

Supplementary Figures and Tables

**CRISPR-DT: designing gRNAs for the CRISPR-Cpf1 system with improved target efficiency and specificity**

Houxiang Zhu<sup>1</sup> and Chun Liang<sup>1,2\*</sup>

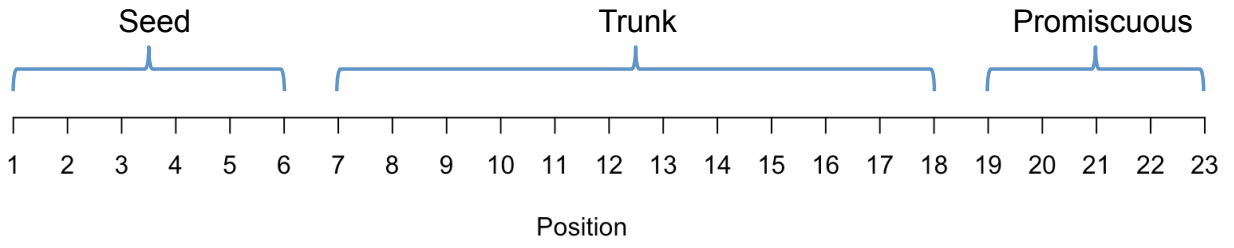
<sup>1</sup>Department of Biology, Miami University, Oxford, OH 45056, USA

<sup>2</sup>Department of Computer Science and Software Engineering, Miami University, Oxford, OH 45056, USA

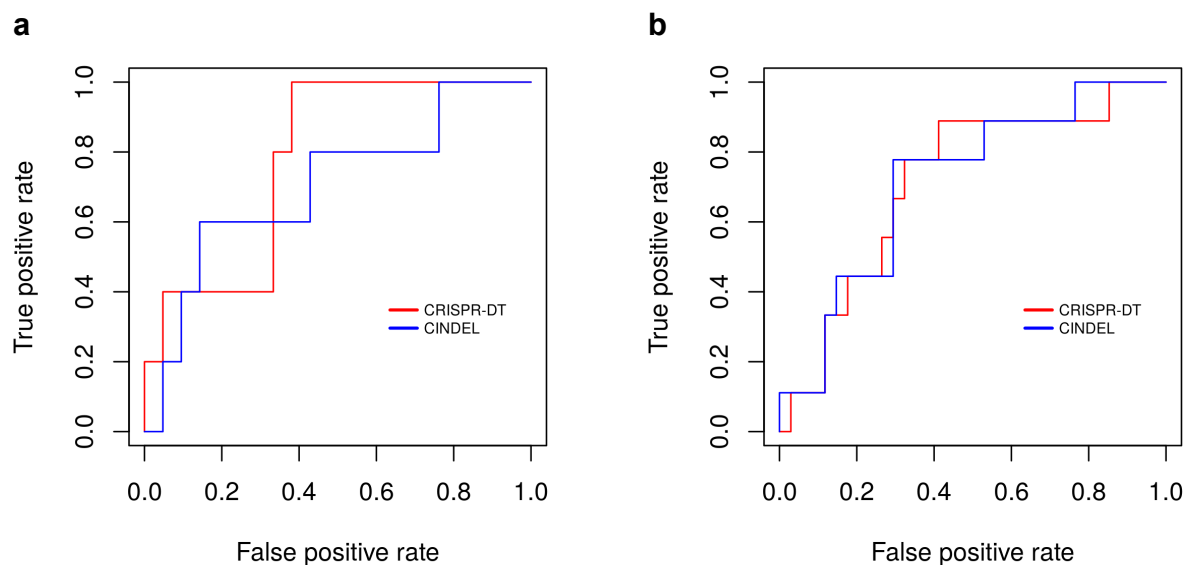
\* Correspondence: [liangc@miamioh.edu](mailto:liangc@miamioh.edu)

HZ: [zhuh4@miamioh.edu](mailto:zhuh4@miamioh.edu)

CL: [liangc@miamioh.edu](mailto:liangc@miamioh.edu)



**Figure S1** Three different regions of the 23-nt gRNA sequence for the CRISPR-Cpf1 system.



**Figure S2** Validation of our final SVM model using independent experimental data. **a** ROC curves comparing the performance of our model (CRISPR-DT) and the Kim et al. (2017) model (CINDEL) in predicting target efficiency using independent gRNAs with the TTTC PAM. **b** ROC curves comparing the performance of our model (CRISPR-DT) and the Kim et al. (2017) model (CINDEL) in predicting target efficiency using independent gRNAs with the TTTG PAM.

# CRISPR-DT: CRISPR DNA-Targeting Prediction and Visualization

Home  
Cpf1  
Cas9  
Cas9n  
RFN

## CRISPR-Cpf1 System

### Step 1: Provide a DNA sequence to target

Input a DNA sequence in FASTA format: [?](#) Example Sequence

### Step 2: Choose the reference genome

Reference genome:  [?](#)

### Step 3: Describe the PAM requirement

PAM sequence [?](#)

On target:  Off target:

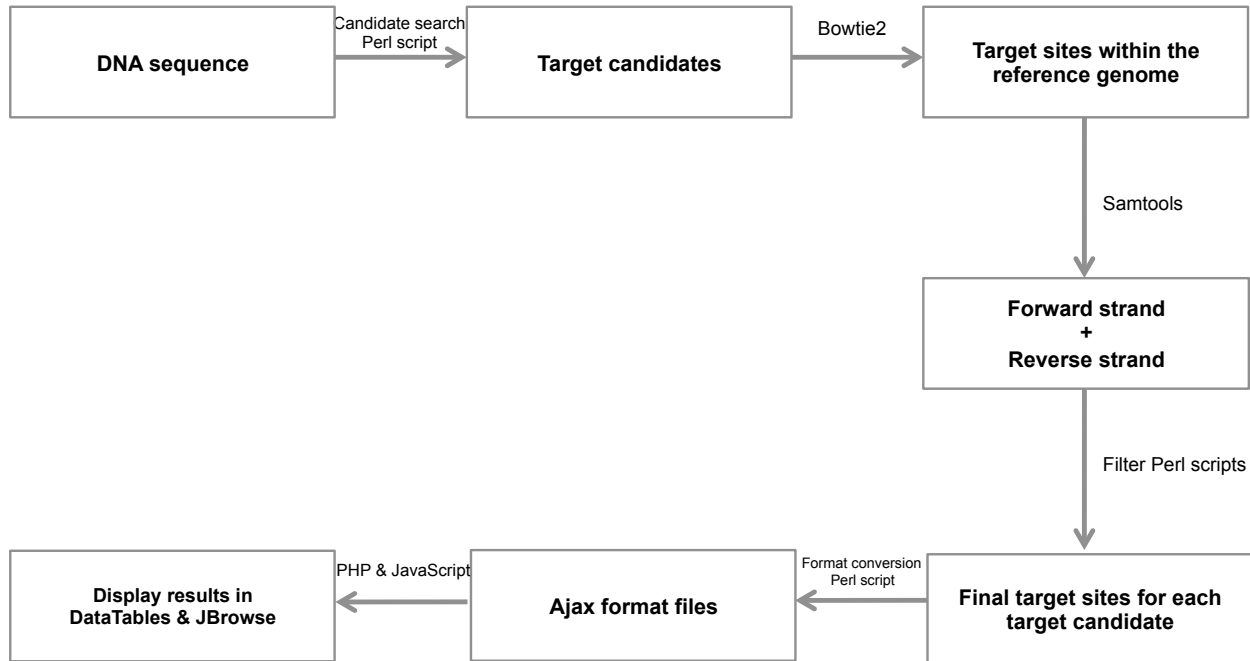
### Step 4: Choose an off-target setting

Basic settings [\(show options\)](#)

Specific settings [\(show options\)](#)

[Advanced Options](#)

**Figure S3** The parameter setting page of the CRISPR-Cpf1 system. First, users input a DNA sequence in FASTA format into the text area. Second, users select a reference genome. Third, users set up the PAM requirement. Fourth, users choose an off-target setting (either “Basic settings” or “Specific settings”).



**Figure S4** The flowchart of CRISPR-DT backend pipeline. CRISPR-DT first processes user input DNA sequence and then displays results in the interfaces.

**Table S1** Significant position-specific dinucleotide probability within the 23-nt gRNA sequences

Position-specific Dinucleotides	Efficient gRNAs	Inefficient gRNAs	Enrichment Ratio <sup>a</sup>	<i>P</i> -value <sup>b</sup>
Position_1_CC	0.26	0.06	4.57	1.12E-05
Position_1_UC	0.01	0.16	0.05	1.51E-05
Position_1_UU	0.00	0.12	0.00	7.08E-05
Position_19_GG	0.04	0.18	0.22	2.92E-04
Position_3_UG	0.16	0.03	5.00	5.78E-04
Position_5_CG	0.02	0.13	0.13	6.16E-04
Position_1_UG	0.01	0.10	0.08	9.91E-04
Position_8_AC	0.14	0.03	4.50	1.76E-03
Position_17_GC	0.01	0.10	0.08	1.77E-03
Position_3_GU	0.02	0.13	0.19	1.91E-03
Position_22_GG	0.05	0.17	0.29	2.20E-03
Position_13_UU	0.00	0.07	0.00	2.38E-03
Position_2_GG	0.02	0.12	0.20	3.33E-03
Position_6_AC	0.16	0.05	3.33	3.68E-03
Position_12_GG	0.05	0.16	0.30	3.68E-03
Position_1_GG	0.17	0.06	3.00	4.94E-03
Position_5_CA	0.13	0.03	4.00	5.15E-03
Position_18_GG	0.02	0.11	0.21	5.75E-03
Position_20_GC	0.02	0.11	0.21	5.75E-03
Position_4_UG	0.05	0.15	0.32	6.10E-03
Position_3_UA	0.06	0.00	N/A	7.63E-03
Position_8_GG	0.06	0.16	0.35	8.04E-03
Position_18_CG	0.02	0.10	0.23	9.84E-03
Position_16_GG	0.05	0.14	0.33	9.98E-03
Position_22_CG	0.05	0.14	0.33	9.98E-03
Position_9_CU	0.07	0.01	9.00	1.00E-02
Position_21_UA	0.07	0.01	9.00	1.00E-02
Position_13_UA	0.09	0.02	5.50	1.05E-02
Position_15_GU	0.02	0.09	0.18	1.05E-02
Position_22_UC	0.15	0.06	2.71	1.29E-02
Position_8_AA	0.05	0.00	N/A	1.37E-02
Position_6_GU	0.03	0.11	0.29	1.45E-02
Position_1_GA	0.10	0.02	4.00	1.67E-02
Position_16_CC	0.10	0.02	4.00	1.67E-02
Position_21_CG	0.02	0.10	0.25	1.67E-02
Position_11_AA	0.06	0.01	8.00	1.78E-02
Position_7_UG	0.03	0.10	0.31	2.40E-02
Position_21_CC	0.03	0.10	0.31	2.40E-02
Position_8_UG	0.05	0.13	0.38	2.57E-02
Position_20_GG	0.05	0.13	0.38	2.57E-02

Position_1_GU	0.09	0.02	3.67	2.81E-02
Position_7_CC	0.09	0.02	3.67	2.81E-02
Position_17_UU	0.09	0.02	3.67	2.81E-02
Position_4_UA	0.07	0.02	4.50	3.13E-02
Position_6_UA	0.07	0.02	4.50	3.13E-02
Position_7_CA	0.07	0.02	4.50	3.13E-02
Position_2_UG	0.01	0.06	0.14	3.17E-02
Position_3_AA	0.01	0.06	0.14	3.17E-02
Position_16_UU	0.01	0.06	0.14	3.17E-02
Position_9_GC	0.04	0.11	0.36	3.20E-02
Position_14_GU	0.06	0.14	0.41	3.20E-02
Position_22_AC	0.17	0.08	2.10	3.50E-02
Position_10_CG	0.03	0.10	0.33	3.91E-02
Position_11_GG	0.05	0.12	0.40	4.05E-02
Position_18_AU	0.03	0.00	N/A	4.51E-02
Position_1_AG	0.02	0.08	0.30	4.67E-02
Position_12_AG	0.08	0.02	3.33	4.67E-02
Position_12_UU	0.02	0.08	0.30	4.67E-02

<sup>a</sup>The enrichment ratio was calculated through dividing the average dinucleotide probability of efficient gRNAs by that of inefficient gRNAs.

<sup>b</sup>The *P*-value was determined by Welch's t-test ( $P < 0.05$ ).

**Table S2** Significant position-nonspecific trinucleotide and tetranucleotide counts within the 23-nt gRNA sequences

Tri- and Tetranucleotides	Efficient gRNAs	Inefficient gRNAs	Enrichment Ratio <sup>a</sup>	<i>P</i> -value <sup>b</sup>
Count_UGG	0.32	0.75	0.43	1.83E-05
Count_CCU	0.58	0.27	2.15	1.10E-04
Count_CCA	0.63	0.31	2.03	8.88E-04
Count_UAU	0.19	0.05	4.00	1.08E-03
Count_GCG	0.10	0.34	0.30	1.33E-03
Count_ACC	0.70	0.40	1.76	1.46E-03
Count_CUC	0.57	0.32	1.78	1.64E-03
Count_CCG	0.25	0.49	0.51	1.71E-03
Count_GUG	0.35	0.64	0.55	2.16E-03
Count_UAC	0.40	0.19	2.08	2.83E-03
Count_CGG	0.27	0.53	0.52	4.61E-03
Count_GGC	0.22	0.47	0.47	6.28E-03
Count_GGU	0.46	0.74	0.62	6.54E-03
Count_AUA	0.14	0.03	4.25	7.07E-03
Count_GGG	0.27	0.48	0.57	8.09E-03
Count_CGU	0.26	0.46	0.56	1.38E-02
Count_AAC	0.38	0.22	1.78	1.49E-02
Count_GUU	0.24	0.40	0.60	1.71E-02
Count_UAA	0.18	0.07	2.56	1.95E-02
Count_GAA	0.33	0.18	1.86	2.24E-02
Count_GUA	0.41	0.25	1.65	2.64E-02
Count_CAU	0.30	0.17	1.76	2.73E-02
Count_UUU	0.08	0.18	0.43	3.20E-02
Count_AGA	0.34	0.21	1.65	3.37E-02
Count_CUA	0.27	0.16	1.70	3.57E-02
Count_UCG	0.32	0.48	0.67	4.36E-02
Count_CCCA	0.17	0.02	7.00	3.84E-04
Count_GGGG	0.05	0.22	0.21	4.48E-04
Count_GUGG	0.10	0.29	0.36	8.01E-04
Count_GCCG	0.01	0.09	0.09	3.17E-03
Count_CGUU	0.02	0.12	0.20	3.33E-03
Count_ACAU	0.10	0.02	6.50	3.43E-03
Count_UGGG	0.07	0.22	0.33	3.84E-03
Count_CUGG	0.06	0.18	0.35	3.94E-03
Count_CCUU	0.16	0.05	3.33	5.36E-03
Count_AAGA	0.10	0.02	6.00	6.02E-03
Count_AGCG	0.02	0.10	0.17	6.02E-03
Count_GGUG	0.11	0.26	0.42	7.25E-03



Count_UUCG	0.04	0.14	0.29	7.38E-03
Count_CCAA	0.12	0.02	5.00	8.36E-03
Count_ACCA	0.20	0.07	2.78	8.84E-03
Count_GACC	0.21	0.09	2.36	9.53E-03
Count_UGGC	0.02	0.10	0.23	9.84E-03
Count_GCUU	0.02	0.10	0.23	9.84E-03
Count_GUAU	0.07	0.01	9.00	1.00E-02
Count_AGAA	0.09	0.02	5.50	1.05E-02
Count_CCCU	0.14	0.04	3.60	1.23E-02
Count_CCUC	0.22	0.10	2.15	1.27E-02
Count_CUAU	0.05	0.00	N/A	1.37E-02
Count_CCGG	0.06	0.17	0.33	1.47E-02
Count_CCCG	0.05	0.15	0.32	1.50E-02
Count_AGGU	0.08	0.19	0.42	1.54E-02
Count_CCAU	0.10	0.02	4.33	1.54E-02
Count_GUAA	0.08	0.02	5.00	1.82E-02
Count_ACUC	0.16	0.06	2.50	2.07E-02
Count_AGUU	0.03	0.10	0.31	2.40E-02
Count_GGGC	0.06	0.17	0.38	2.52E-02
Count_GAGC	0.06	0.14	0.39	2.64E-02
Count_GGAA	0.09	0.02	3.67	2.81E-02
Count_UCGG	0.06	0.14	0.41	3.20E-02
Count_UGGU	0.10	0.22	0.48	3.42E-02
Count_GCGG	0.02	0.11	0.21	3.79E-02
Count_ACUU	0.10	0.03	3.00	3.91E-02
Count_UACA	0.09	0.02	3.67	4.08E-02
Count_UCCG	0.04	0.11	0.36	4.21E-02
Count_CGGG	0.07	0.16	0.45	4.37E-02
Count_AUAU	0.03	0.00	N/A	4.51E-02
Count_UAAA	0.03	0.00	N/A	4.51E-02
Count_UAUA	0.03	0.00	N/A	4.51E-02
Count_UAAC	0.08	0.02	3.33	4.67E-02
Count_CUCC	0.13	0.06	2.29	4.93E-02

<sup>a</sup> The enrichment ratio was calculated through dividing the average nucleotide counts of efficient gRNAs by that of inefficient gRNAs.

<sup>b</sup> The *P*-value was determined by Welch's t-test ( $P < 0.05$ ).

**Table S3** The 48 most important features used to build the final SVM model.

Feature Name	Feature Type	Mean Gini
Position_1_U	Binary	4.58
Minimum_Free_Energy	Numerical	3.90
Count_G	Numerical	2.45
MT_Promiscuous	Numerical	2.43
MT_Entire	Numerical	2.40
Count_A	Numerical	1.79
Position_1_G	Binary	1.32
Count_C	Numerical	1.28
MT_Trunk	Numerical	1.17
Count_GG	Numerical	1.03
Position_23_C	Binary	1.00
Position_1_UC	Binary	0.98
Position_1_CC	Binary	0.97
Count_CG	Numerical	0.95
GC_Entire	Numerical	0.81
MT_Seed	Numerical	0.80
Count_CU	Numerical	0.79
Position_1_UU	Binary	0.68
Count_AC	Numerical	0.67
GC_Promiscuous	Numerical	0.65
Position_1_C	Binary	0.65
Count_CC	Numerical	0.64
Count_UGG	Numerical	0.64
Position_23_G	Binary	0.64
Count_CCU	Numerical	0.62
Count_U	Numerical	0.61
Position_19_G	Binary	0.55
Count_UA	Numerical	0.53
Count_GC	Numerical	0.53
Count_CCA	Numerical	0.52
GC_Trunk	Numerical	0.52
Count_GU	Numerical	0.52
Count_GA	Numerical	0.49
Position_3_UG	Binary	0.49
Count_AG	Numerical	0.49
GC_Seed	Numerical	0.49
Position_3_U	Binary	0.48
Position_9_G	Binary	0.48
Position_1_GG	Binary	0.46
Count_CCG	Numerical	0.45

Count_UU	Numerical	0.45
Count_CA	Numerical	0.44
Count_CUC	Numerical	0.43
Position_3_GU	Binary	0.42
Count_AU	Numerical	0.42
Count_GCG	Numerical	0.42
Count_ACC	Numerical	0.42
Count_CGG	Numerical	0.40

---