FEBS Letters

FEBS PRESS science publishing by scientists

Chromatin accessibility and guide sequence secondary structure affect CRISPR-Cas9 gene editing efficiency

Kristopher Torp Jensen[1,2], Lasse Fløe[1], Trine Skov Petersen[1], Jinrong Huang[3,4], Fengping Xu[3,4,5], Lars Bolund[1,3,4], Yonglun Luo[1] and Lin Lin[1]

1 Department of Biomedicine, Aarhus University, Denmark
2 University of Cambridge, UK
3 BGI-Shenzhen, China
4 China National GeneBank-Shenzhen, BGI-Shenzhen, China
5 Biology Department, Copenhagen University, Denmark

Correspondence
Y. Luo and L. Lin, Department of Biomedicine, Aarhus University, Wilhelm Meyer Alle 4, DK-8000, Aarhus C, Denmark
Fax: +45 86123173
Tel: +45 87167761
E-mails: alun@biomed.au.dk; lin.lin@biomed.au.dk

(Received 26 February 2017, revised 29 May 2017, accepted 30 May 2017, available online 28 June 2017)

doi:10.1002/1873-3468.12707

Edited by Ned Mantei

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-associated protein 9 (CRISPR-Cas9) systems have emerged as the method of choice for genome editing, but large variations in on-target efficiencies continue to limit their applicability. Here, we investigate the effect of chromatin accessibility on Cas9-mediated gene editing efficiency for 20 gRNAs targeting 10 genomic loci in HEK293T cells using both SpCas9 and the eSpCas9(1.1) variant. Our study indicates that gene editing is more efficient in euchromatin than in heterochromatin, and we validate this finding in HeLa cells and in human fibroblasts. Furthermore, we investigate the gRNA sequence determinants of CRISPR-Cas9 activity using a surrogate reporter system and find that the efficiency of Cas9-mediated gene editing is dependent on guide sequence secondary structure formation. This knowledge can aid in the further improvement of tools for gRNA design.

Keywords: chromatin accessibility; CRISPR; efficiency; gRNA

Over the past 4 years, Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-associated protein 9 (CRISPR-Cas9) has emerged as the preferred genome editing tool in a wide variety of organisms, with the *Streptococcus pyogenes* Cas9 nuclease (SpCas9) being the preferred choice in most situations [1–3]. However, the applicability of SpCas9 and Cas9 proteins in general is still limited by large variations in gene editing efficiencies at different genomic loci, which renders the system ineffective toward some potential targets [4]. More importantly, despite the development of predictive tools with respect to gRNA efficiency [5–7], there is still a lack of knowledge regarding the underlying reasons for the variation in Cas9 efficiency. This leads to limited predictability of whether the CRISPR-Cas9 system will be able to effectively target a given region of interest.

A further challenge when using Cas9 in large metazoan genomes is a lack of specificity in the presence of DNA sequences partially complementary to the guide sequence [8]. This has been addressed by several approaches aimed at decreasing the Cas9-DNA-gRNA binding energy [9–11]. Two notable rationally engineered Cas9 variants designed to improve specificity are the Cas9-HF1 [10] and the eSpCas9(1.1) [9] (hereafter eSpCas9). They were designed by introducing selective point mutations in sequences encoding SpCas9 DNA-interacting regions to minimize nonspecific Cas9–DNA interactions. This decreases the Cas9-DNA-RNA complex formation energy with the

Abbreviations
CRISPR-Cas9, Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-associated protein 9; DMEM, Dulbecco's modified Eagle's medium; DSBs, DNA double-strand breaks; RPKM, reads-per-kilobase million; SpCas9, *Streptococcus pyogenes* Cas9 nuclease.

resulting Cas9 proteins being dependent on complete or near-complete complementarity between gRNA and target site for strand separation and cleavage to occur. Determinants of gRNA activity for these rationally engineered proteins and how they compare to native SpCas9 have not been thoroughly investigated.

dCas9 binding [12] and SpCas9 nuclease activity [13–15] have previously been shown to correlate with an absence of repressive histone marks and increased DNase I sensitivity in human cancer cells. This was shown by Chen *et al.* [14].and Daer *et al.* [13] using inducible model systems, and by Chari *et al.* [15] using a library-on-library approach. Additionally, previous reports show a dependence of Cas9 endonuclease activity on the ability of the gRNA to form internal secondary structures [16,17].

In the present study, we further investigate the effects of chromatin structure and gRNA properties on both SpCas9 and eSpCas9 activity in human cancer cells and fibroblasts. We do this by a comparative analysis of the activities of 20 gRNAs when targeting either 10 endogenous regions, as quantified by TIDE assay [18], or the corresponding target sequences incorporated into the plasmid-based C-Check reporter system [19]. This is done to distinguish between chromatin-specific factors and guide-specific factors affecting the endogenous cleavage rate. Additionally, we use a nonlibrary-based approach in order to avoid competition between gRNAs and relative RNA stabilities affecting our results.

We confirm that SpCas9 is more efficient at inducing DNA double-strand breaks (DSBs) in euchromatin than in heterochromatin in HEK293T cells, HeLa cells, and human fibroblasts. We further show that the effect of chromatin accessibility is similar for both SpCas9 and eSpCas9 in the human cancer cell lines. We also show that strong internal secondary structure formation of the gRNA guide sequence is detrimental to SpCas9-mediated DNA cleavage in the chromatin-independent C-Check system. These low activities can be rescued by selective point mutations that destabilize strong guide sequence secondary structures.

Materials and methods

CRISPR gRNA design

Clustered Regularly Interspaced Short Palindromic Repeats gRNAs were designed using the online prediction tool DNA2.0 (https://www.dna20.com/eCommerce/cas9/input) for the selected open and closed chromatin regions. The remaining target sites were selected from the articles describing eSpCas9 and SpCas9-HF1 [9,10]. All gRNA target sites and the synthetic gRNA-DNA oligonucleotides are listed in Table S1.

DNA oligonucleotide synthesis and Sanger sequencing

All DNA oligonucleotides in this study were synthesized by Sigma Aldrich (Brøndby, Denmark). Sanger sequencing was conducted using the Mix2seq kit from Eurofins Genomics (Ebersberg, Germany).

Generation of gRNA expression vectors, and C-Check vectors

The SpCas9 (Addgene plasmid # 48139) and eSpCas9(1.1) (Addgene plasmid # 71814) plasmids were a gift from Feng Zhang. To generate the gRNA expression vectors, complementary gRNA oligos were annealed in 1X NEB buffer 2, cloned into the lentiGuide-Puro plasmid (a gift from Feng Zhang, Addgene plasmid # 52963), and subsequently validated by Sanger sequencing. To generate the C-Check reporter vectors, each genomic region was amplified by PCR from genomic DNA isolated from HEK293T cells and cloned into the C-Check vector (Addgene plasmid #66817) by Golden Gate assembly. All C-Check vectors were validated by Sanger sequencing. PCR conditions and PCR primers are listed in Table S2.

Cell culture

HEK293T cells (ATCC), HeLa cells (ATCC), and human fibroblasts (MJ2646, a gift from Dr. Bin Liu, KB, Denmark) were cultured in Dulbecco's modified Eagle's medium (DMEM) (LONZA) supplied with 10% FBS (Gibco, Taastrup, Denmark), 1% penicillin–streptomycin (Sigma), and 1% GlutaMAX (Gibco) in a 37 °C incubator with a 5% CO_2 atmosphere and maximum humidity. At approximately 80% confluence the cells were detached by 0.05% Trypsin-EDTA and passaged at a ratio of 1 : 8.

Transfection

Transfections were performed using the X-tremeGENE9 DNA transfection reagent (Roche). Cells were seeded into 24-well plates 1 day before transfection with a cell density of 5×10^4 cells per well for HEK293T cells and fibroblasts and 1×10^4 cells per well for HeLa cells. A pUC19 plasmid was used for transfection of control groups in the same amount as for the combined plasmids of noncontrol groups.

C-Check analysis

Forty-eight hours after transfection according to the X-tremeGENE9 DNA transfection reagent protocol, cells were

harvested using 0.025% trypsin-PBS-EDTA (phenol red free) and washed twice with PBS-5%FBS, fixed with 4% Formaldehyde-PBS solution for 10 min at room temperature, and washed twice with PBS before analysis. The percentage of GFP+ cells among the AsRED+ cells was quantified by flow cytometry (LSRFortessa analyser, FACS CORE facility, Department of Biomedicine, Aarhus University). For GFP, a 488 nm laser and 530/30 bandpass filter was applied. For AsRED, a 561 nm laser and 586/15 bandpass filter was applied. At least 20 000 events were recorded for each sample. All experiments were performed in triplicates or more, and data were analyzed with Flowjo 10.

TIDE assay

Two days after transfection, HEK293T and HeLa cells were harvested in 50 μL of a complete cell lysis solution (50 mM KCl, 1.5 mM MgCl2, 10 mM Tris–Cl, pH 8.5, 0.5% Nonidet P40, 0.5% Tween 20, 400 $\mu g \cdot mL^{-1}$ proteinase K), followed by heating at 65 °C for 30 min and 95 °C for 10 min. Fibroblasts were treated with puromycin (1 $\mu g \cdot mL^{-1}$) at day 1 after transfection followed by removal of puromycin with two PBS washes and change of medium at day 3 after transfection. They were cultured for an additional 3 days followed by three washes with PBS and then harvested as described above. 1 μL (HEK293T & HeLa) or 2 μL (fibroblasts) of lysate was used as template for PCR with the high fidelity Platinum® Pfx DNA Polymerase. PCR products were purified using NucleoSpin Gel and PCR Clean Up kit and directly sequenced using Sanger sequencing (Mix2seq, Eurofins Genomics). Easy quantitative assessments of genome editing (TIDE) were performed to determine the cleavage efficiency [18]. A P value cutoff of $P < 0.001$ and indel size range from −10 to +10 was used for all analyses, and decomposition windows were optimized for each group according to TIDE guidelines. For analysis of TIDE data, control values from groups transfected with a plasmid expressing an empty gRNA backbone were subtracted, as these indicate the nonspecific error from the TIDE calculation at each site. These were generally low with most control values being 0.1–2% and the highest being 3.55%.

DNase I-seq, CTCF ChIP-seq, H3K4me3, RNA-seq, and DNA methylation analysis for selected genomic loci

For DNase I-seq (HEK293T, Hela, human fibroblasts), CTCF ChIP-seq (HEK293T) and H3K4me3 (HEK293T) data, bam files were downloaded from the ENCODE project (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/).

For RNA sequencing in HEK293T cells (GSM1937997), sequencing was conducted on the Ion Proton platform and clean reads were aligned to the human reference genome (hg19) by tmap.

Next, the total reads and RPKM/FPKM of the investigated regions (Table S3) were calculated by read count (version 0.01).

For DNA methylation, clean reads from whole-genome bisulfide sequencing in HEK293T cells were aligned to the human reference genome (hg19) by BSMAP(v2.74) with the parameter '-u -v 5 -z 33 -p 6 -n 0 -w 20 -s 16 -r 0 -f 10 -L 140'. The CpG methylation levels of the investigated regions were then calculated.

Open chromatin regions were selected from regions with a significantly enriched DNase-seq signal (P value = $10e^{-16}$). Closed chromatin regions had no enriched Dnase-seq signal and little to no RNA-seq read coverage. Regions that met these criteria in both HEK293T and HeLa cells were used. The chromatin accessibility of regions repeated from previous studies was quantified for 1 kb of DNA encompassing the target sites.

Analysis of Gibbs free energy for gRNA guide sequences and melting temperature of target sites

The Gibbs free energies for formation of gRNA secondary structures were determined using the MFOLD RNA (3.0) software (http://unafold.rna.albany.edu/?q=mfold/rna-folding-form). DNA target site melting temperatures were determined using the MATLAB oligoprop feature based on Sugimoto *et al.* [20] nearest neighbor calculations.

Statistical analysis

All data are represented as mean ± standard deviation. Unless stated elsewhere, statistical analyses were carried out using Student's t-test for comparisons between pairs of groups and one-way analysis of variance (ANOVA) with Bonferroni correction for multiple comparisons. All statistical analyses were conducted using STATA (version 10). P values less than 0.05 were considered statistically significant.

Results

Local chromatin accessibility affects Cas9 efficiency

To perform initial screens of the activities of SpCas9 [21] and eSpCas9 [9], we used the plasmid-based dual-fluorescent reporter system C-Check [19]. The C-Check plasmid constitutively expresses AsRED, and two CRISPR-Cas9 target sites are incorporated into the

plasmid between two sections of GFP, both including a 500 nucleotide region of homology. Single-strand annealing-mediated repair of a potential DSB in the target region reconstitutes GFP expression and green fluorescence. As expected, these preliminary investigations suggested that eSpCas9 exhibits capacity similar to that of SpCas9 for inducing DNA double-strand breaks while having reduced mismatch tolerance (Fig. S1).

We then proceeded to investigate whether the local chromatin environment affects Cas9 efficiency when targeting endogenous loci. Ten endogenous regions were selected based on their chromatin accessibility in HEK293T cells according to hierarchical clustering (Fig. 1A) using data on DNaseI-sensitivity (DNase-seq), H3K4me3, and CFTC binding frequency from the ENCODE project, as well as gene expression levels (RNase-seq, GSM1937997) and CpG methylation (GSM2425586) from an ongoing study. For this analysis, between 1 and 5 kilobases of DNA encompassing the target sites was considered for each region (Table S3). DNase I-seq reads-per-kilobase million (RPKM) values were found to correlate well with the other parameters investigated and are used as an indicator of chromatin accessibility in the following (Fig. 1A). We next designed 20 gRNAs targeting these 10 different genomic regions (two gRNAs per region, Fig. S2). The endogenous efficiencies of the 20 gRNAs for SpCas9 and eSpCas9 were assessed using Sanger sequencing and TIDE assay in HEK293T cells (Materials and methods). Control values from cells transfected with the relevant Cas9 and a plasmid expressing the empty gRNA backbone were subtracted to yield indel frequencies for analysis. A significant variation in indel mutation rates was observed between the 20 gRNAs (Fig. 1B). eSpCas9 generally had an efficiency comparable to that of SpCas9 for most targets, but as observed in the initial publication on eSpCas9 it had a much lower cleavage efficiency than SpCas9 for certain target sites (Fig. 1B, Chr22.1_C2 and Chr18_C1 in particular; Chr9_O1 and Chr9_O2 to a lesser extent). However, eSpCas9 also had a higher efficiency than SpCas9 for some targets and generally appears to be a viable high-specificity alternative to SpCas9. Furthermore, we observed large differences in activity between gRNAs targeting the same genomic region (Fig. 1B). This illustrates that in addition to chromatin-associated effects, gRNA sequence also affects efficiency.

Based on the DNase-seq data, we proceeded to investigate whether indel frequency was correlated with chromatin accessibility. There was a notable correlation between DNase-Seq RPKM values for the targeted regions and cleavage efficiency in genomic DNA using the 20 gRNAs (Fig. 1C, $R^2 = 0.44$ and $R^2 = 0.55$ for a linear RPKM-dependence for SpCas9 and eSpCas9, respectively). A comparison of the groups of gRNAs targeting open and closed regions also showed that there is a significant difference in activity between these two groups (Fig. 1D, $P < 0.01$ for SpCas9, $P < 0.001$ for eSpCas9, ANOVA). This suggests that Cas9 efficiency is affected by local chromatin accessibility, with Cas9 activity generally being higher in DNase I hypersensitive regions.

To confirm that the chromatin dependence of Cas9-mediated DSBs is not specific to HEK293T cells, the data were further validated by targeting two DNase I-sensitive and two DNase I-insensitive chromatin regions (8 gRNAs in total) in HeLa cells and human fibroblasts. In HeLa cells, a correlation between DNase I RPKM values and indel frequencies similar to that for HEK293T cells was observed (Fig. 1E, $R^2 = 0.72$ for SpCas9, $R^2 = 0.83$ for eSpCas9, and Fig. S3A,B). In human fibroblasts, nontransfected cells were eliminated by transient puromycin selection (see Materials and methods). The SpCas9 efficiency similarly appeared to depend strongly on chromatin accessibility (Fig. 1F, $R^2 = 0.85$, and Fig. S3C,D). The efficiency of eSpCas9 across the investigated targets was too low compared to control values in fibroblasts to say anything conclusive about its chromatin dependence (data not shown). Taken together, our results suggest a positive correlation between chromatin accessibility and CRISPR-Cas9 gene editing efficiency.

The C-Check system facilitates investigations of gRNA-specific effects on Cas9 efficiency

We then proceeded to verify that the observed trends in guide efficiency are in fact due to differences in chromatin structure rather than differences in the inherent activities of the gRNAs. We assayed the activity of the 20 investigated gRNAs, and an additional 7 gRNAs, using a plasmid-based dual-fluorescent reporter system (C-Check). Our C-Check analyses showed that eSpCas9 and SpCas9 exhibited similar activity for most targets (Fig. 2A). However, consistent with the observations when endogenous loci were targeted, there are notable differences in efficiency between SpCas9 and eSpCas9 for some target sites. When gRNA activity in the C-Check system was compared to DNase I RPKM values, no correlation was found (Fig. S4A). Similarly, comparing gRNAs targeting open and closed regions showed that there is no significant difference in the activities between the two groups in the C-Check system (Fig. 2B, $P = 0.65$ for SpCas9, $P = 0.54$ for eSpCas9). This further supports

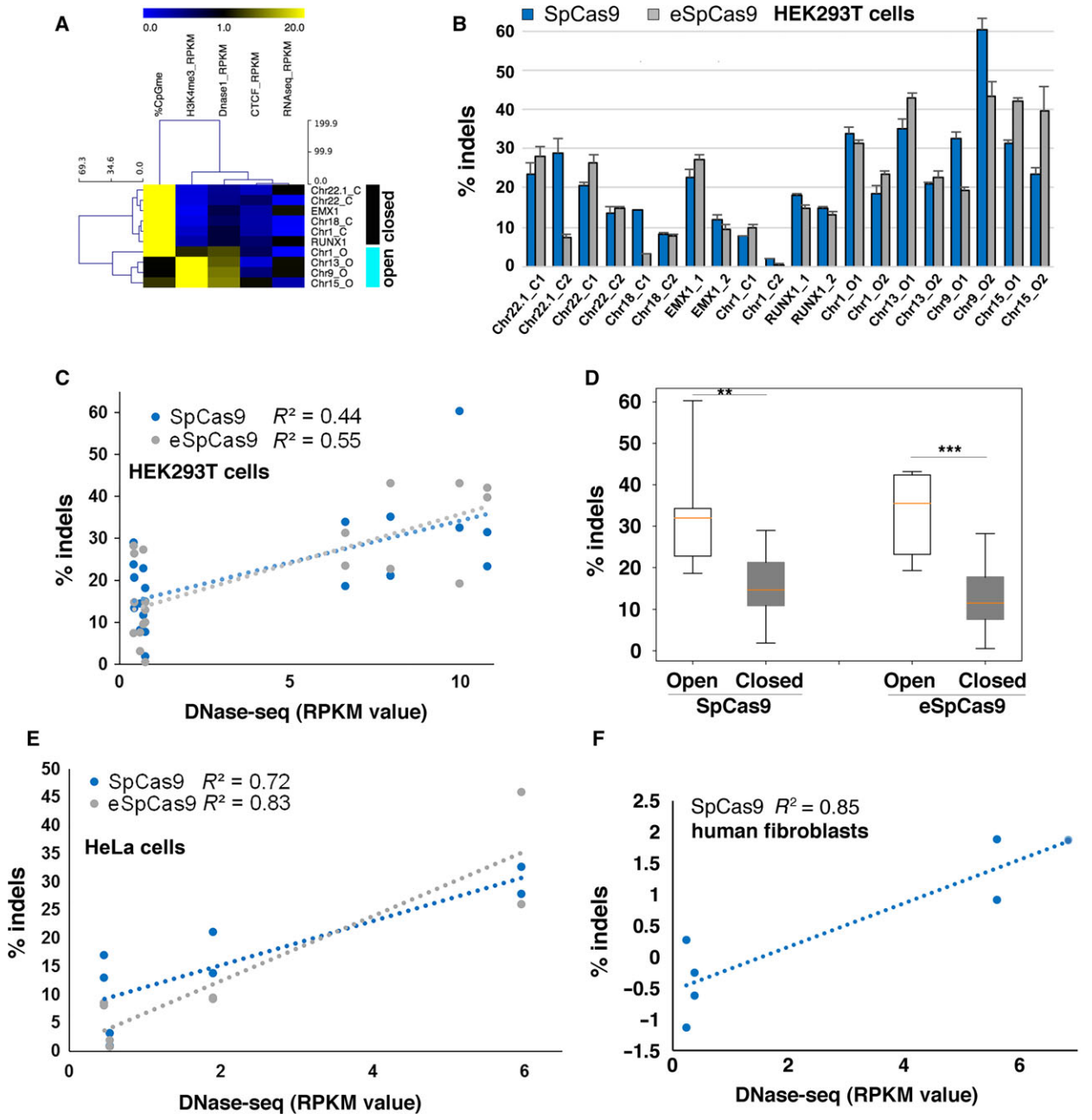

Fig. 1. Effect of chromatin accessibility on SpCas9 and eSpCas9 efficiencies. (A) Hierarchical clustering (average linkage with Euclidean Distance) of the regions targeted in the experiments. Regions were clustered according to data from DNase I-seq (ENCODE), RNase-seq (GSM1937997), H3K4me3 ChIP-seq (ENCODE), CTCF ChIP-seq (ENCODE), and CpG methylation by whole genome bisulfite sequencing (GSM2425586) in HEK293T cells. (B) Efficiency of SpCas9 and eSpCas9 for 20 gRNAs targeting 10 genomic loci in HEK293T cells. Indels were quantified at each genomic locus by TIDE and ordered according to chromatin DNase-Seq RPKM for the region. Annotation indicates closed chromatin (_C#), open chromatin (_O#) or reference (_#) target. Values represent mean and one SD ($n = 2$–3). (C) Dot plot of % indels against HEK293T DNase I-seq RPKM values for each target. Regression line represents linear line of best fit. (D) Box plot of % indels for gRNAs grouped according to target regions in open or closed chromatin. Asterisk represents $P < 0.01$ (**) or < 0.001 (***) (ANOVA). (E) Dot plot of % indels against DNase-seq RPKM values for two open and two closed chromatin regions in HeLa cells. Two gRNAs were used to target each of the Chr9_O, Chr15_O, Chr1_C, and Chr22_C regions. Regression line represents linear line of best fit. (F) Dot plot of % indels against DNase I-seq RPKM values for two open and two closed chromatin regions in human fibroblasts. Two gRNAs were used to target each of the Chr9_O, Chr15_O, Chr1_C, and Chr22_C regions. Regression line represents linear line of best fit.

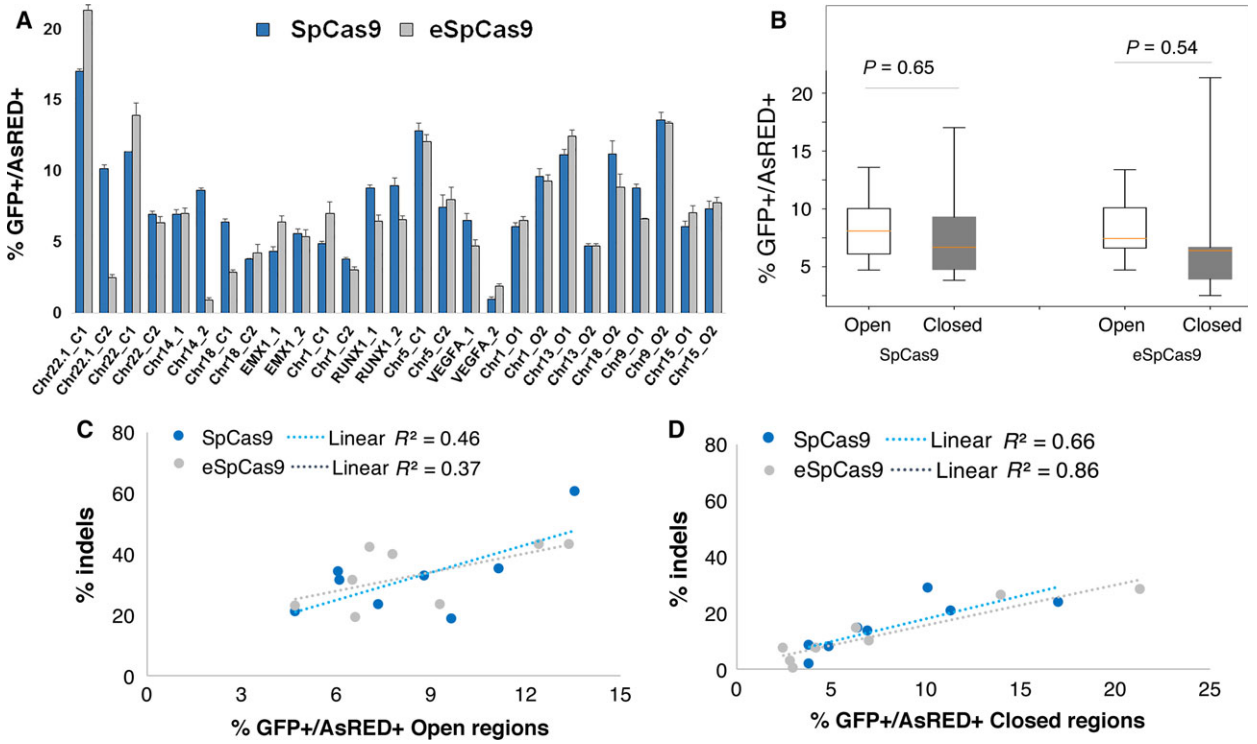

Fig. 2. Comparison of Cas9 activity in endogenous and plasmid-based assays. (A) Efficiency of SpCas9 and eSpCas9 in the C-Check system for 27 gRNAs. Sorted from lowest to highest RPKM, values represent mean and one SD ($n = 3$). (B) Box plot of % GFP+/AsRED+ cells, grouped by gRNAs targeting open or closed chromatin regions. *P* values from ANOVA analysis. (C, D) Dot plot of % indels against C-Check efficiency for each target for SpCas9 and eSpCas9. Data have been separated into regions from open chromatin (C) and regions from closed chromatin (D). Regression lines represent linear lines of best fit.

the observed chromatin dependence when targeting endogenous loci, since the endogenous pattern in gRNA activities is not replicated in the plasmid-based system. The DNase I-seq data further allowed us to compare Cas9 efficiencies in the C-Check system with the TIDE assay results for DNase I-sensitive and -insensitive chromatin regions separately. This revealed that while there is only a rough correlation [$R^2 = 0.32$ (SpCas9) and 0.34 (eSpCas9)] between C-Check and TIDE efficiencies for the combined data (Fig. S4B), there is a higher linear correlation when only comparing targets in regions of similar DNase I sensitivity (Fig. 2D,E, $R^2 = 0.46$ (SpCas9) and 0.37 (eSpCas9) for open regions, $R^2 = 0.66$ (SpCas9) and 0.86 (eSpCas9) for closed regions). This indicates that while such plasmid-based reporter systems may be useful tools for predicting endogenous Cas9 efficiencies, this is only the case when comparing targets in similar chromatin environments. After accounting for DNase I sensitivity, the correlation between C-Check and TIDE data is still not as strong as might have been expected. This could be because the C-Check system represents a nonphysiological environment with multiple targets per cell and only gives an indirect readout of DSB frequency via protein expression. However, it could also be due to the number of confounding factors influencing Cas9-mediated cleavage frequency for endogenous targets. While we have sorted target regions according to DNase I sensitivity and this correlates well with a number of other chromatin parameters (Fig. 1A), these additional factors introduce more variability in the Cas9 efficiency than can be accounted for merely by gRNA sequence and DNase I sensitivity.

gRNA guide sequence secondary structure formation influences Cas9 efficiency

Since the C-Check results indicate how the efficiency of the two Cas9 proteins vary with gRNA sequence independently of chromatin, the assay was used to identify gRNA sequence-specific factors which influence Cas9-mediated cleavage. Among the features analyzed, our data indicate that secondary structure formation of the guide sequence alone (Fig. 3A,B, $P < 0.05$ for SpCas9, One-way ANOVA), but not the complete gRNA including scaffold (Fig. S5A),

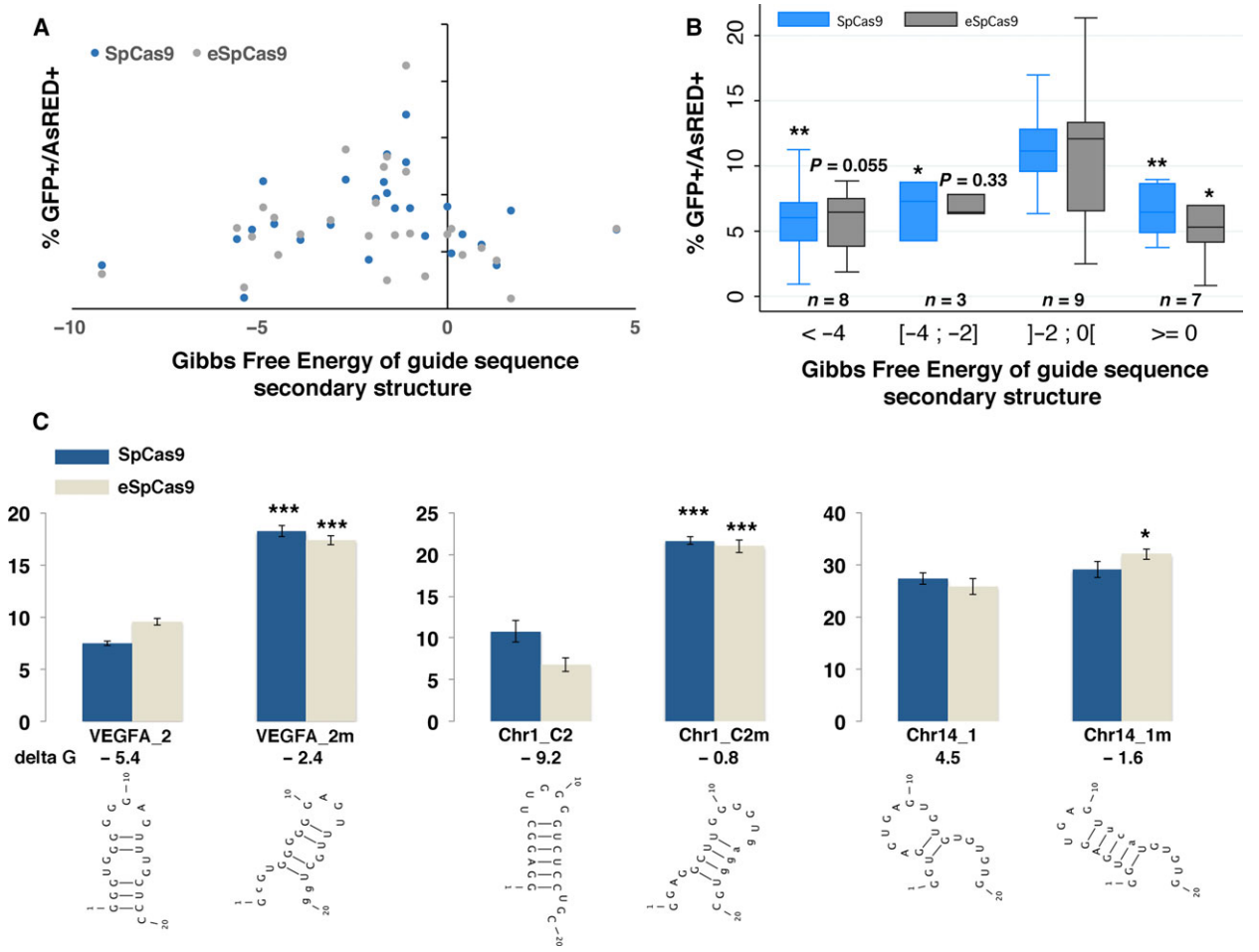

Fig. 3. Effect of gRNA guide sequence secondary structure on SpCas9 and eSpCas9 efficiencies. (A) Dot plot of SpCas9 and eSpCas9 efficiency quantified by C-Check against the Gibbs free energy change ($\Delta_r G$, kJ·mol^{-1}) for formation of the most stable guide sequence secondary structure for each gRNA. (B) Box plot showing the efficiency of SpCas9 and eSpCas9 grouped according to the Gibbs free energy change ($\Delta_r G$, kJ·mol^{-1}) for formation of the most stable guide sequence secondary structure. Asterisk represents $P < 0.05$ (*) or < 0.01 (**) compared to $\Delta_r G \in [-2; 0]$. (C) Predicted most stable secondary structure and corresponding $\Delta_r G$ for three pairs of original and modified gRNAs together with their efficiencies in the C-Check system. Asterisk represents $P < 0.05$ (*) or < 0.001 (***) compared to original gRNAs.

significantly affects the cleavage efficiency for both SpCas9 and eSpCas9. More specifically, cleavage efficiencies are highest when the Gibbs free energy ($\Delta_r G$) for formation of the most stable gRNA guide secondary structure is between -2 and 0 kJ·mol^{-1}. Our data indicate that weak secondary structure formation might in fact be beneficial to cleavage ($P < 0.05$ compared to no secondary structure) while formation of very stable secondary structures is detrimental ($P < 0.05$ compared to $\Delta_r G \in [-4; -2]$, $P < 0.01$ compared to $\Delta_r G < -4$ kJ·mol^{-1}). We propose a model where the scaffold of the gRNA binds strongly to the Cas9 protein while the guide sequence has more conformational freedom. This allows the guide sequence to not only invade the target DNA strand but also to form secondary RNA structures relatively independently of the scaffold, thus explaining why secondary structure stability of the guide sequence alone affects Cas9 activity. We further validated our observation that stable guide sequence secondary structure formation is detrimental to Cas9 activity by making selective point mutations in two of our previously investigated gRNAs (VEGFA_2, Chr1_C2) which exhibited strong secondary structures (-5.4 kJ·mol^{-1} and -9.2 kJ·mol^{-1}) and correspondingly low efficiencies in the original assay. The point mutations yielded modified guides with Gibbs free energies of -2.4 kJ·mol^{-1} and -0.8 kJ·mol^{-1}, respectively, for formation of the most stable guide sequence secondary structure. We designed corresponding C-Check

plasmids and assayed the efficiency of SpCas9- and eSpCas9-mediated cleavage with these mutated guides and their original counterparts in triplicates. As expected, the gRNAs with weaker secondary structure formation exhibited much higher efficiencies (Fig. 3C, 144% (SpCas9) and 81% (eSpCas9) increase for VEGFA_2; 101% and 210% increase for Chr1_C2, $P < 0.001$). Altering gRNA Chr14_1 from having a Gibbs free energy of 4.5 kJ·mol^{-1} to −1.6 kJ·mol^{-1} for formation of the most stable secondary structure resulted in a slight increase in activity which was only significant for eSpCas9 (24%, $P < 0.05$). Other factors such as DNA melting temperature of the target site and CG-content at either the seed region or whole guide sequence were not found to affect cleavage efficiency significantly (Fig. S5B–I).

Discussion

In the present study, we have validated previous data showing that eSpCas9 [9] generally exhibits wild-type-level on-target activity in human cancer cells. However, eSpCas9 exhibits significantly reduced activity for certain target sites and in fibroblasts [9,10]; a phenomenon which is dependent on intrinsic properties of the gRNA sequence since it is repeated in the endogenous cleavage activity and the plasmid-based surrogate reporter system C-Check. The chromatin-related factors might have a higher impact on eSpCas9 activity than SpCas9, which, however, should be investigated thoroughly in future studies. The C-Check system and similar systems described by Kim *et al.* and Ren *et al.* may provide a method for preselection of target sites with higher on-target activity for eSpCas9 [22,23].

We also show that both eSpCas9 and SpCas9 operate more efficiently in euchromatin than in heterochromatin in both HeLa and HEK293T cells when targeting endogenous loci, and that this is also the case for SpCas9 in human fibroblasts. This is in agreement with previous studies using inducible model systems and library-based approaches in human cancer cells [13–15]. The increased Cas9 activity in DNase I hypersensitive regions is also consistent with previous findings that nucleosomes impede Cas9 access [24], which may be explained by the relatively large Cas9 proteins potentially having easier access in less condensed chromatin.

We further found that the Gibbs free energy change for folding of the gRNA guide sequence alone, but not the gRNA in its entirety, influences the activity of both SpCas9 and eSpCas9. This is consistent with previous observations for SpCas9 [16,17]. Peak cleavage efficiencies were observed when $\Delta_r G \approx -1$ kJ·mol^{-1} for formation of the most stable guide sequence secondary structure [25], with very stable guide sequence secondary structures being particularly detrimental to Cas9 activity. We show this using 27 singly transfected gRNA-expressing plasmids in a plasmid-based reporter system in order to normalize for local chromatin context while retaining a cellular environment. In this assay, it is important to note that results may be influenced by variations in plasmid copy number, that the large number of targets in a single cell represents a nonphysiological scenario, and that protein expression is only an indirect readout of DSB frequency.

Thyme *et al.* hypothesize that the effect of internal gRNA secondary structures on Cas9 activity could be due to cotranscriptional folding of the gRNA, which is supported by data showing an increase in activity for some gRNAs following heating and reannealing. We propose an alternative and complementary hypothesis that there might be potential for folding of the guide sequence independently of the scaffold following formation of the Cas9-gRNA complex. A combination of these two processes may explain why only a subset of the low-activity gRNAs show increased activity upon reannealing in the assay of Thyme *et al.*, while also explaining why inactive gRNAs compete as effectively as active gRNAs for Cas9 binding. While formation of very stable secondary structures appears adversely to affect heteroduplex formation, weak secondary structures could potentially prevent interactions of an otherwise naked RNA strand with other cellular factors while still allowing for efficient strand invasion at the target site, something which would not necessarily have been detected in an *in vitro* environment. Our observation that a lack of guide sequence secondary structure may negatively affect Cas9 activity is thus also consistent with a model in which the guide sequence can form secondary structures following binding to Cas9. The folding energy of gRNAs including scaffold was not found to affect Cas9 activity although it is related to gRNA stability, and Moreno-Mateos *et al.* [5] found gRNA stability to correlate with gRNA-Cas9 efficiency. We suggest that this difference is due to our use of single gRNA-expressing plasmids as opposed to their coinjection of 80 *in vitro* transcribed gRNAs, in which case the more stable gRNAs might out-compete the less stable gRNAs.

In summary, we demonstrate that in addition to gRNA-specific factors including guide sequence secondary structure formation, the endogenous efficiency of Cas9 proteins is also dependent on local chromatin architecture for both SpCas9 and eSpCas9. This may aid in the future design of gRNAs for high-efficiency

Cas9 targeting, while also providing insights into the nature of gRNA-DNA dynamics. This knowledge suggests that given a choice of target sequences, a gRNA with potential for weak secondary structures is likely to be more efficient than alternatives with strong secondary structure, and targets in open chromatin regions are more likely to be effectively cleaved than alternatives in less accessible regions. We thereby hope to contribute to increasing the applicability and predictability of this and other CRISPR-Cas9 systems.

Acknowledgements

We thank Prof. Jacob Giehm Mikkelsen for the critical reading and comments, and the FACS Core Facility of Aarhus University for technical help with all FCM analysis. This work was funded by grants from the Danish Research Council for Independent Research (DFF-1337-00128), the Sapere Aude Young Research Talent Prize (DFF-1335-00763A), the Innovation Fund Denmark (BrainStem) and the Lundbeck Foundation (R173-2014-1105) to YL. The Lundbeck Foundation (R151-2013-14439) supported LB through the DREAM project and YL through the Casmere project. LL is supported by the Lundbeck Foundation (R219-2016-1375). JH, and FX are supported by a grant from the Shenzhen Municipal Government of China (NO.CXZZ20150330171838997).

Author contributions

YL, LB, and LL conceived the idea. KTJ, YL, and LL planned and prepared the manuscript and figures, and YL oversaw the study. KTJ, LF, TSP, JH, FX, YL, and LL performed experiments and analyzed the data. All authors revised the manuscript.

References

1 Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA and Charpentier E (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821.

2 Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823.

3 Mali P, Yang L, Esvelt KM, Aach J, Guell M, Dicarlo JE, Norville JE and Church GM (2013) RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826.

4 Li W, Teng F, Li T and Zhou Q (2013) Simultaneous generation and germline transmission of multiple gene mutations in rat using CRISPR-Cas systems. *Nat Biotechnol* **31**, 684–686.

5 Moreno-Mateos MA, Vejnar CE, Beaudoin JD, Fernandez JP, Mis EK, Khokha MK and Giraldez AJ (2015) CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nat Methods* **12**, 982–988.

6 Doench JG, Hartenian E, Graham DB, Tothova Z, Hegde M, Smith I, Sullender M, Ebert BL, Xavier RJ and Root DE (2014) Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat Biotechnol* **32**, 1262–1267.

7 Xu H, Xiao T, Chen CH, Li W, Meyer CA, Wu Q, Wu D, Cong L, Zhang F, Liu JS *et al.* (2015) Sequence determinants of improved CRISPR sgRNA design. *Genome Res* **25**, 1147–1157.

8 Fu Y, Foden JA, Khayter C, Maeder ML, Reyon D, Joung JK and Sander JD (2013) High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat Biotechnol* **31**, 822–826.

9 Slaymaker IM, Gao L, Zetsche B, Scott DA, Yan WX and Zhang F (2016) Rationally engineered Cas9 nucleases with improved specificity. *Science* **351**, 84–88.

10 Kleinstiver BP, Pattanayak V, Prew MS, Tsai SQ, Nguyen NT, Zheng Z and Joung JK (2016) High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490–495.

11 Fu Y, Sander JD, Reyon D, Cascio VM and Joung JK (2014) Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat Biotechnol* **32**, 279–284.

12 Wu X, Scott DA, Kriz AJ, Chiu AC, Hsu PD, Dadon DB, Cheng AW, Trevino AE, Konermann S, Chen S *et al.* (2014) Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nat Biotechnol* **32**, 670–676.

13 Daer RM, Cutts JP, Brafman DA and Haynes KA (2017) The impact of chromatin dynamics on Cas9-mediated genome editing in human cells. *ACS Synth Biol* **6**, 428–438.

14 Chen X, Rinsma M, Janssen JM, Liu J, Maggio I and Goncalves MA (2016) Probing the impact of chromatin conformation on genome editing tools. *Nucleic Acids Res* **44**, 6482–6492.

15 Chari R, Mali P, Moosburner M and Church GM (2015) Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. *Nat Methods* **12**, 823–826.

16 Thyme SB, Akhmetova L, Montague TG, Valen E and Schier AF (2016) Internal guide RNA interactions interfere with Cas9-mediated cleavage. *Nat Commun* **7**, 11750.

17 Wong N, Liu W and Wang X (2015) WU-CRISPR: characteristics of functional guide RNAs for the CRISPR/Cas9 system. *Genome Biol* **16**, 218.

18 Brinkman EK, Chen T, Amendola M and van Steensel B (2014) Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Res* **42**, e168.
19 Zhou Y, Liu Y, Hussmann D, Brøgger P, Al-Saaidi RA, Tan S, Lin L, Petersen TS, Zhou GQ, Bross P *et al.* (2016) Enhanced genome editing in mammalian cells with a modified dual-fluorescent surrogate system. *Cell Mol Life Sci* **73**, 2543–2563.
20 Sugimoto N, Nakano S, Yoneyama M and Honda K (1996) Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Res* **24**, 4501–4505.
21 Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA and Zhang F (2013) Genome engineering using the CRISPR-Cas9 system. *Nat Protoc* **8**, 2281–2308.
22 Ren C, Xu K, Liu Z, Shen J, Han F, Chen Z and Zhang Z (2015) Dual-reporter surrogate systems for efficient enrichment of genetically modified cells. *Cell Mol Life Sci* **72**, 2763–2772.
23 Kim H, Um E, Cho SR, Jung C, Kim H and Kim JS (2011) Surrogate reporters for enrichment of cells with nuclease-induced mutations. *Nat Methods* **8**, 941–943.
24 Horlbeck MA, Witkowsky LB, Guglielmi B, Replogle JM, Gilbert LA, Villalta JE, Torigoe SE, Tjian R and Weissman JS (2016) Nucleosomes impede Cas9 access to DNA in vivo and in vitro. *eLife*. pii: e12677.
25 Wolfsheimer S and Hartmann AK (2010) Minimum-free-energy distribution of RNA secondary structures: entropic and thermodynamic properties of rare events. *Phys Rev E Stat Nonlin Soft Matter Phys* **82**, 021902.

Supporting information

Additional Supporting Information may be found online in the supporting information tab for this article:

Fig. S1. (A) Schematic illustration of the C-Check system. Upon formation of a DNA double-strand break, repair by single-strand annealing can reconstitute GFP expression. CDS, coding sequence. (B) Comparison of on-target activity (ON) and mismatch tolerance (OT) for SpCas9 and eSpCas9. Values represent mean and one SD ($n = 3$). (C) Schematic illustration of endogenous targets for SpCas9 and eSpCas9, one fully complementary and one with an A:G mismatch at nucleotide +17 from the PAM. (D,E) SpCas9 and eSpCas9 activities for the targets in Fig. 1C as measured by C-Check (D) and TIDE (E). Controls were transfected with a plasmid expressing the empty gRNA scaffold. Error bars represent one SD ($n = 3$). Asterisks represent P value < 0.001 (***, Student's t-test) compared to the control.

Fig. S2. Genome browser illustration of targeted regions including CRISPR gRNA target sites (highlighted in blue), reference gene information (if any), DNase I hypersensitivity clusters, and hypersensitive sites (in 125 cell types) from ENCODE.

Fig. S3. (A) DNase I-seq reads-per-kilobase million (RPKM) values for the target regions investigated in HeLa cells (data retrieved from ENCODE). (B) Efficiency of SpCas9 and eSpCas9 targeting eight genomic loci in HeLa cells. The percentage of indels at each genomic locus was quantified by TIDE. Error bars represent one SD ($n = 3$). (C) DNase I-seq reads-per-kilobase million (RPKM) values for the target regions investigated in human fibroblasts (data retrieved from ENCODE). (D) Efficiency of SpCas9 targeting eight genomic loci in human fibroblasts together with control values from experimental groups transfected with an empty gRNA backbone plasmid. The percentage of indels at each genomic locus was quantified by TIDE. Error bars represent one SD ($n = 3$).

Fig. S4. (A) Dot plot of efficiency in the plasmid-based C-Check system for 20 gRNAs with either SpCas9 or eSpCas9 against HEK293T DNase-seq RPKM values for the corresponding endogenous regions. Regression lines represent linear line of best fit. (B) Dot plot of % indels in the TIDE assay against C-Check efficiency for all 20 targets investigated in HEK293T cells. Regression line represents linear line of best fit.

Fig. S5. (A) Dot plot of SpCas9 and eSpCas9 efficiency quantified by C-Check against the Gibbs free energy change ($\Delta_r G$, $kJ \cdot mol^{-1}$) for formation of the most stable secondary structure for each gRNA including scaffold. (B–I) Dot plots of SpCas9 (left) or eSpCas9 (right) efficiency in the C-Check reporter system against the following parameters: B-C, DNA melting temperature of each target sequence. D-E, GC-content of the 6 most PAM-proximal nucleotides of each gRNA. F-G, GC-content of the 12 most PAM-proximal nucleotides of each gRNA. H-I, GC-content of the entire gRNAs. Linear regression analyses were performed for all correlation comparisons.

Table S1. gRNA sequences.

Table S2. PCR primers and program.

Table S3. Values of DNase I-seq (HEK293T, Hela, and human fibroblasts), CpG methylation (HEK293T), H3K4me3 ChIP-seq (HEK293T), CTCF ChIP-seq (HEK293T), and RNA-seq (HEK293T) of targeted regions in this study.