Supplementary Information

DNA Targeting Specificity of the RNA-guided Cas9 Nuclease

Patrick D. Hsu[1,2,3*], David A. Scott[1,2*], Joshua A. Weinstein[1,2‡], F. Ann Ran[1,2,3‡], Silvana Konermann[1,2], Vineeta Agarwala[1], Yinqing Li[1,2], Eli J. Fine[4], Xuebing Wu[5], Ophir Shalem[1,2], Thomas J. Cradick[4], Luciano A. Marraffini[6], Gang Bao[4], and Feng Zhang[1,2†]

[1] Broad Institute of MIT and Harvard
7 Cambridge Center
Cambridge, MA 02142, USA

[2] McGovern Institute for Brain Research
Department of Brain and Cognitive Sciences
Department of Biological Engineering
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

[3] Department of Molecular and Cellular Biology
Harvard University
Cambridge, MA 02138, USA

[4] Department of Biomedical Engineering
Georgia Institute of Technology and Emory University
Atlanta, Georgia 30332, USA

[5] Koch Institute for Integrative Cancer Research
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

[6] Laboratory of Bacteriology
The Rockefeller University
1230 York Ave.
New York, NY 10065, USA

[*]These authors contributed equally to this work.
[†]To whom correspondence should be addressed: zhang@broadinstitute.org.

Nature Biotechnology: doi:10.1038/nbt.2647

SUPPLEMENTARY METHODS

Cell culture and transfection

Human embryonic kidney (HEK) cell line 293FT (Life Technologies) was maintained in Dulbecco's modified Eagle's Medium (DMEM) supplemented with 10% fetal bovine serum (HyClone), 2mM GlutaMAX (Life Technologies), 100U/mL penicillin, and 100μg/mL streptomycin at 37°C with 5% CO_2 incubation.

293FT cells were seeded onto 6-well plates, 24-well plates, or 96-well plates (Corning) 24 hours prior to transfection. Cells were transfected using Lipofectamine 2000 (Life Technologies) at 80-90% confluency following the manufacturer's recommended protocol. For each well of a 6-well plate, a total of 1 ug of Cas9+sgRNA plasmid was used. For each well of a 24-well plate, a total of 500ng Cas9+sgRNA plasmid was used unless otherwise indicated. For each well of a 96-well plate, 65 ng of Cas9 plasmid was used at a 1:1 molar ratio to the U6-sgRNA PCR product.

Human embryonic stem cell line HUES9 (Harvard Stem Cell Institute core) was maintained in feeder-free conditions on GelTrex (Life Technologies) in mTesR medium (Stemcell Technologies) supplemented with 100ug/ml Normocin (InvivoGen). HUES9 cells were transfected with Amaxa P3 Primary Cell 4-D Nucleofector Kit (Lonza) following the manufacturer's protocol.

SURVEYOR nuclease assay for genome modification

293FT and HUES9 cells were transfected with DNA as described above. Cells were incubated at 37ºC for 72 hours post-transfection prior to genomic DNA extraction. Genomic DNA was extracted using the QuickExtract DNA Extraction Solution (Epicentre) following the manufacturer's protocol. Briefly, pelleted cells were resuspended in QuickExtract solution and incubated at 65ºC for 15 minutes, 68°C for 15 minutes, and 98ºC for 10 minutes.

The genomic region flanking the CRISPR target site for each gene was PCR amplified (primers listed in Supplementary Table 2), and products were purified using QiaQuick Spin Column (Qiagen) following the manufacturer's protocol. 400ng total of the purified PCR products were mixed with 2□l 10X Taq DNA Polymerase PCR buffer (Enzymatics) and ultrapure water to a final volume of 20□l, and subjected to a re-annealing process to enable heteroduplex formation: 95°C for 10min, 95°C to 85°C ramping at −2°C/s, 85°C to 25°C at −

0.25ºC/s, and 25ºC hold for 1 minute. After re-annealing, products were treated with SURVEYOR nuclease and SURVEYOR enhancer S (Transgenomics) following the manufacturer's recommended protocol, and analyzed on 4-20% Novex TBE poly-acrylamide gels (Life Technologies). Gels were stained with SYBR Gold DNA stain (Life Technologies) for 30 minutes and imaged with a Gel Doc gel imaging system (Bio-rad). Quantification was based on relative band intensities. Indel percentage was determined by the formula, 100 x (1 - (1 - (b + c) / (a + b + c))$^{1/2}$), where a is the integrated intensity of the undigested PCR product, and b and c are the integrated intensities of each cleavage product.

Northern blot analysis of tracrRNA expression in human cells

Northern blots were performed as previously described[1]. Briefly, RNAs were extracted using the mirPremier microRNA Isolation Kit (Sigma) and heated to 95ºC for 5 min before loading on 8% denaturing polyacrylamide gels (SequaGel, National Diagnostics). Afterwards, RNA was transferred to a pre-hybridized Hybond N+ membrane (GE Healthcare) and crosslinked with Stratagene UV Crosslinker (Stratagene). Probes were labeled with [gamma-32P] ATP (Perkin Elmer) with T4 polynucleotide kinase (New England Biolabs). After washing, membrane was exposed to phosphor screen for one hour and scanned with phosphorimager (Typhoon).

Bisulfite sequencing to assess DNA methylation status

Genomic DNA from 293FT cells was isolated with the DNeasy Blood & Tissue Kit (Qiagen) and bisulfite converted with EZ DNA Methylation-Lightning Kit (Zymo Research). Bisulfite PCR was conducted using KAPA2G Robust HotStart DNA Polymerase (KAPA Biosystems) with primers designed using the Bisulfite Primer Seeker (Zymo Research, Supplementary Table 2). Resulting PCR amplicons were gel-purified, digested with EcoRI and HindIII, and ligated into a pUC19 backbone prior to transformation. Individual clones were then Sanger sequenced to assess DNA methylation status.

Nature Biotechnology: doi:10.1038/nbt.2647

In vitro transcription and cleavage assay

Whole cell lysates from 293FT cells were prepared with lysis buffer (20 mM HEPES, 100 mM KCl, 5 mM MgCl2, 1 mM DTT, 5% glycerol, 0.1% Triton X-100) supplemented with Protease Inhibitor Cocktail (Roche). T7-driven sgRNA was transcribed *in vitro* using custom oligos (Supplementary Sequences) and HiScribe T7 *In Vitro* Transcription Kit (NEB), following the manufacturer's recommended protocol. To prepare methylated target sites, pUC19 plasmid was methylated by M.*Sss*I and tested by digestion with HpaII. Unmethylated and successfully methylated pUC19 plasmids were linearized by NheI. The *in vitro* cleavage assay was performed as follows: for a 20 uL cleavage reaction, 10 uL of cell lysate was incubated with 2 uL cleavage buffer (100 mM HEPES, 500 mM KCl, 25 mM MgCl2, 5 mM DTT, 25% glycerol), 1 ug *in vitro* transcribed RNA, and 300 ng pUC19 plasmid DNA.

Deep sequencing to assess targeting specificity

HEK 293FT cells plated in 96-well plates were transfected with Cas9 plasmid DNA and single guide RNA (sgRNA) PCR cassette 72 hours prior to genomic DNA extraction (Supplementary Fig. 4). The genomic region flanking the CRISPR target site for each gene was amplified (Supplementary Fig. 6, Supplementary Table 5, Supplementary Sequences) by a fusion PCR method to attach the Illumina P5 adapters as well as unique sample-specific barcodes to the target amplicons (schematic described in Supplementary Figure 5). PCR products were purified using EconoSpin 96-well Filter Plates (Epoch Life Sciences) following the manufacturer's recommended protocol.

Barcoded and purified DNA samples were quantified by Quant-iT PicoGreen dsDNA Assay Kit or Qubit 2.0 Fluorometer (Life Technologies) and pooled in an equimolar ratio. Sequencing libraries were then sequenced with the Illumina MiSeq Personal Sequencer (Life Technologies).

Sequencing data analysis and indel detection

MiSeq reads were filtered by requiring an average Phred quality (Q score) of at least 23, as well as perfect sequence matches to barcodes and amplicon forward primers. Reads from on- and off-target loci were analyzed by first performing Smith-Waterman alignments against amplicon sequences that included 50 nucleotides upstream and downstream of the target site (a

Nature Biotechnology: doi:10.1038/nbt.2647

total of 120 bp). Alignments, meanwhile, were analyzed for indels from 5 nucleotides upstream to 5 nucleotides downstream of the target site (a total of 30 bp). Analyzed target regions were discarded if part of their alignment fell outside the MiSeq read itself, or if matched base-pairs comprised less than 85% of their total length.

Negative controls for each sample provided a gauge for the inclusion or exclusion of indels as putative cutting events. For each sample, an indel was counted only if its quality score exceeded $\mu - \sigma$, where μ was the mean quality-score of the negative control corresponding to that sample and σ was the standard deviation of the same. This yielded whole target-region indel rates for both negative controls and their corresponding samples. Using the negative control's per-target-region-per-read error rate, q, the sample's observed indel count n, and its read-count R, a maximum-likelihood estimate for the fraction of reads having target-regions with true-indels, p, was derived by applying a binomial error model, as follows.

Letting the (unknown) number of reads in a sample having target regions incorrectly counted as having at least 1 indel be E, we can write (without making any assumptions about the number of true indels)

$$\text{Prob}(E|p) = \binom{R(1-p)}{E} q^E (1-q)^{R(1-p)-E}$$

since $R(1-p)$ is the number of reads having target-regions with no true indels. Meanwhile, because the number of reads observed to have indels is n, $n = E + Rp$, i.e. the number of reads having target-regions with errors but no true indels *plus* the number of reads whose target-regions *correctly* have indels. We can then re-write the above

$$\text{Prob}(E|p) = \text{Prob}(n = E + Rp|p) = \binom{R(1-p)}{n-Rp} q^{n-Rp} (1-q)^{R-n}$$

Taking all values of the frequency of target-regions with true-indels p to be equally probable *a priori*, $\text{Prob}(n|p) \propto \text{Prob}(p|n)$. The maximum-likelihood estimate (MLE) for the frequency of target regions with true-indels was therefore set as the value of p that maximized $\text{Prob}(n|p)$. This was evaluated numerically.

Nature Biotechnology: doi:10.1038/nbt.2647

In order to place error bounds on the true-indel read frequencies in the sequencing libraries themselves, Wilson score intervals[2] were calculated for each sample, given the MLE-estimate for true-indel target-regions, Rp, and the number of reads R. Explicitly, the lower bound l and upper bound u were calculated as

$$l = \left(Rp + \frac{z^2}{2} - z\sqrt{Rp(1-p) + z^2/4} \right) / (R + z^2)$$

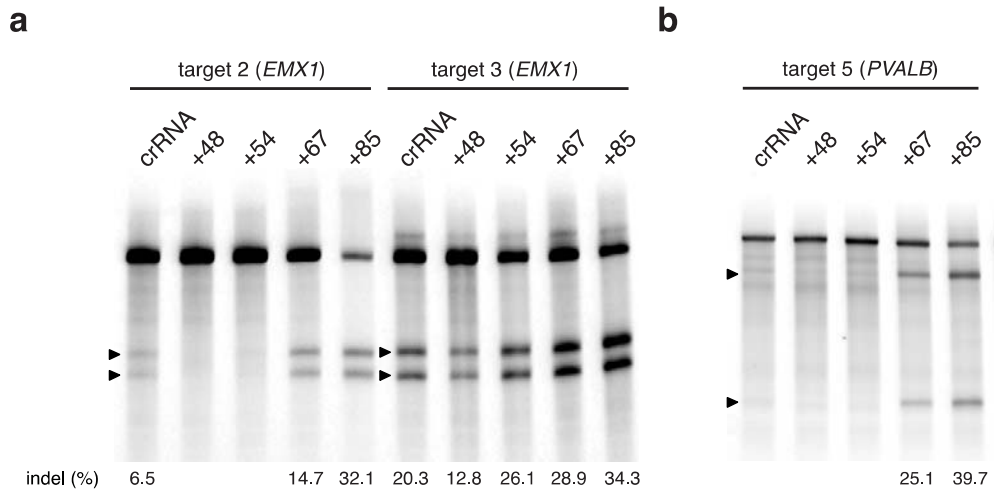
$$u = \left(Rp + \frac{z^2}{2} + z\sqrt{Rp(1-p) + z^2/4} \right) / (R + z^2)$$

where z, the standard score for the confidence required in normal distribution of variance 1, was set to 1.96, meaning a confidence of 95%. The maximum upper bounds and minimum lower bounds for each biological replicate are listed in Supplementary Tables 5-8.

qRT-PCR analysis of relative Cas9 and sgRNA expression

72 hours post-transfection, total RNA from 293FT cells was harvested with miRNeasy Micro Kit (Qiagen). Reverse-strand synthesis for sgRNAs was performed with qScript Flex cDNA kit (VWR) and custom first-strand synthesis primers (Supplementary Table 2). qPCR analysis was performed with Fast SYBR Green Master Mix (Life Technologies) and custom primers (Supplementary Table 2), using GAPDH as an endogenous control. Relative quantification was calculated by the ΔΔCT method.

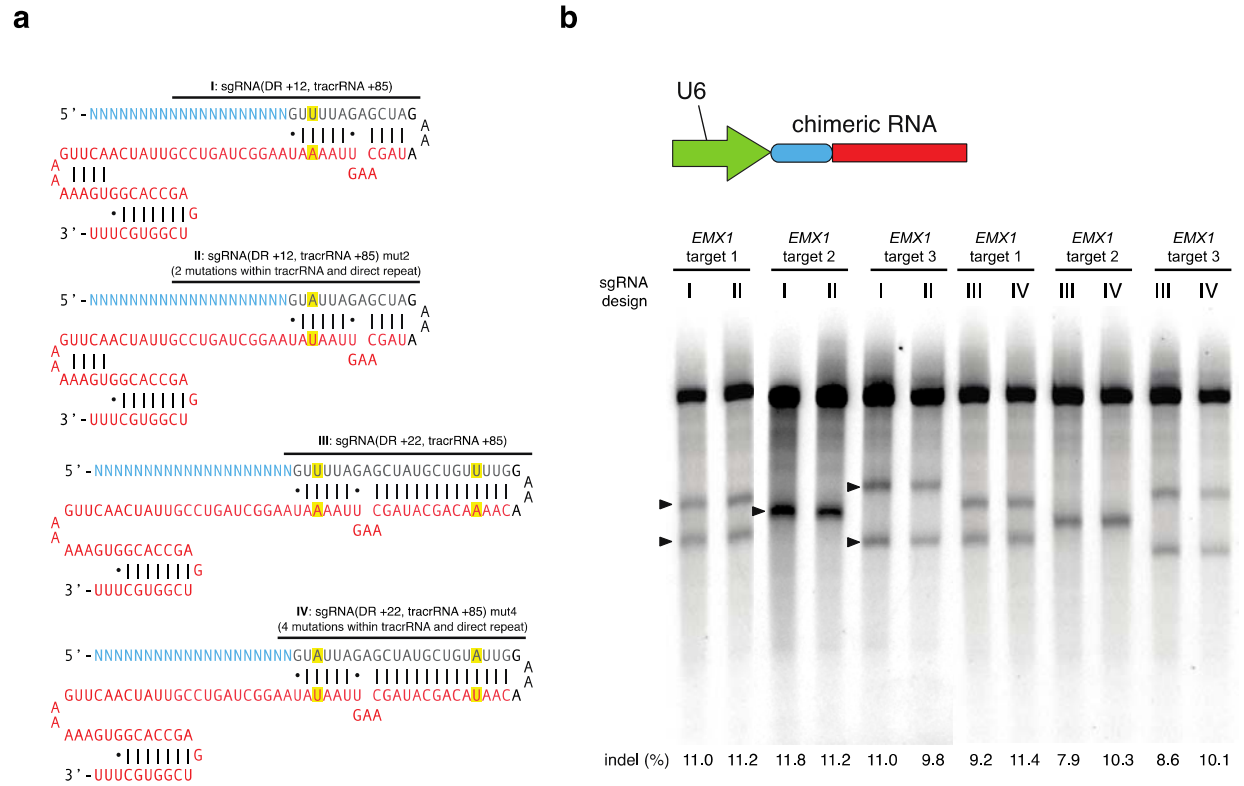
SUPPLEMENTARY FIGURE 1



Supplementary Figure 1 | Modification efficiencies of CRISPR-Cas system for additional human genomic targets. DNA expression vectors carrying SpCas9 and crRNA-tracrRNA pair or single guide RNA (sgRNA) are co-transfected into 293FT cells. Cleavage efficiency (% indel) is assessed using the SURVEYOR nuclease assay as described[1]. Modification efficiencies at **a,** 2 *EMX1* loci and **b,** 1 *PVALB* locus are shown. All target site sequences are listed in **Supplementary Table 1**. Arrows indicate the expected SURVEYOR fragments.

Nature Biotechnology: doi:10.1038/nbt.2647

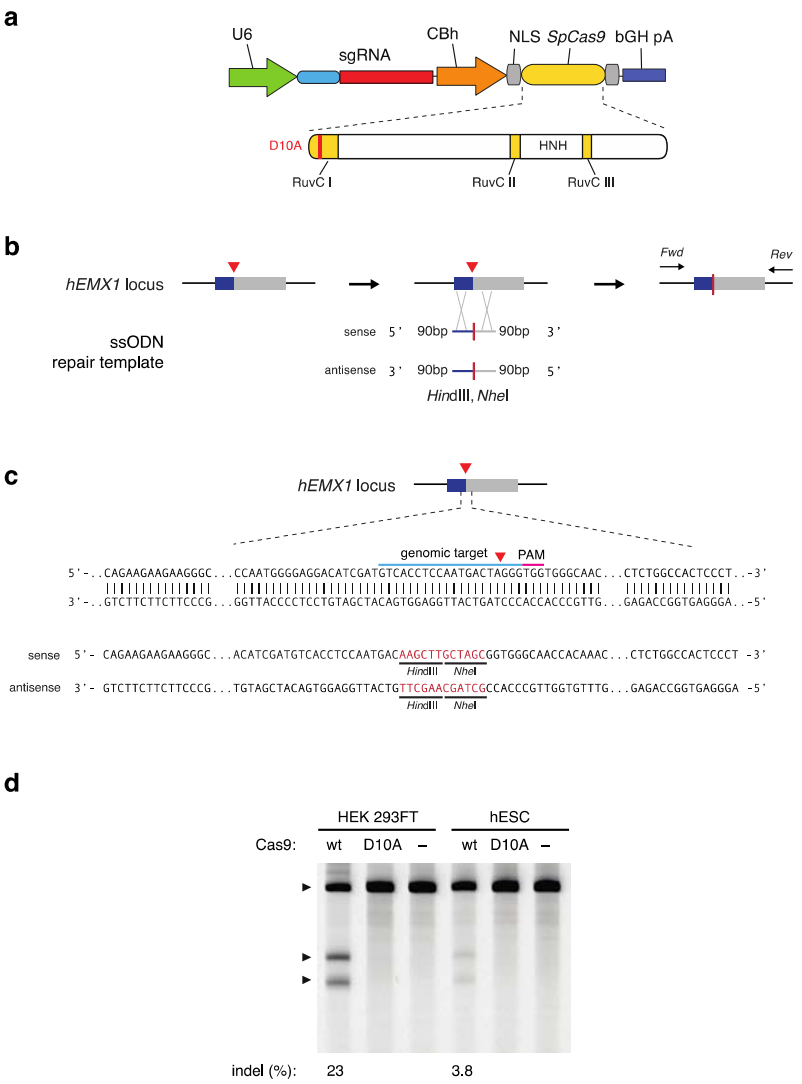
SUPPLEMENTARY FIGURE 2



Supplementary Figure 2 | Further optimization of CRISPR-Cas sgRNA architecture. a, Schematic of four additional sgRNA architectures, I-IV. Each consists of a 20-nt guide sequence (blue) joined to the direct repeat (DR, grey), which base-pairs with the tracrRNA (red). The DR-tracrRNA hybrid is truncated at +12 or +22, as indicated, with an artificial GAAA stem loop. tracrRNA truncation positions are numbered according to the previously reported transcription start site for tracrRNA (Supplementary Figure 11 of reference)[3]. sgRNA architectures II and IV carry mutations within their poly-U tracts, which could serve as premature transcriptional terminators for U6 promoter. **b,** SURVEYOR assay for SpCas9-mediated indels at the human *EMX1* locus for target sites 1-3. Arrows indicate the expected SURVEYOR fragments ($n = 3$).

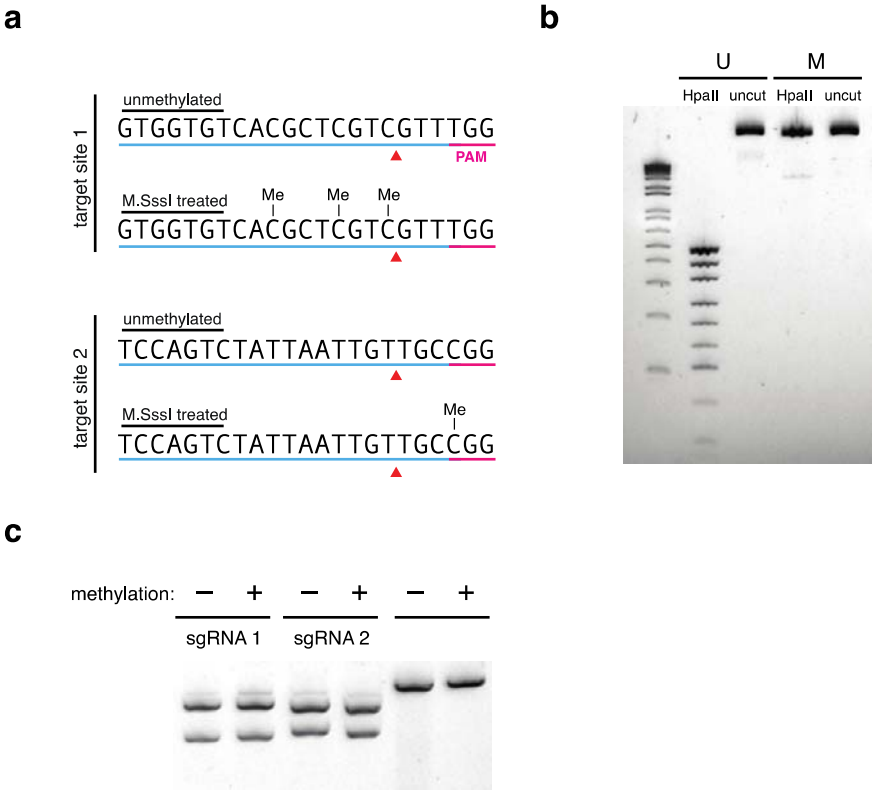
Nature Biotechnology: doi:10.1038/nbt.2647

SUPPLEMENTARY FIGURE 3



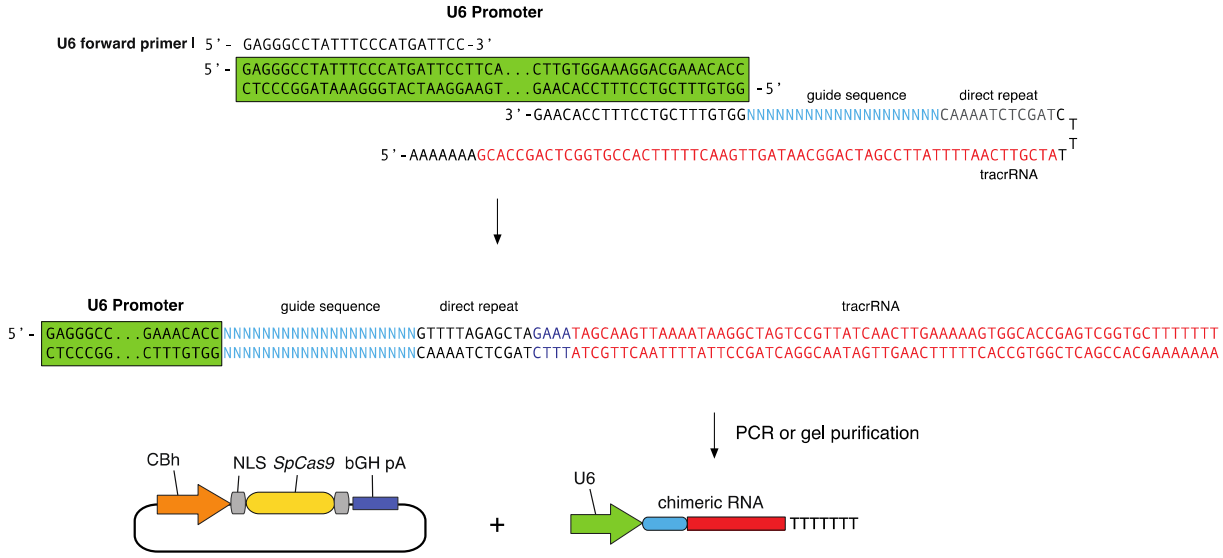
Supplementary Figure 3 | Genome editing via homologous recombination. a, Schematic of SpCas9 nickase, with D10A mutation in the RuvC I catalytic domain. **b,** Schematic representing homologous recombination (HR) at the human *EMX1* locus using either sense or antisense single stranded oligonucleotides as repair templates. Red arrow above indicates sgRNA cleavage site; PCR primers for genotyping (Supplementary Table 2) are indicated as arrows in right panel. **c,** Sequence of region modified by HR. **d,** SURVEYOR assay for wildtype (wt) and nickase (D10A) SpCas9-mediated indels at the *EMX1* target 1 locus ($n = 3$). Arrows indicate positions of expected fragment sizes.

SUPPLEMENTARY FIGURE 4



Supplementary Figure 4 | SpCas9 cleaves methylated targets *in vitro*. a, Sequence of CpG dinucleotide-containing targets in pUC19 plasmid methylated *in vitro* by M.*SssI*. Methyl-CpGs in either the target sequence or PAM are indicated; arrows indicate expected cleavage site. **b,** Unmethylated (U) or methylated (M) pUC19 was subjected to restriction digest by the methylation-sensitive restriction enzyme HpaII. Unmethylated pUC19 is digested into a ladder while M.SssI-treated pUC19 is protected from HpaII digestion. **c,** Cleavage of either unmethylated or methylated targets 1 and 2 on linearized pUC19 by SpCas9. No sgRNAs are present in negative control lanes.

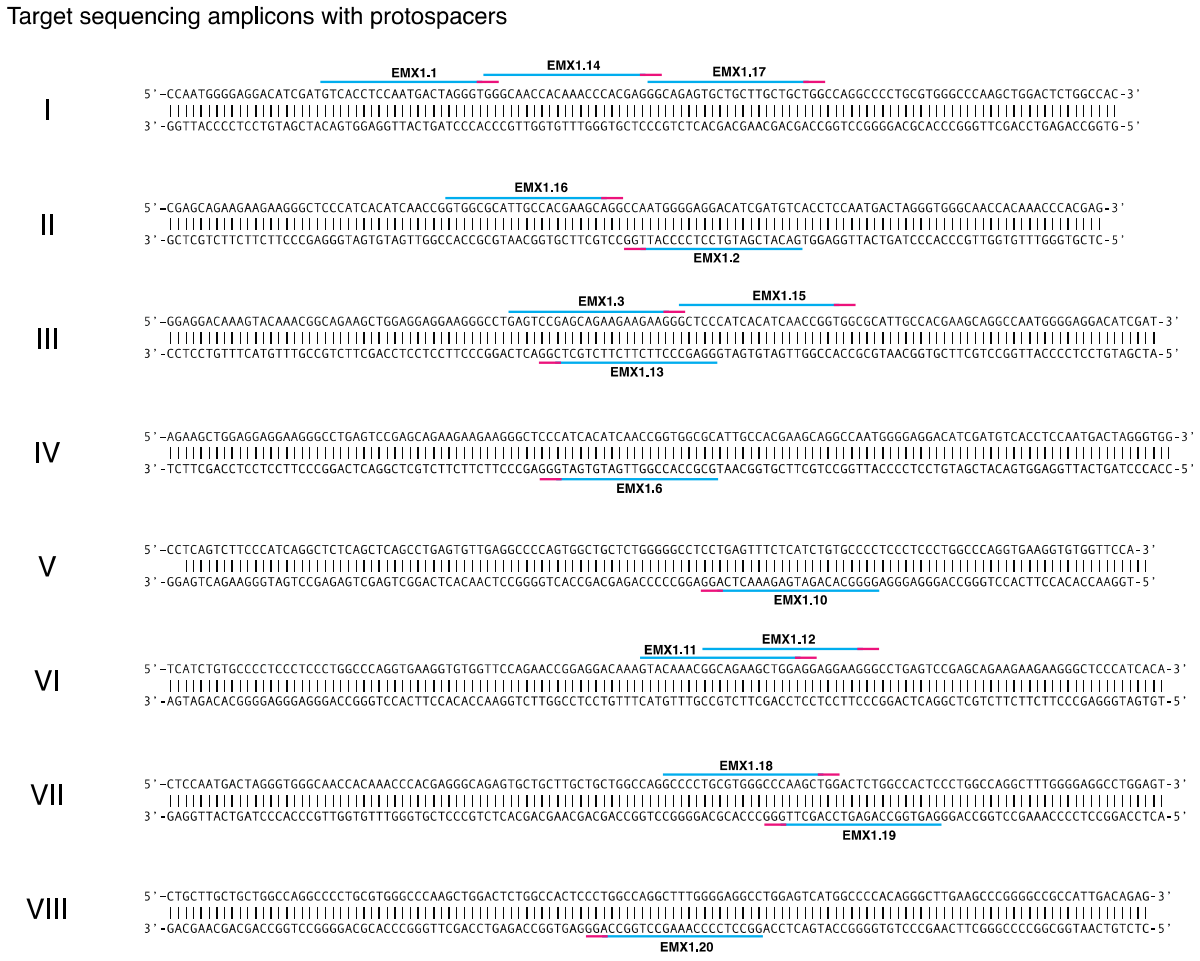
SUPPLEMENTARY FIGURE 5



Supplementary Figure 5 | PCR cassette for sgRNA expression. a, Schematic of a PCR-based method for rapid and efficient CRISPR targeting in mammalian cells. A plasmid containing the human RNA polymerase III promoter U6 is PCR-amplified using a U6-specific forward primer and a reverse primer carrying the reverse complement of part of the U6 promoter, the sgRNA(+85) scaffold with guide sequence, and 7 T nucleotides for transcriptional termination. The resulting PCR product is purified and co-delivered with a plasmid carrying Cas9 driven by the CBh promoter.

Nature Biotechnology: doi:10.1038/nbt.2647

SUPPLEMENTARY FIGURE 6



Supplementary Figure 6 | The human EMX1 locus with target sites. Schematic of the human *EMX1* locus showing the location of 15 target DNA sites, indicated by blue lines with corresponding PAM in magenta.