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

Supplemental information

Table of contents:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

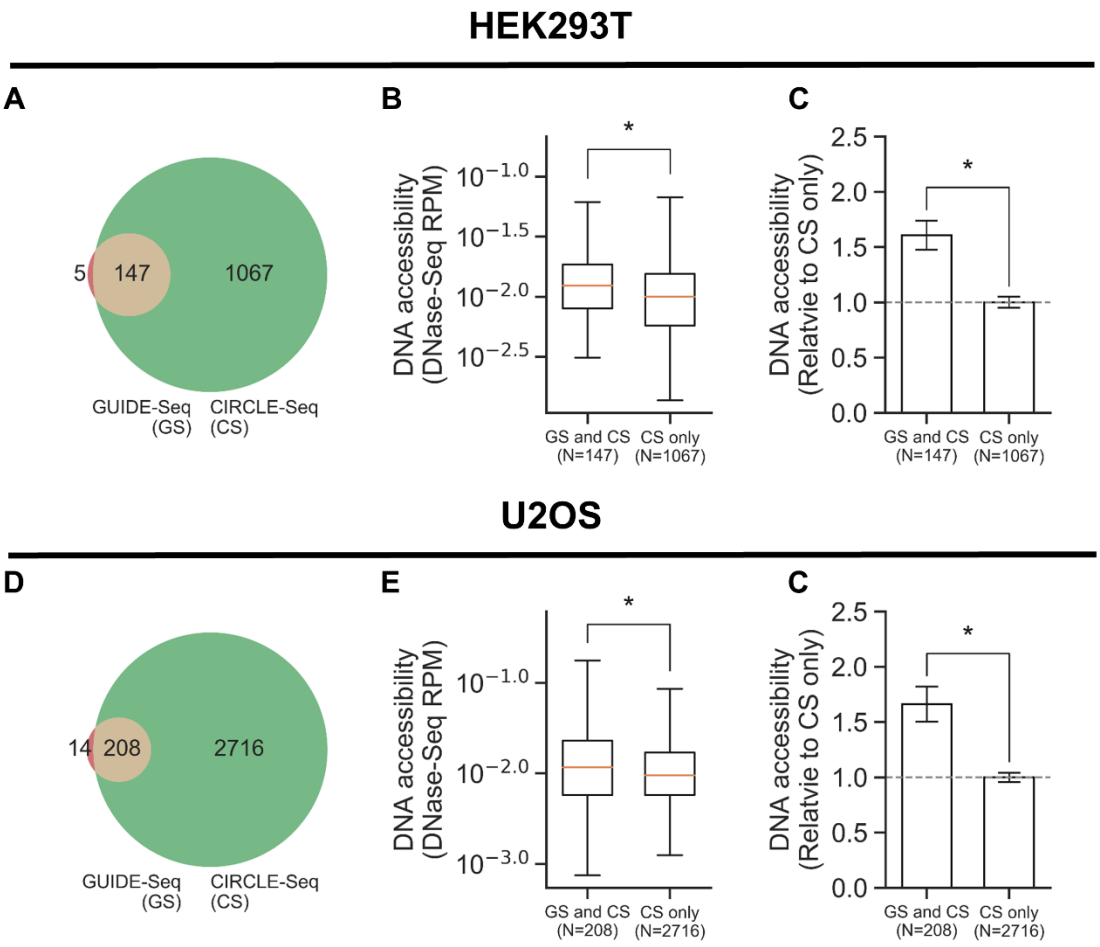

Figure S1. CRISPR-Cas9 targeted more accessible regions in either HEK293T or U2OS cells. (A and D) The Venn diagram displays the number of cleavage sites identified by both GUIDE-seq (GS) and CIRCLE-seq (CS) for the indicated cell type. (B and E) The cleavage sites that both GS and CS identified (GS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

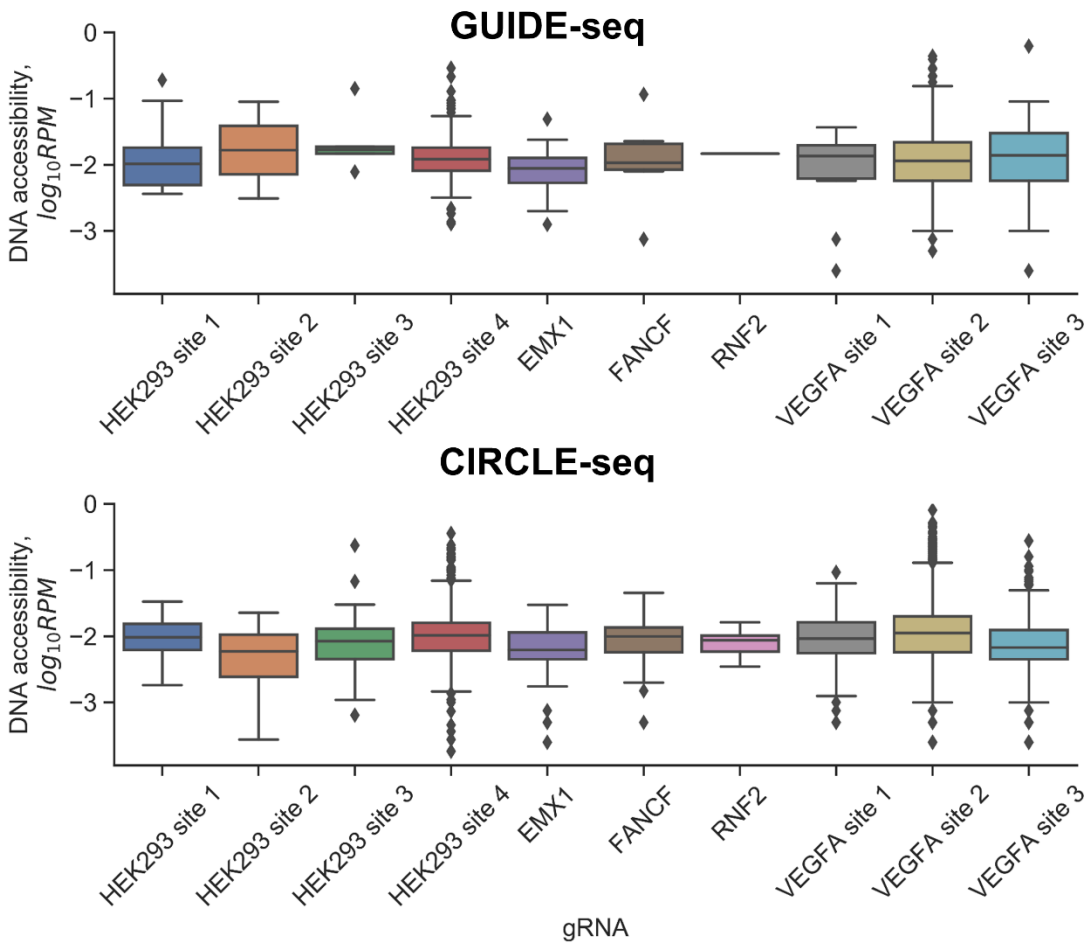

Figure S2. The distributions of DNA accessibility at cleavage sites were similar across individual gRNAs. The box plot shows the distribution of DNA accessibility for individual gRNAs in both assays. The box represents 50% quantile and the line inside the box represents the median.

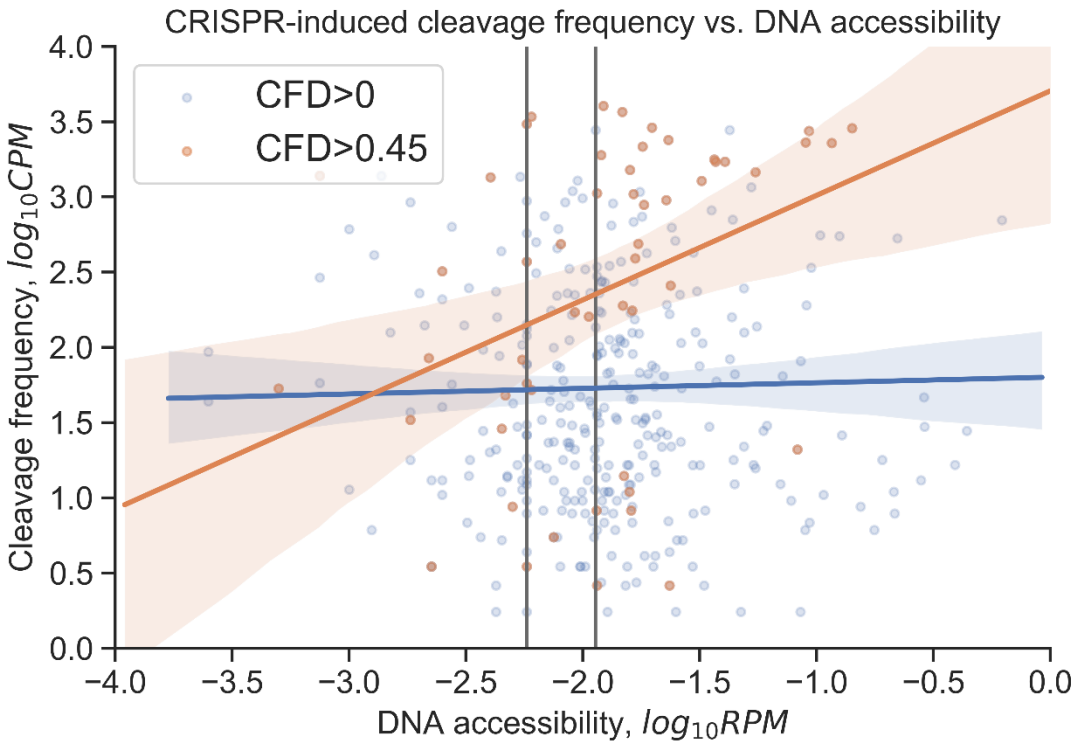

Figure S3. DNA accessibility impacts CRISPR-induced cleavage frequency among cleavage sites with high sequence similarity. Cleavage sites with high sequence similarity were selected as the top 15% of ranked CFD (N=53) in GS and CS subset, which contains cleavage sites with CFD>0.45 (orange). These cleavage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

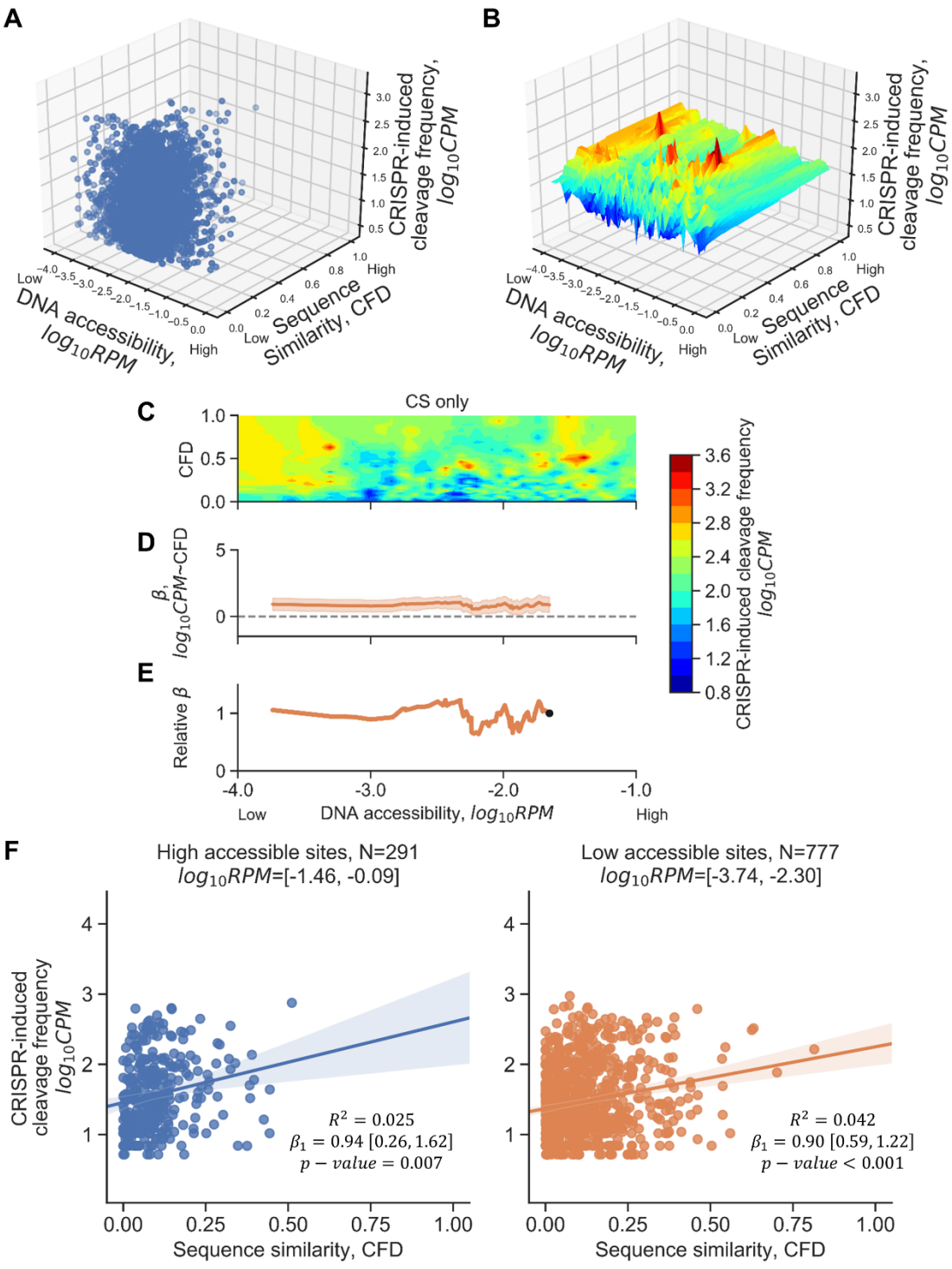

Figure S4. The correlation between gRNA:target sequence similarity and CRISPR-induced cleavage frequency was not affected by DNA accessibility in CS only subset (N=3783). (A) The three-dimensional scatter plot of sequence similarity, DNA

accessibility and CRISPR-induced cleavage frequency using the CRISPR-induced cleavage sites listed in the CS only subset. Each dot represents a CRISPR-induced cleavage site identified by CIRCLE-seq and absent in GUIDE-seq result. CPM represents the number of cleavage events at a CRISPR-induced cleavage site detected by CIRCLE-seq; Sequence similarity represents the likelihood of CRISPR cutting based on the sequence between gRNA and target using CFD matrix; RPM represents the DNA accessibility at a CRISPR-induced cleavage site. (B) The surface plot estimated by the nearest-neighbor method described in the Methods. The sequence similarity is estimated by the position-specific matrix of Cutting Frequency Determination (CFD) score [0,1] that describes the cleavage possibility of gRNA:target pair at the off-target sites. Red color represents high cleavage frequency represented in CPM while blue color represents low cleavage frequency identified by the CIRCLE-seq technique. (C) Contour map of CRISPR-induced cleavage frequency based on the grids of CFD score and DNase-seq RPM; a top-down view of (B). (D) The beta coefficient between CFD and CRISPR-induced cleavage frequency at given 15% quantile of DNA accessibility. Note that the data point was the lower boundary of a given quantile. The shaded regions represent 95% confidence intervals of the t-test. The horizontal dashed line at beta coefficient equal to 0 represents the threshold of the significance of the beta coefficient. The correlation was not significant when the 95% confidence interval covers the horizontal line. (E) The beta coefficient relative to the first quantile that contained the cleavage sites with the top 15% DNA accessibility in the CS only subset. The dashed line represents the regions that were not significant in the Wald Test (D). Note that the CS only subset does not have insignificant quantile therefore no dashed line was indicated. (F) Correlation between CRISPR-induced cleavage frequency and CFD score of 15% most accessible sites (left panel) or 15% least accessible sites (right panel) in the CS only subset. β_1: beta coefficient of simple linear regression. p-value of Wald Test for a hypothesis test that the slope is 0.

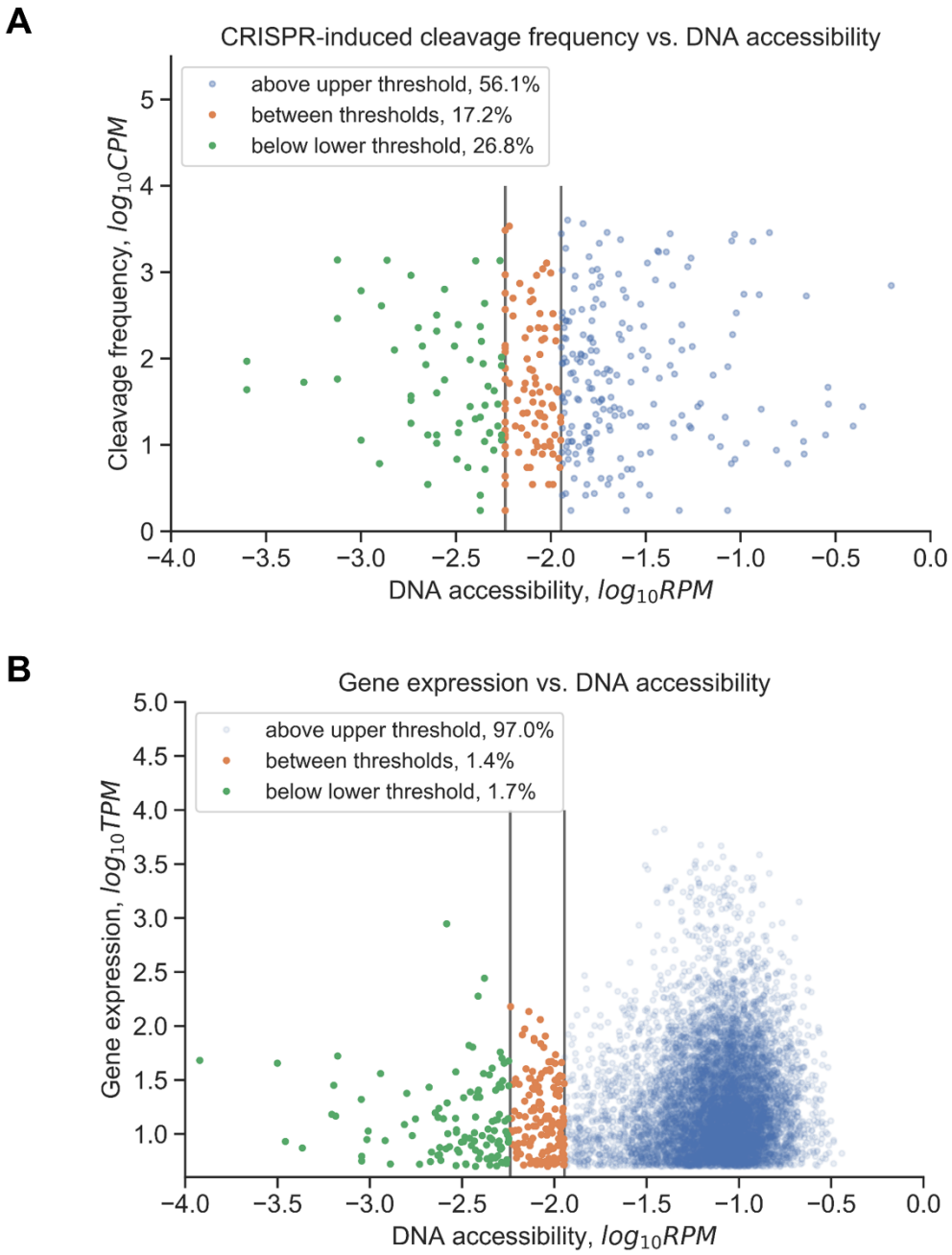

Figure S5. Higher proportion of CRISPR-induced cleavage sites were located at regions with low DNA accessibility as compared to that of endogenous gene loci. (A) Scatter plot of CRISPR-induced cleavage frequency measured by GUIDE-seq and DNA accessibility measured by DNase-seq in both HEK293T and U2OS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

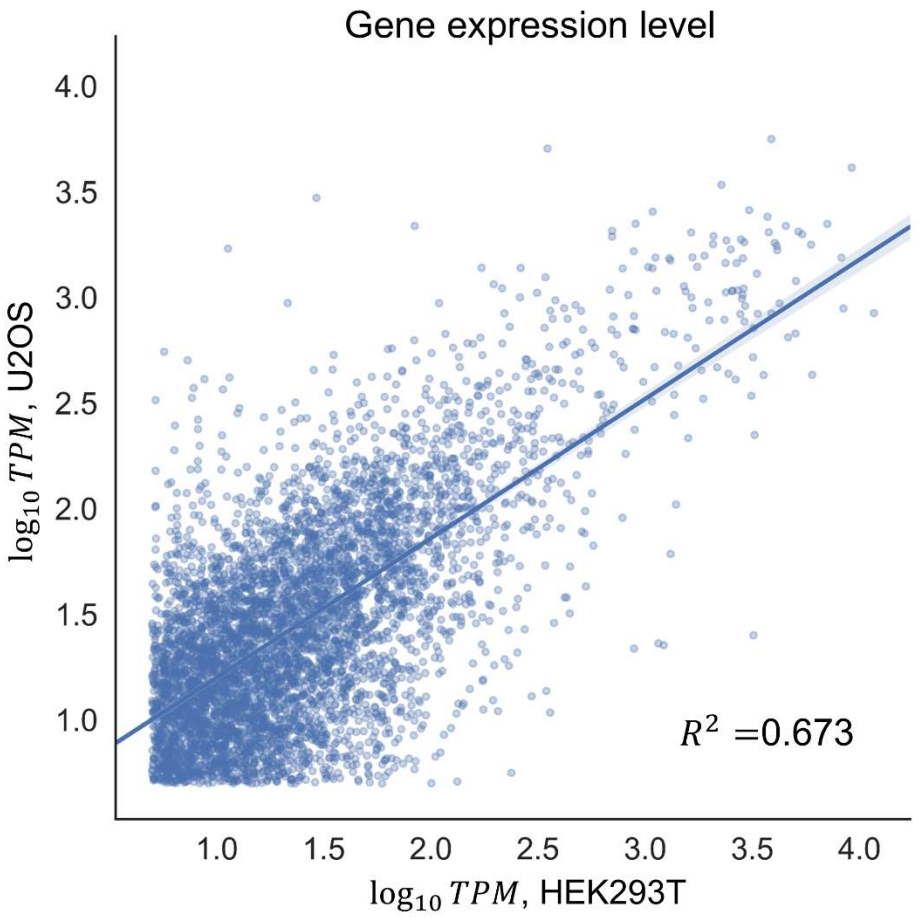

Figure S6. The gene expression profiles were positively correlated between untreated HEK293T and U2OS cells (N=8619). Transcripts with predicted expression level above 5 TPM in both cells were included in this analysis. The R-square was estimated by Pearson correlation coefficient test, p-value<0.001.

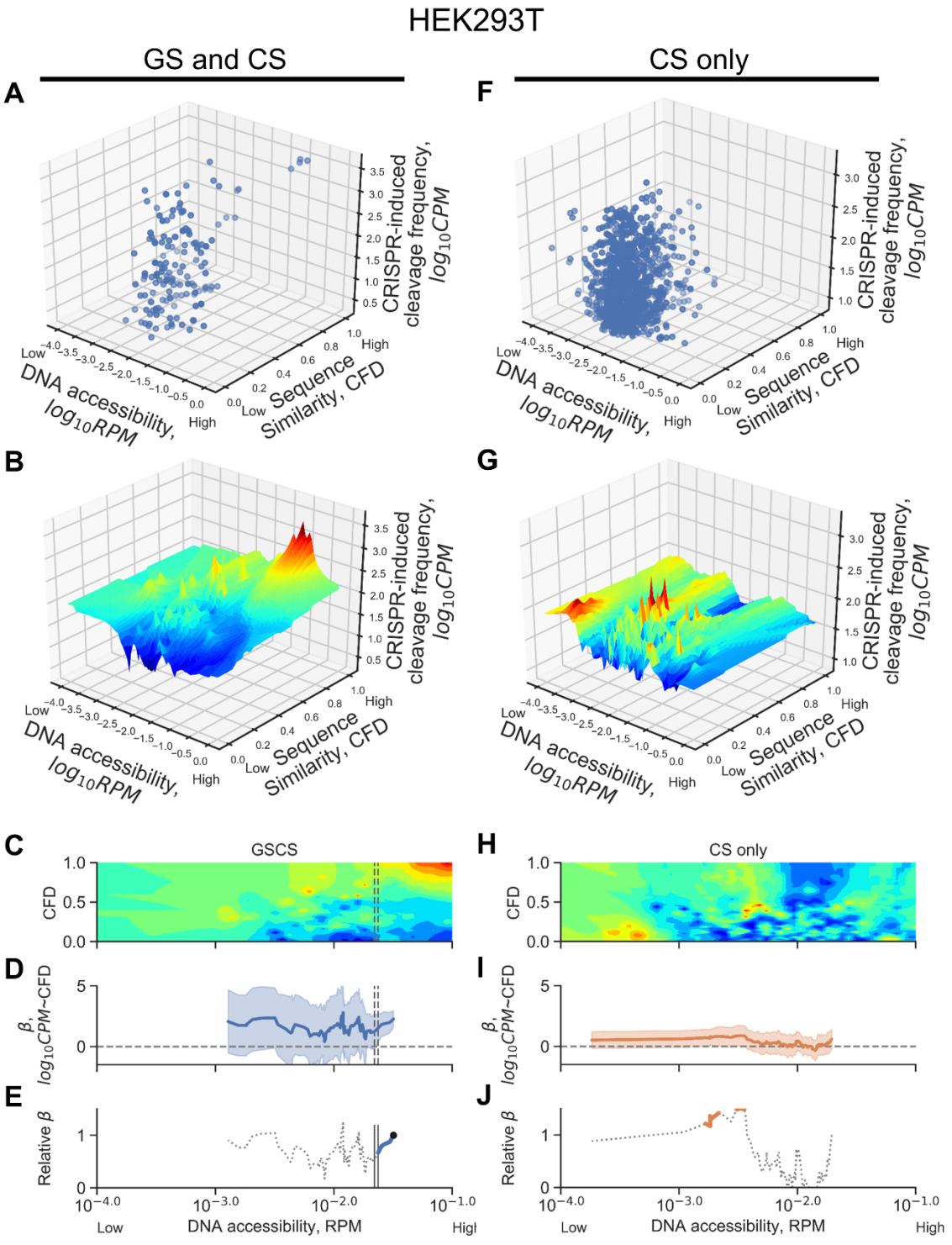

Figure S7. The DNA accessibility abrogated the correlation between gRNA:target sequence similarity and CRISPR-induced cleavage frequency when the chromosomal regions are less accessible in GS and CS subset but not CS only subset in HEK293T cells. The three-dimensional scatter plot of sequence similarity,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

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

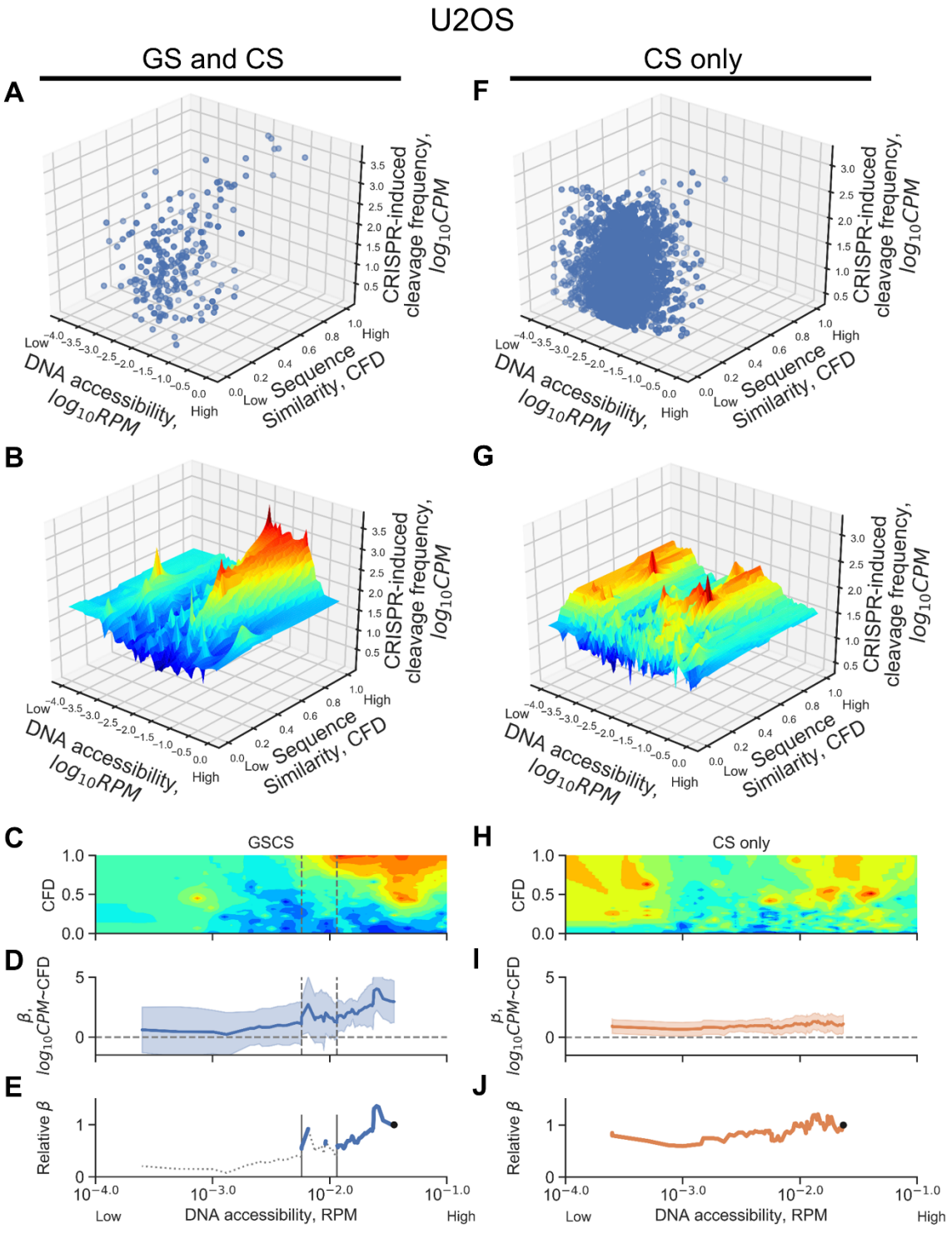

Figure S8. The DNA accessibility abrogated the correlation between gRNA:target sequence similarity and CRISPR-induced cleavage frequency when the chromosomal regions are less accessible in GS and CS subset but not CS only subset in U2OS cells. The three-dimensional scatter plot of sequence similarity,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

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

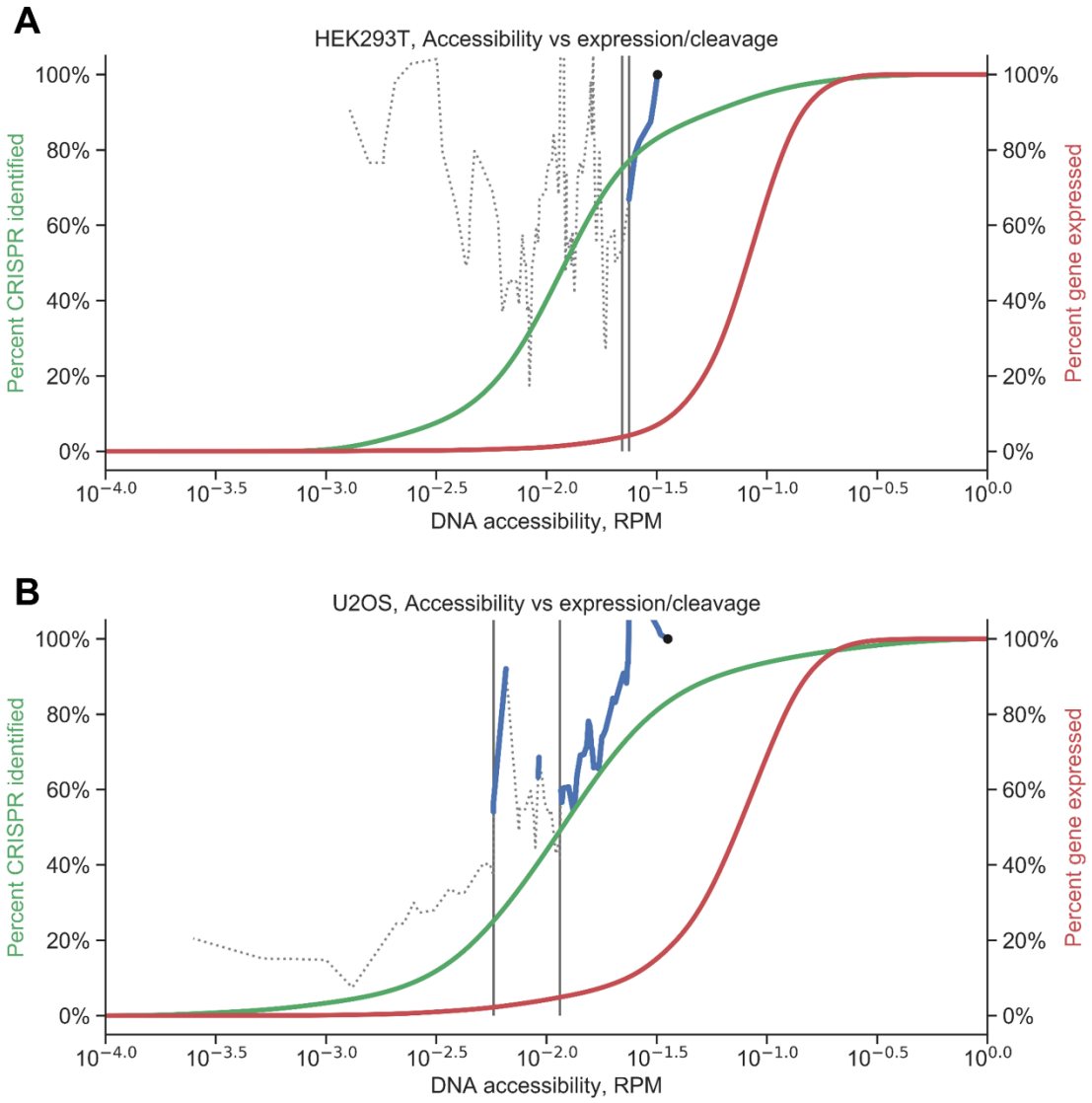

Figure S9. Chromatin accessibility required for CRISPR-mediated cleavage reaction was significantly less than that for endogenous gene to express. (A) Analysis using cleavage sites only identified in HEK293T cells in GS assay and HEK293T RNA-seq. Green curve represents the cumulative percentage of CRISPR-induced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

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

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

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*	
$\log_{10} CPM \sim \mathrm{CFD} + \log_{10} RPM + \mathrm{CFD} \times \log_{10} RPM$	Sequence similarity	<0.001*	0.028
	DNA accessibility	0.053	
	Sequence similarity × DNA accessibility	0.526	

Table S2. Frequency table of GUIDE-seq detected cleavage sites by individual gRNAs. *All 5 cleavage sites with 7 mismatches were not detected by CIRCLE-seq; 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		

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

Table S4. Counts of CRISPR-mediated cleavage sites intersected between GS and 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

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

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