

**YMTHE, Volume 28**

**Supplemental Information**

**Computational Analysis Concerning  
the Impact of DNA Accessibility  
on CRISPR-Cas9 Cleavage Efficiency**

**Cheng-Han Chung, Alexander G. Allen, Neil T. Sullivan, Andrew Atkins, Michael R. Nonnemacher, Brian Wigdahl, and Will Dampier**

1 **Supplemental information**

2 Table of contents:

3 **Figure S1.** CRISPR-Cas9 targets more accessible regions in either HEK293T or U2OS  
4 cells

5 **Figure S2.** The distributions of DNA accessibility at cleavage sites are similar across  
6 individual gRNAs

7 **Figure S3.** DNA accessibility impacts CRISPR-induced cleavage frequency among  
8 cleavage sites with high sequence similarity

9 **Figure S4.** The correlation between gRNA:target sequence similarity and CRISPR-  
10 induced cleavage frequency is not affected by DNA accessibility in CS only subset

11 **Figure S5.** Higher proportion of CRISPR-induced cleavage sites are located at regions  
12 with low DNA accessibility than that of endogenous gene loci

13 **Figure S6.** The gene expression profiles are positively correlated between untreated  
14 HEK293T and U2OS cells

15 **Figure S7.** The DNA accessibility abrogates the correlation between gRNA:target  
16 sequence similarity and CRISPR-induced cleavage frequency when the chromosomal  
17 regions are less accessible in GS and CS subset but not CS only subset in HEK293T  
18 cells

19 **Figure S8.** The DNA accessibility abrogates the correlation between gRNA:target  
20 sequence similarity and CRISPR-induced cleavage frequency when the chromosomal  
21 regions are less accessible in GS and CS subset but not CS only subset in U2OS cells

22 **Figure S9.** Chromatin accessibility required for CRISPR-mediated cleavage reaction is  
23 significantly less than that for endogenous gene to express

24 **Table S1.** The interaction between CFD score and DNA accessibility does not impact  
25 the CRISPR-induced cleavage frequency in CS only subset

26 **Table S2.** Frequency table of GUIDE-seq detected cleavage sites by individual gRNAs

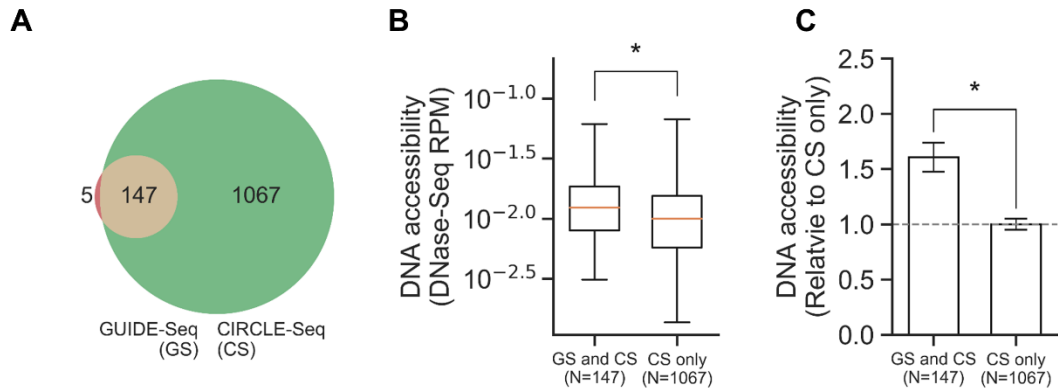
27 **Table S3.** Frequency table of CIRCLE-seq detected cleavage sites by individual gRNAs

28 **Table S4.** Counts of CRISPR-mediated cleavage sites intersected between GS and CS  
29 datasets

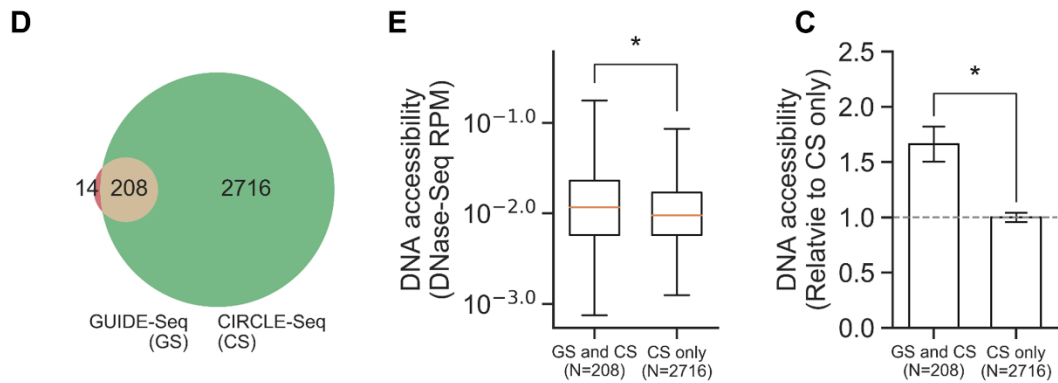
30 **Table S5.** List of cleavage sites and corresponding characteristics including CPM, RPM,  
31 CFD score detected by GUIDE-seq

32 **Table S6.** List of cleavage sites and corresponding characteristics including CPM, RPM,  
33 CFD score detected by CIRCLE-seq

## HEK293T



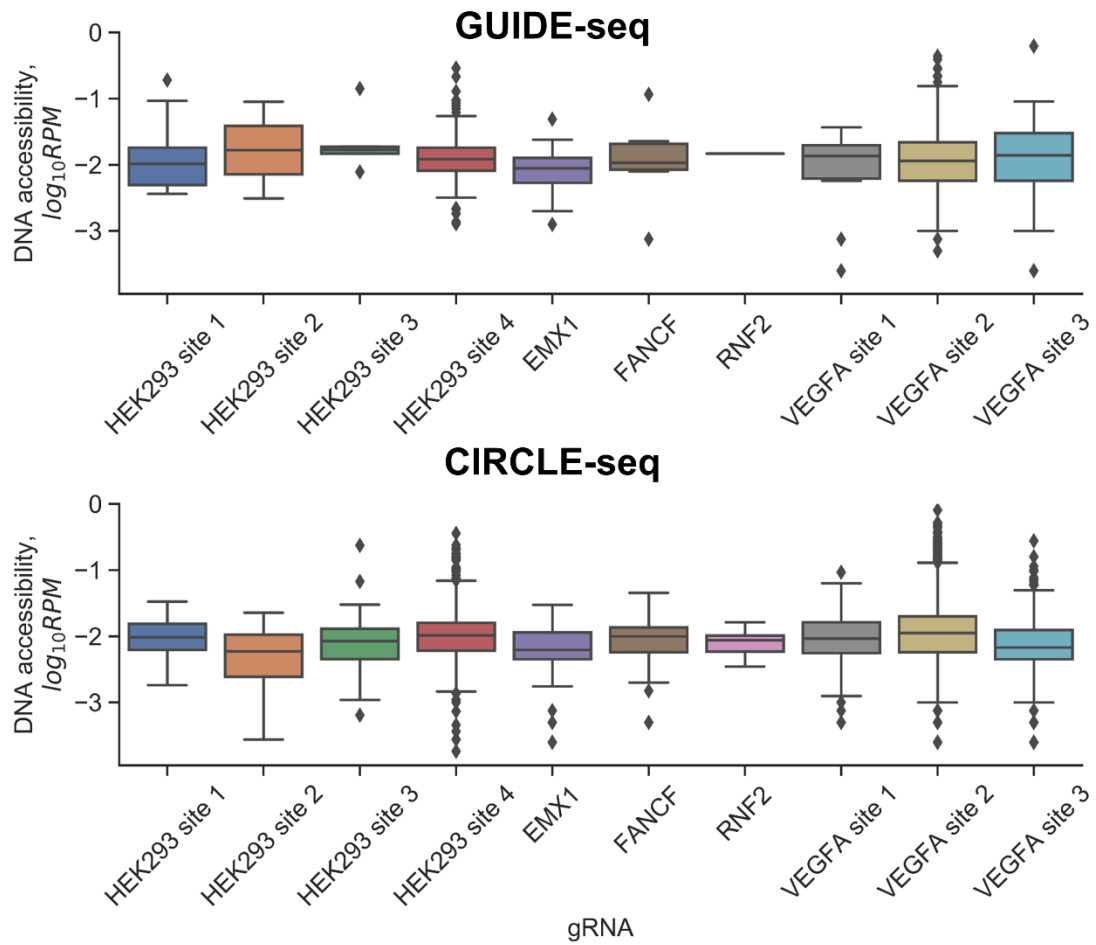
## U2OS



35

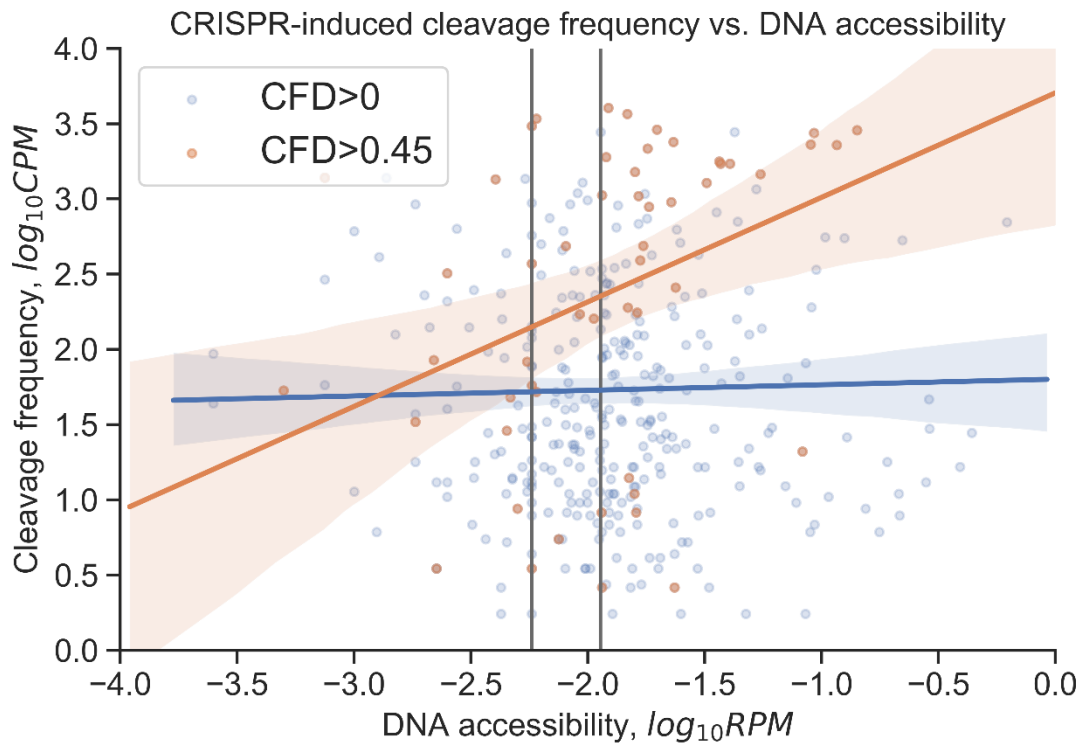
36 **Figure S1. CRISPR-Cas9 targeted more accessible regions in either HEK293T or**  
 37 **U2OS cells.** (A and D) The Venn diagram displays the number of cleavage sites  
 38 identified by both GUIDE-seq (GS) and CIRCLE-seq (CS) for the indicated cell type. (B  
 39 and E) The cleavage sites that both GS and CS identified (GS and CS) shows higher  
 40 DNA accessibility than those sites only identified by CS (CS only). The DNA accessibility  
 41 of cleavage sites were the DNase-seq read depth per million mapped reads within 50 bp  
 42 window flanking by the DSB positions (termed RPM). \* p-value < 0.001 two-tailed t-test.  
 43 (C and F) The DNA accessibility normalized to the mean DNase-seq RPM of CS only  
 44 subset. \* p-value < 0.001 two-tailed t-test.

45



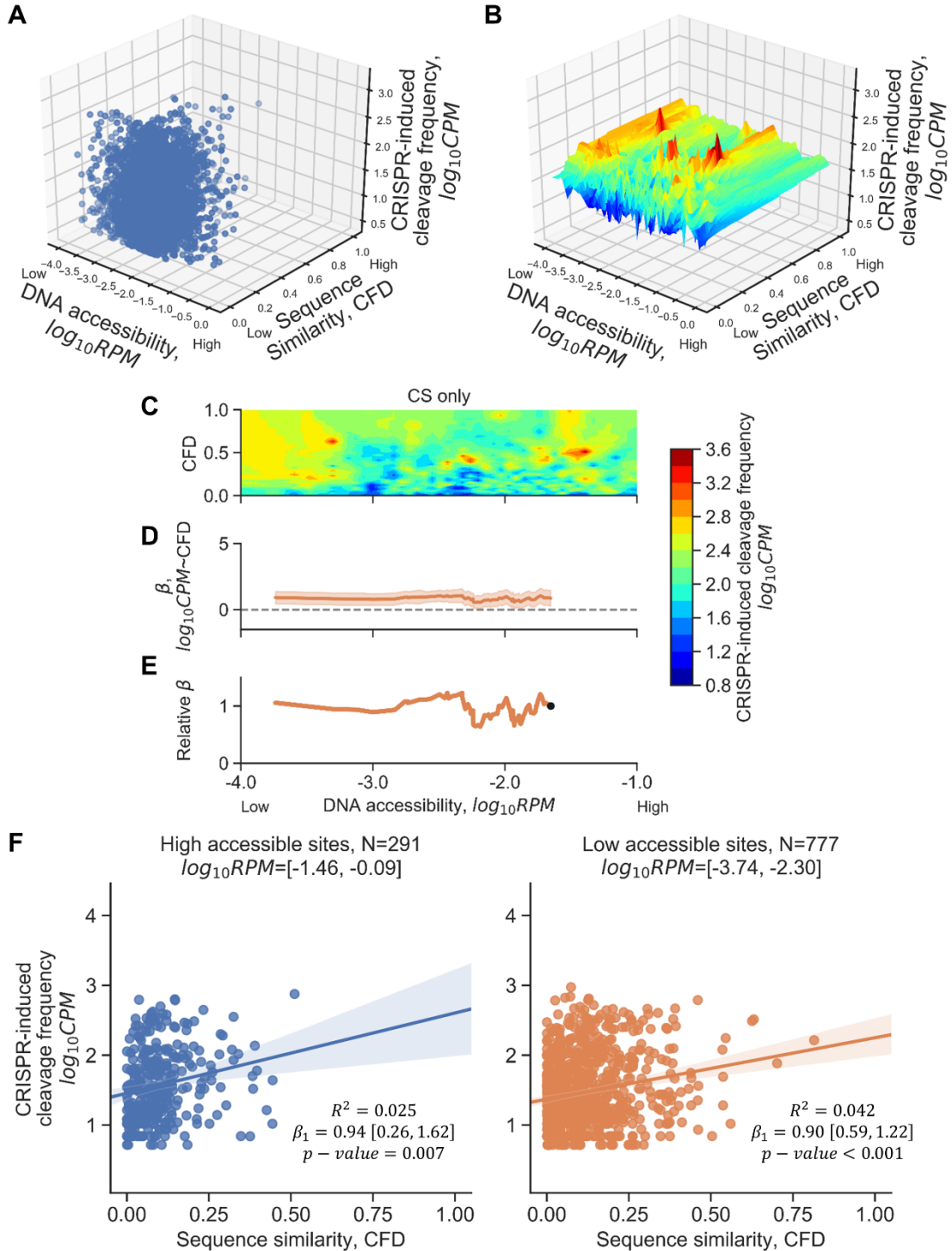
46

47 **Figure S2. The distributions of DNA accessibility at cleavage sites were similar**  
 48 **across individual gRNAs.** The box plot shows the distribution of DNA accessibility for  
 49 individual gRNAs in both assays. The box represents 50% quantile and the line inside  
 50 the box represents the median.



51

52 **Figure S3. DNA accessibility impacts CRISPR-induced cleavage frequency among**  
 53 **cleavage sites with high sequence similarity.** Cleavage sites with high sequence  
 54 similarity were selected as the top 15% of ranked CFD (N=53) in GS and CS subset,  
 55 which contains cleavage sites with CFD>0.45 (orange). These cleavage sites show  
 56 positive correlation between DNA accessibility and CRISPR-induced cleavage frequency.  
 57 This relationship was not observed in the correlation test using all data points in GS and  
 58 CS subset (N=355). This result indicates that even with high sequence similarity, low  
 59 DNA accessibility reduces CRISPR-induced cleavage frequency.

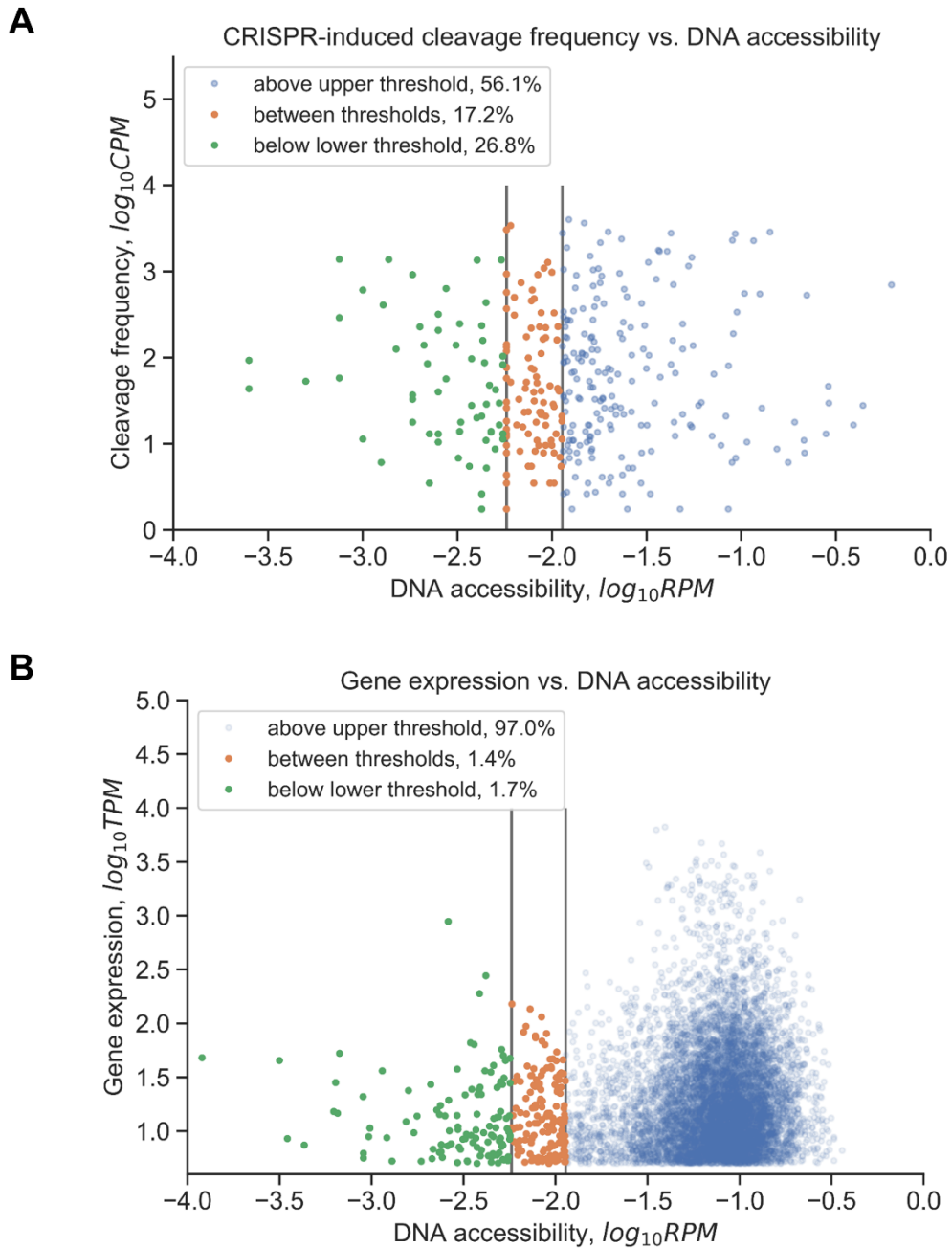


61

62 **Figure S4. The correlation between gRNA:target sequence similarity and CRISPR-**  
 63 **induced cleavage frequency was not affected by DNA accessibility in CS only**  
 64 **subset (N=3783). (A) The three-dimensional scatter plot of sequence similarity, DNA**

65 accessibility and CRISPR-induced cleavage frequency using the CRISPR-induced  
66 cleavage sites listed in the CS only subset. Each dot represents a CRISPR-induced  
67 cleavage site identified by CIRCLE-seq and absent in GUIDE-seq result. CPM  
68 represents the number of cleavage events at a CRISPR-induced cleavage site detected  
69 by CIRCLE-seq; Sequence similarity represents the likelihood of CRISPR cutting based  
70 on the sequence between gRNA and target using CFD matrix; RPM represents the DNA  
71 accessibility at a CRISPR-induced cleavage site. (B) The surface plot estimated by the  
72 nearest-neighbor method described in the Methods. The sequence similarity is  
73 estimated by the position-specific matrix of Cutting Frequency Determination (CFD)  
74 score [0,1] that describes the cleavage possibility of gRNA:target pair at the off-target  
75 sites. Red color represents high cleavage frequency represented in CPM while blue  
76 color represents low cleavage frequency identified by the CIRCLE-seq technique. (C)  
77 Contour map of CRISPR-induced cleavage frequency based on the grids of CFD score  
78 and DNase-seq RPM; a top-down view of (B). (D) The beta coefficient between CFD and  
79 CRISPR-induced cleavage frequency at given 15% quantile of DNA accessibility. Note  
80 that the data point was the lower boundary of a given quantile. The shaded regions  
81 represent 95% confidence intervals of the t-test. The horizontal dashed line at beta  
82 coefficient equal to 0 represents the threshold of the significance of the beta coefficient.  
83 The correlation was not significant when the 95% confidence interval covers the  
84 horizontal line. (E) The beta coefficient relative to the first quantile that contained the  
85 cleavage sites with the top 15% DNA accessibility in the CS only subset. The dashed  
86 line represents the regions that were not significant in the Wald Test (D). Note that the  
87 CS only subset does not have insignificant quantile therefore no dashed line was  
88 indicated. (F) Correlation between CRISPR-induced cleavage frequency and CFD score  
89 of 15% most accessible sites (left panel) or 15% least accessible sites (right panel) in  
90 the CS only subset.  $\beta_1$ : beta coefficient of simple linear regression. p-value of Wald Test  
91 for a hypothesis test that the slope is 0.

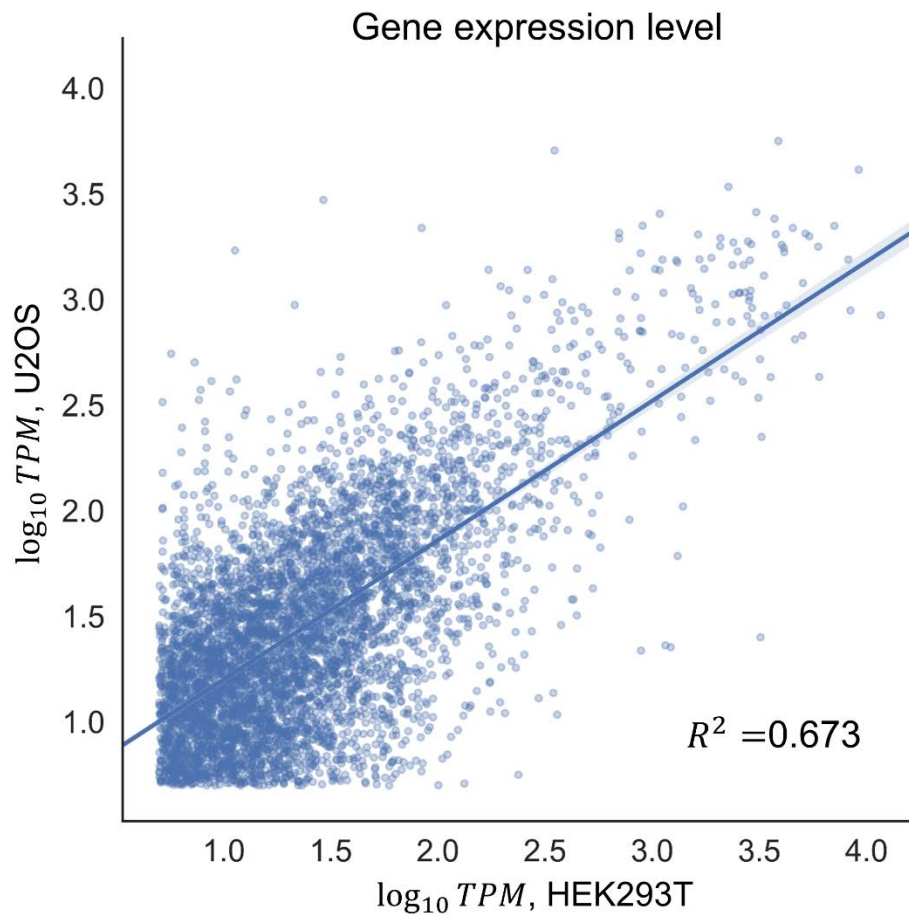
92



93

94 **Figure S5. Higher proportion of CRISPR-induced cleavage sites were located at**  
 95 **regions with low DNA accessibility as compared to that of endogenous gene loci.**  
 96 (A) Scatter plot of CRISPR-induced cleavage frequency measured by GUIDE-seq and  
 97 DNA accessibility measured by DNase-seq in both HEK293T and U2OS cells using GS  
 98 and CS subset. Vertical lines correspond to the thresholds as determined in Figure 4. (B)  
 99 Scatter plot of gene expression level measured by RNA-seq and DNA accessibility  
 100 measure by DNase-seq in untreated HEK293T and U2OS cells. Expressed gene was  
 101 defined as any protein-coding genes with > 5 TPM. Gray vertical lines represent the  
 102 thresholds where DNA accessibility abrogates the significance between CFD and  
 103 CRISPR-induced cleavage frequency, which were adopted from Figure 4C.

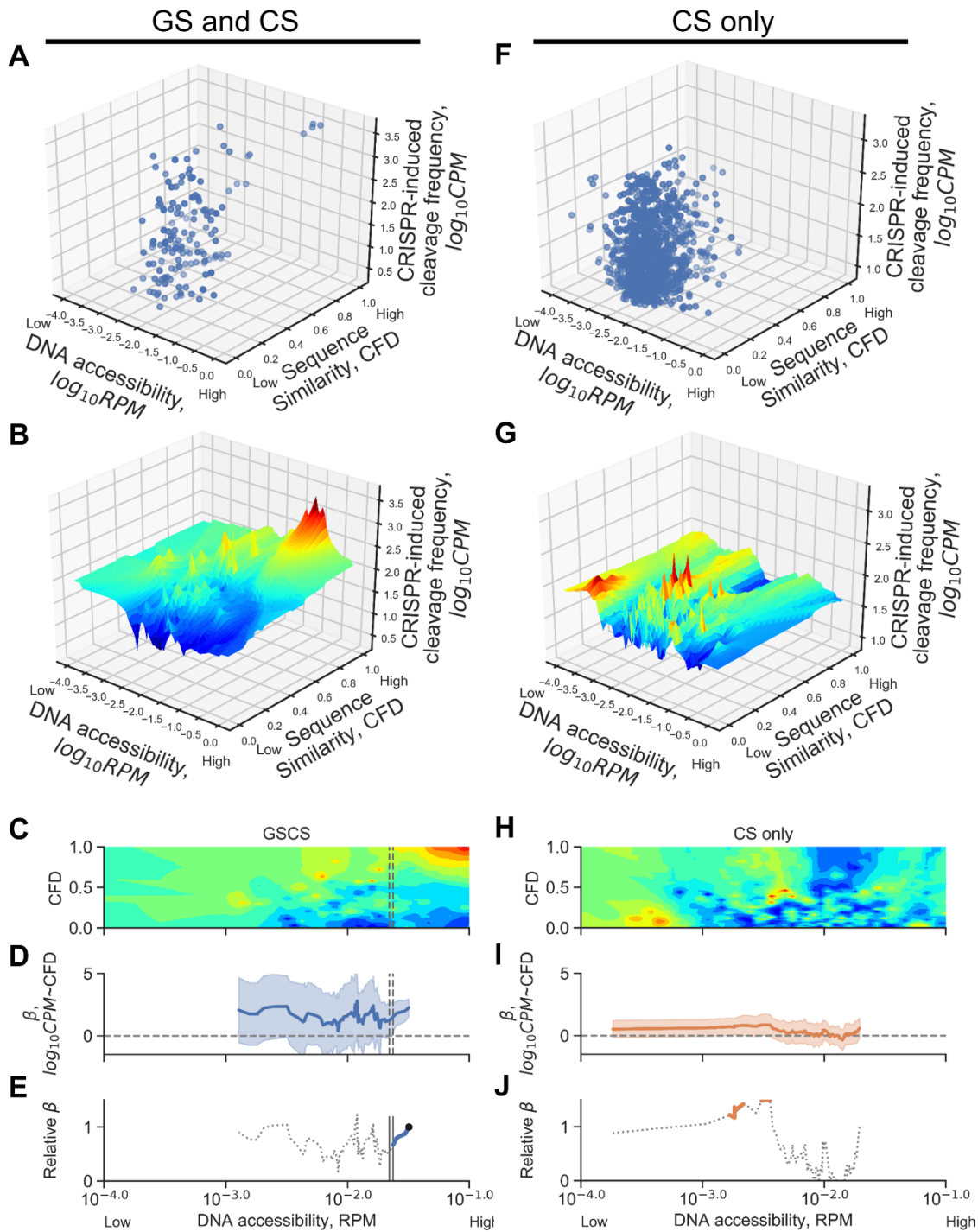




104

105 **Figure S6. The gene expression profiles were positively correlated between**  
106 **untreated HEK293T and U2OS cells (N=8619).** Transcripts with predicted expression  
107 level above 5 TPM in both cells were included in this analysis. The R-square was  
108 estimated by Pearson correlation coefficient test, p-value<0.001.

HEK293T

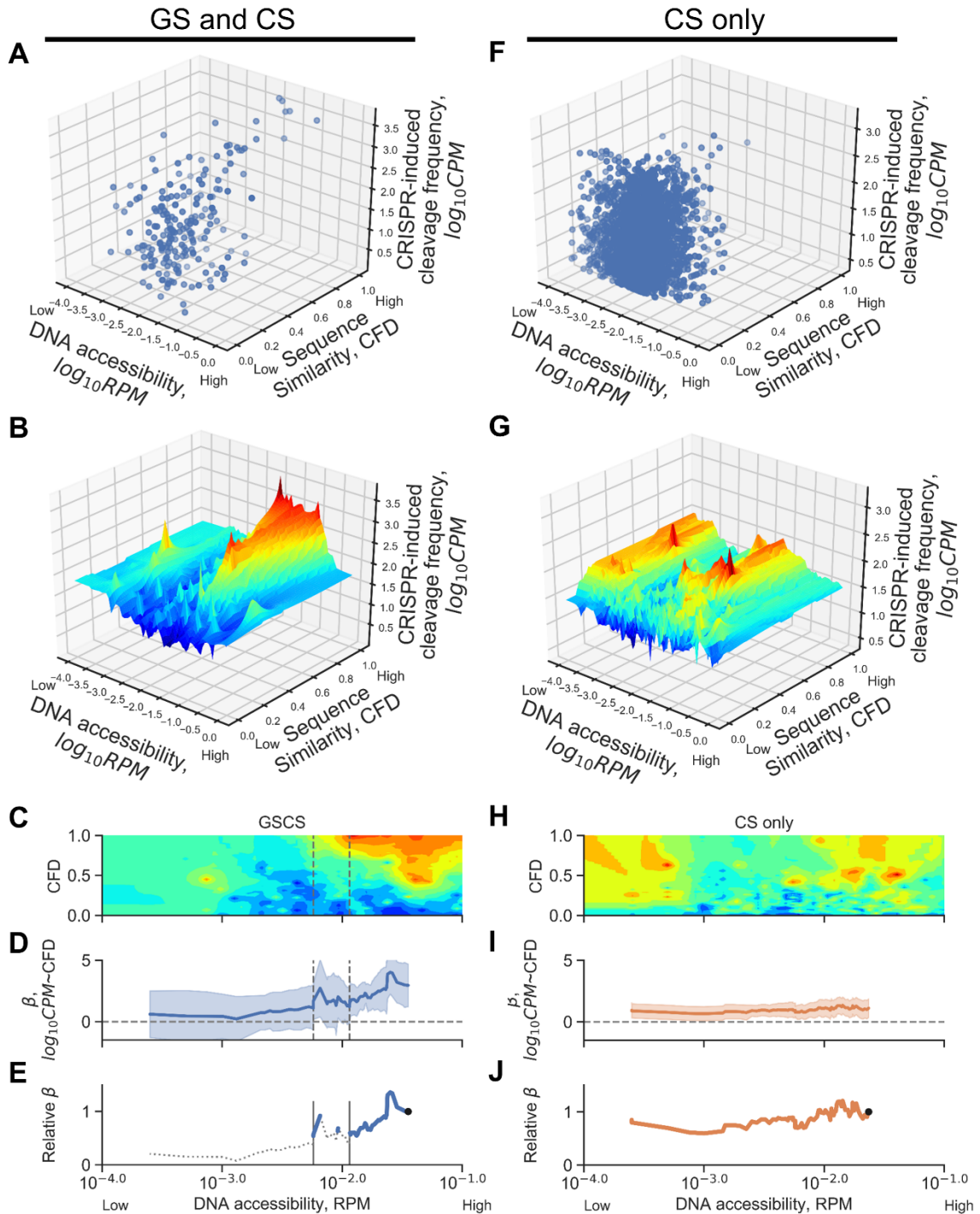


109

110 **Figure S7. The DNA accessibility abrogated the correlation between gRNA:target**  
 111 **sequence similarity and CRISPR-induced cleavage frequency when the**  
 112 **chromosomal regions are less accessible in GS and CS subset but not CS only**  
 113 **subset in HEK293T cells. The three-dimensional scatter plot of sequence similarity,**  
 114 **DNA accessibility and CRISPR-induced cleavage frequency using the CRISPR-induced**

115 cleavage sites listed in either the GS and CS subset (A) or CS only subset (F). Each dot  
116 represents a CRISPR-induced cleavage site identified by both GUIDE-seq and CIRCLE-  
117 seq. CPM represents the number of cleavage events at a CRISPR-induced cleavage  
118 site detected by GUIDE-seq; sequence similarity represents the likelihood of CRISPR  
119 cutting based on the sequence between gRNA and target using CFD matrix; RPM  
120 represents the DNA accessibility at a CRISPR-induced cleavage site. (B, G) The surface  
121 plot estimated by the nearest-neighbor method described in the Methods. The sequence  
122 similarity was estimated by the position-specific matrix of Cutting Frequency  
123 Determination (CFD) score [0,1] that described the cleavage possibility of gRNA:target  
124 pair at the off-target sites. Red color represents high cleavage frequency while blue color  
125 represents low cleavage frequency identified by the GUIDE-seq technique. (C) Contour  
126 map of CRISPR-induced cleavage frequency based on the grids of CFD score and  
127 DNase-seq RPM derived from Fig. 3B using the GS and CS subsets. (D) The beta  
128 coefficient between CFD and CRISPR-induced cleavage frequency at given 15%  
129 quantiles of DNA accessibility. Note that the data point was the lower boundary of a  
130 given quantile. The shaded regions represent 95% confidence intervals of the t-test. The  
131 horizontal dashed line at beta coefficient equal to 0 represents the threshold of the  
132 significance of the beta coefficient. The correlation was not significant when the 95%  
133 confidence interval covers the horizontal line. (E) The beta coefficient relative to the first  
134 quantile that contains the cleavage sites with the top 15% DNA accessibility in GS and  
135 CS subset. The dashed line represents the regions that were not significant in the  
136 Pearson correlation coefficient test (D). The right vertical lines represent the threshold of  
137 DNA accessibility that started to affect the significance between CFD and CRISPR-  
138 induced cleavage frequency. The left vertical line represents the threshold such that the  
139 correlation between homology and CRISPR-induced cleavage efficiency is insignificant  
140 anywhere below the DNA accessibility. (H, I, J) The equivalent analysis using the CS  
141 only subset. The  $\beta$  between gRNA:target homology and CRISPR-induced cleavage  
142 frequency is always significant across different DNA accessibility.

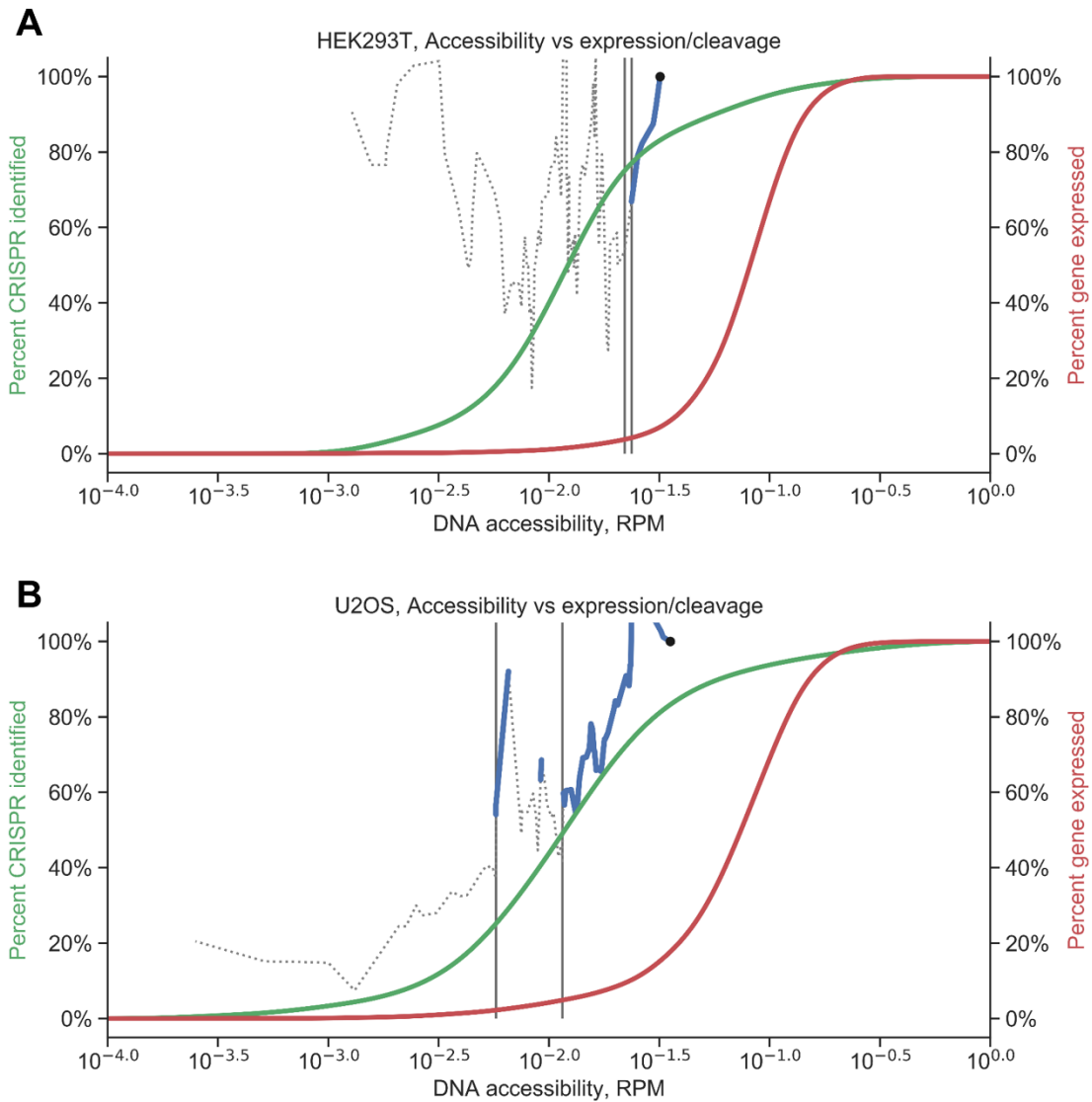
U2OS



143

144 **Figure S8. The DNA accessibility abrogated the correlation between gRNA:target**  
 145 **sequence similarity and CRISPR-induced cleavage frequency when the**  
 146 **chromosomal regions are less accessible in GS and CS subset but not CS only**  
 147 **subset in U2OS cells. The three-dimensional scatter plot of sequence similarity, DNA**  
 148 **accessibility and CRISPR-induced cleavage frequency using the CRISPR-induced**

149 cleavage sites listed in either the GS and CS subset (A) or CS only subset (F). Each dot  
150 represents a CRISPR-induced cleavage site identified by both GUIDE-seq and CIRCLE-  
151 seq. CPM represents the number of cleavage events at a CRISPR-induced cleavage  
152 site detected by GUIDE-seq; Sequence similarity represents the likelihood of CRISPR  
153 cutting based on the sequence between gRNA and target using CFD matrix; RPM  
154 represents the DNA accessibility at a CRISPR-induced cleavage site. (B, G) The surface  
155 plot estimated by the nearest-neighbor method described in the Methods. The sequence  
156 similarity is estimated by the position-specific matrix of Cutting Frequency Determination  
157 (CFD) score [0,1] that describes the cleavage possibility of gRNA:target pair at the off-  
158 target sites. Red color represents high cleavage frequency while blue color represents  
159 low cleavage frequency identified by the GUIDE-seq technique. (C) Contour map of  
160 CRISPR-induced cleavage frequency based on the grids of CFD score and DNase-seq  
161 RPM derived from Fig. 3B using the GS and CS subset. (D) The beta coefficient  
162 between CFD and CRISPR-induced cleavage frequency at given 15% quantiles of DNA  
163 accessibility. Note that the data point is the lower boundary of a given quantile. The  
164 shaded regions represent 95% confidence intervals of the t-test. The horizontal dashed  
165 line at beta coefficient equal to 0 represents the threshold of the significance of the beta  
166 coefficient. The correlation is not significant when the 95% confidence interval covers the  
167 horizontal line. (E) The beta coefficient relative to the first quantile that contains the  
168 cleavage sites with the top 15% DNA accessibility in GS and CS subset. The dashed  
169 line represents the regions that were not significant in the Pearson correlation coefficient  
170 test (D). The right vertical lines represent the threshold of DNA accessibility that started  
171 to affect the significance between CFD and CRISPR-induced cleavage frequency. The  
172 left vertical line represents the threshold such that the correlation between homology and  
173 CRISPR-induced cleavage efficiency is insignificant anywhere below the DNA  
174 accessibility. (H, I, J) The equivalent analysis using the CS only subset. The correlation  
175 between gRNA:target homology and CRISPR-induced cleavage frequency is always  
176 significant across different DNA accessibility.



177

178 **Figure S9. Chromatin accessibility required for CRISPR-mediated cleavage**  
 179 **reaction was significantly less than that for endogenous gene to express.** (A)  
 180 Analysis using cleavage sites only identified in HEK293T cells in GS assay and  
 181 HEK293T RNA-seq. Green curve represents the cumulative percentage of CRISPR-  
 182 induced cleavage sites identified in GS and CS subset (N=152). Red curve represents  
 183 the cumulative percentage of expressed genes detected in HEK293T cells (N=7984).  
 184 Blue curve represents the relative  $\beta$  to the first quantile that contains the cleavage sites  
 185 with the top 15% DNA accessibility in GS and CS subset. The blue curve, gray lines and  
 186 thresholds were adopted from Figure S5E. (B) Analysis using cleavage sites only  
 187 identified in U2OS cells in GS assay and U2OS RNA-seq. Green curve represents the  
 188 cumulative percentage of CRISPR-induced cleavage sites identified in GS and CS  
 189 subset (N=222). Red curve represents the cumulative percentage of expressed genes  
 190 detected in HEK293T cells (N=7883). Blue curve represents the relative beta coefficient  
 191 to the first quantile that contains the cleavage sites with the top 15% DNA accessibility in

192 GS and CS subset. Gray vertical lines represent the thresholds where DNA accessibility  
193 abrogates the significance between CFD and CRISPR-induced cleavage frequency.  
194 The blue curve, gray lines and thresholds were adopted from Figure S6E.

195 **Table S1. The interaction between CFD score and DNA accessibility does not**  
 196 **impact the CRISPR-induced cleavage frequency in CS only subset.** The multiple  
 197 regression analysis was performed by adding independent variables and interaction of  
 198 independent variables sequentially to the models. CPM: number of cleavage events per  
 199 million mapped reads; CFD: nucleotide-specific scoring matrix for gRNA:target pair.  
 200 RPM: DNase-seq read depth per million mapped reads within 50 bp window flanking by  
 201 the DSB positions. <sup>†</sup>Adjusted R-square was used to adjust the correlation coefficient by  
 202 accounting for the number of independent variables each model has. \*: The beta  
 203 coefficient is significantly different from zero under **t-test** with a two-tailed p-value<0.05.

204

Model	Parameters	p-values	Adjusted R-square
$\log_{10} CPM \sim CFD$	Sequence similarity	<0.001*	0.027
$\log_{10} CPM \sim \log_{10} RPM$	DNA accessibility	0.113	0.0004
$\log_{10} CPM \sim CFD$ + $\log_{10} RPM$	Sequence similarity	<0.001*	0.028
	DNA accessibility	0.031*	
$\log_{10} CPM \sim CFD$ + $\log_{10} RPM$ + CFD $\times \log_{10} RPM$	Sequence similarity	<0.001*	0.028
	DNA accessibility	0.053	
	Sequence similarity $\times$ DNA accessibility	0.526	

205

206



207 **Table S2. Frequency table of GUIDE-seq detected cleavage sites by individual**  
 208 **gRNAs.** \*All 5 cleavage sites with 7 mismatches were not detected by CIRCLE-seq;  
 209 hence they do not affect the subsequent analysis.

Mismatches	0	1	2	3	4	5	6	7	Subtotal	Alias	Cell line
gRNA	Detected cleavage sites										
HEK293 site 1	1	0	1	5	2	1	0	0	10	HEKgRNA1	HEK293T
HEK293 site 2	1	0	1	0	1	0	0	0	3	HEKgRNA2	HEK293T
HEK293 site 3	1	0	0	2	2	0	0	0	5	HEKgRNA3	HEK293T
HEK293 site 4	1	0	9	50	55	13	5	1	134	HEKgRNA4	HEK293T
EMX1	1	0	1	7	5	0	0	0	14	EMX1	U2OS
FANCF	1	0	1	3	3	0	0	0	8	FANCF	U2OS
RNF2	1	0	0	0	0	0	0	0	1	RNF2	U2OS
VEGFA site 1	1	1	2	2	6	2	1	0	15	VEGFA_site1	U2OS
VEGFA site 2	1	0	0	10	49	47	22	3	132	VEGFA_site2	U2OS
VEGFA site 3	1	1	7	26	12	3	1	1	52	VEGFA_site3	U2OS
Subtotal	10	2	22	105	135	66	29	5	374		

210

211 **Table S3. Frequency table of CIRCLE-seq detected cleavage sites by individual**  
 212 **gRNAs.**

Mismatches	0	1	2	3	4	5	6	Subtotal	Alias	Cell line	
gRNA	Detected cleavage sites										
HEK293 site 1	1	0	1	9	17	18	5	51	HEK293_Adli_site1	HEK293T	
HEK293 site 2	1	0	1	13	21	5	1	42	HEK293_Adli_site_2	HEK293T	
HEK293 site 3	1	0	2	10	26	44	26	109	HEK293_Adli_site_3	HEK293T	
HEK293 site 4	1	0	13	100	352	385	160	1011	HEK293_combined_Adli_site_4	HEK293T	
EMX1	1	0	1	11	26	26	4	69	U2OS_exp2_EMX1	U2OS	
FANCF	1	0	1	10	18	16	4	50	U2OS_exp2_FANCF	U2OS	
RNF2	1	0	1	0	4	1	1	8	U2OS_exp2_RNF2	U2OS	
VEGFA site 1	1	1	3	15	59	124	113	316	U2OS_exp2_VEGFA_site_1	U2OS	
VEGFA site 2	1	0	6	46	254	558	816	1681	U2OS_combined_VEGFA_site_2	U2OS	
VEGFA site 3	1	1	15	167	371	205	40	800	U2OS_combined_VEGFA_site_3	U2OS	
Subtotal	10	2	44	381	1148	1382	1170	4137			

213

214 **Table S4. Counts of CRISPR-mediated cleavage sites intersected between GS and**  
 215 **CS datasets.**

Detected cleavage sites	CS only	GS and CS	GS only
gRNA			
HEKgRNA1	41	10	0
HEKgRNA2	40	2	1
HEKgRNA3	104	5	0
HEKgRNA4	882	130	4
EMX1	57	12	2
FANCF	43	7	1
RNF2	7	1	0
VEGFA_site1	302	14	1
VEGFA_site2	1553	128	4
VEGFA_site3	754	46	6

216

217 **Table S5. List of cleavage sites and corresponding characteristics including CPM,**  
218 **RPM, CFD score detected by GUIDE-seq. (Available for download)**

219 **Table S6. List of cleavage sites and corresponding characteristics including CPM,**  
220 **RPM, CFD score detected by CIRCLE-seq. (Available for download)**