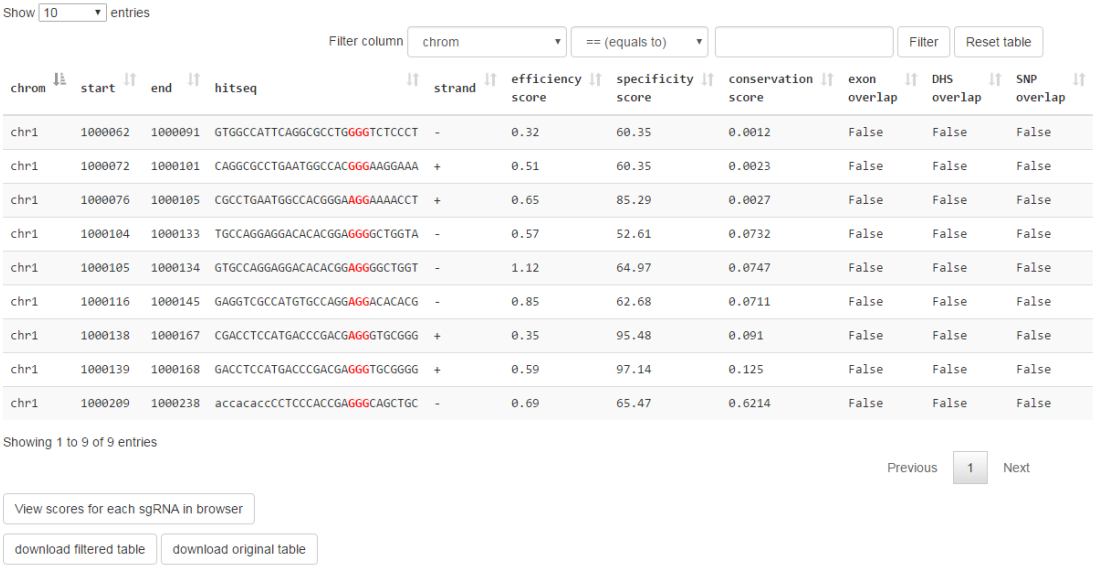
Supplementary Figure S3. The distribution of efficiency score for the total sgRNA with efficiency larger than zero in hg38, mm10, danRer7, dm6 and ce10.



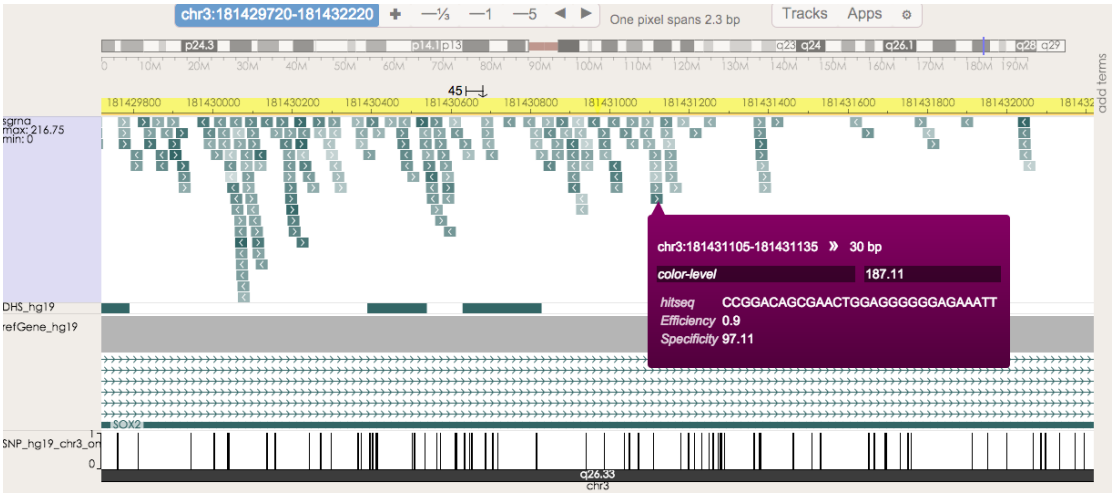
Supplementary Figure S4. The distribution of specificity score for the total sgRNA with efficiency larger than zero in hg38, mm10, danRer7, dm6 and ce10.



Supplementary Figure S5. Correlation between time cost for running CRISPR-DO and the number of sgRNAs in the region.



Supplementary Figure S6. Example table output for the target sequence scan result. The first 5 column is the basic information of the target sequence located in genome. The 6th, 7th and 8th column is the score for each target sequence calculated by our tool respectively. And the last 3 column is the overlap of the guide sequence with these genomic features.



Supplementary Figure S7. Example epigenome browser view for the output of the sgRNA target sequence scan result. The color level of each target sequence is calculated by combining its efficiency and specificity score.

Supplementary Tables

Supplementary Table S1. The total number in millions of target sequence identified by NGG in hg38, mm10 and danRer7 whole genome.

Chromosome	hg38 (m)	mm10 (m)	danRer7 (m)
chr1	24.96	19.48	4.14
chr2	24.13	18.91	4.19
chr3	19.28	15.34	4.43
chr4	17.09	16.38	4.54
chr5	17.41	16.04	5.19
chr6	16.50	15.04	4.13
chr7	16.25	15.88	5.39
chr8	14.39	13.48	3.91
chr9	12.88	13.16	4.01
chr10	14.20	13.06	3.22
chr11	14.41	13.64	3.19
chr12	13.70	12.21	3.46
chr13	8.92	12.09	3.72
chr14	9.36	12.35	3.75
chr15	9.23	10.66	3.31
chr16	10.00	9.53	4.08
chr17	10.50	10.03	3.75
chr18	7.68	8.95	3.46
chr19	8.25	6.35	3.45
chr20	7.56	-	3.87
chr21	4.11	-	3.08
chr22	5.30	-	3.00
chr23	-	-	3.24
chr24	-	-	3.00
chr25	-	-	2.67
chrM	0.0022	0.0015	0.0017
chrX	14.97	15.71	-
chrY	2.60	8.29	-
total	303.67	276.57	94.18

Supplementary Table S2. The total number in millions of target sequence identified by NGG in ce10 whole genome.

Chromosome	ce10 (m)
chrI	1.04
chrII	1.07
chrIII	0.961
chrIV	1.12
chrV	1.39
chrX	1.13
chrM	0.00059
total	6.71

Supplementary Table S3. The total number in millions of target sequence identified by NGG in dm6 whole genome.

Chromosome	dm6 (m)
chr2L	2.16
chr2R	2.38
chr3L	2.56
chr3R	3.03
chr4	0.0894
chrX	2.22
chrY	0.275
chrM	0.000587
total	12.72

Supplementary Table S4. The phastCons score used to calculate average conservation of each target sequence.

Genome version	Links
hg19	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/phastCons46way/vertebrate/
hg38	http://hgdownload.cse.ucsc.edu/goldenPath/hg38/phastCons7way/
mm9	http://hgdownload.cse.ucsc.edu/goldenPath/mm9/phastCons30way/vertebrate/
mm10	http://hgdownload.cse.ucsc.edu/goldenPath/mm10/phastCons60way/placental/
ce10	http://hgdownload.cse.ucsc.edu/goldenPath/ce10/phastCons7way/
dm6	http://hgdownload.cse.ucsc.edu/goldenPath/dm6/phastCons27way/
danRer7	http://hgdownload.cse.ucsc.edu/goldenPath/danRer7/phastCons8way/

Supplementary Table S5. the time cost matrix for running jobs.

Genome	Start position	Length	sgRNA number	Time (s)
hg19	chr1.1000000	3k	261	254
hg19	chr1.1000000	4k	369	300
hg19	chr1.5000000	15k	455	390
hg19	chr1.3000000	10k	665	460
hg19	chr1.3000000	12k	815	560
hg19	chr1.3000000	15k	1097	768
hg19	chr1.3000000	18k	1371	920
hg19	chr1.3000000	20k	1520	1042
hg19	chr1.1000000	30k	3024	1588
hg19	chr1.3000000	50k	3727	2756
mm9	chr1.3000000	6k	127	131
mm10	chr3.34650000	6k	288	204
mm10	chr3.34650000	30k	1229	720
danRer7	chr11.36881200	3k	42	45
danRer7	chr11.36881200	30k	400	240
dm6	chrX.22495600	3k	76	11
dm6	chrX.22495600	30k	663	70
ce10	chrIII.1258100	4k	56	8
ce10	chrIII.1258100	30k	330	38

Supplementary Table S6. β matrix from Elastic-Net model for sgRNA efficiency prediction.

Index	A	C	G	T
0	0	0	0	0
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	0	0	0	0
5	0	0	0	0
6	0	0	0	0
7	0	0	0	0
8	0	0	0	0
9	0	0	0	0
10	0	0	0	0
11	0	0	0	0
12	0	-0.114	0	0
13	0	0	0.080	0
14	0.026	0	0.073	0
15	0	0	0.101	0
16	0	0	0.083	-0.071
17	0	0	0	0
18	0.022	0	0	0
19	0.129	0	0	0
20	0.030	0	0	-0.170
21	0.094	0	0	0
22	0	0	0	0
23	0	0	-0.214	0.074
24	0.203	0	0	0
25	0.129	0	0	0
26	0	0	0.108	-0.349
27	0	0.235	0	-0.145
28	0	0	0.239	-0.301
29	0	-0.126	0.353	-0.222
30 (PAM-N)	0	0	0	-0.156
31 (PAM-G)	0	0	0	0
32 (PAM-G)	0	0	0	0
33	0	0.180	0	0
34	0	0	0	0
35	0	0	0	0
36	0	0	0	-0.117
37	0	0	0	0
38	0	0	0	0
39	0	0	0	0

REFERENCES

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

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

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.

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

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

Zhou, X*., et al.* The Human Epigenome Browser at Washington University. *Nat Methods* 2011;8(12):989-990.

Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2005;67(2):301-320.