nature
biotechnology

# Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells

Xuebing Wu[1,2], David A Scott[3,4,8], Andrea J Kriz[5,8], Anthony C Chiu[1,5], Patrick D Hsu[3,4,6], Daniel B Dadon[5,7], Albert W Cheng[2,7], Alexandro E Trevino[3,4], Silvana Konermann[3,4], Sidi Chen[1], Rudolf Jaenisch[7], Feng Zhang[3,4] & Phillip A Sharp[1,5]

**Bacterial type II CRISPR-Cas9 systems have been widely adapted for RNA-guided genome editing and transcription regulation in eukaryotic cells, yet their *in vivo* target specificity is poorly understood. Here we mapped genome-wide binding sites of a catalytically inactive Cas9 (dCas9) from *Streptococcus pyogenes* loaded with single guide RNAs (sgRNAs) in mouse embryonic stem cells (mESCs). Each of the four sgRNAs we tested targets dCas9 to between tens and thousands of genomic sites, frequently characterized by a 5-nucleotide seed region in the sgRNA and an NGG protospacer adjacent motif (PAM). Chromatin inaccessibility decreases dCas9 binding to other sites with matching seed sequences; thus 70% of off-target sites are associated with genes. Targeted sequencing of 295 dCas9 binding sites in mESCs transfected with catalytically active Cas9 identified only one site mutated above background levels. We propose a two-state model for Cas9 binding and cleavage, in which a seed match triggers binding but extensive pairing with target DNA is required for cleavage.**

© 2014 Nature America, Inc. All rights reserved.

Many bacterial and archaeal genomes encode clustered, regularly interspaced, short palindromic repeats (CRISPRs), which are transcribed and processed into short RNAs that guide CRISPR-associated (Cas) proteins to cleave foreign nucleic acids[1–5]. To target particular genomic loci in eukaryotic cells, the type II CRISPR-Cas system from *Streptococcus pyogenes* has been adapted so that it requires the nuclease Cas9 and one sgRNA[6–9]. The first ~20 nucleotides of the sgRNA (the guide region) are complementary to the target DNA site, which also needs to contain a sequence called the protospacer adjacent motif (PAM), typically NGG[10].

The simplicity of targeting any locus with a single protein and a programmable sgRNA has quickly led to widespread use of Cas9 (refs. 11,12) in applications such as genome editing[7,8,13–16], disease gene repair[17,18] and knock-in of specific tags[8,19]. The catalytically inactive dCas9 (containing D10A and H840A mutations) alone or when fused to activators or repressors has been used to modulate transcription[20–25], and dCas9 has also been fused to GFP to allow imaging of genomic loci in living cells[26].

However, the mechanism of target recognition and target specificity of the Cas9 protein remains poorly understood[8,9,24,27–32]. Most previous studies have analyzed a set of candidate off-target sites with up to five mismatches to the designed on-target site. These studies have examined *in vitro* cleavage, cleavage-induced indels or reporter gene expression change as the read-out, rather than direct binding[9,24,27,32]. Base pairing in the first 10–12 nucleotides adjacent to PAM (defined as the 'seed') was found to be generally more important than pairing in the rest of the guide region[6,8,16,33]. However, large variations were observed across target sites, cell types and species regarding the importance of base pairing at each position[28]. Some studies have shown that Cas9 is highly specific[21,30,31], whereas other studies have demonstrated substantial Cas9 off-target activity[9,24,27,29,32]. Epigenetic features such as CpG methylation and chromatin accessibility have been reported to have little effect on targeting[9,23].

To our knowledge, there has been no previous report of genome-wide binding maps of dCas9. Using chromatin immunoprecipitation followed by sequencing (ChIP-seq) for dCas9 binding in mESCs, our data reveal a well-defined seed region for target binding and a very large number of off-target binding sites, most of which do not seem to undergo substantial cleavage by Cas9. Our observations explain some of the previously observed heterogeneity, provide insights into target recognition and the cleavage process, and could guide future target design.

## RESULTS

### Genome-wide binding of dCas9-sgRNA

To map dCas9 *in vivo* binding sites, we generated mESCs with a stably integrated vector encoding hemagglutinin (HA)-tagged dCas9 (**Fig. 1a**), and performed ChIP-seq on cells transfected with either no sgRNA or one of four sgRNAs (Phc1-sg1, Phc1-sg2, Nanog-sg2 and Nanog-sg3) targeting the promoters of *Phc1* or *Nanog*, respectively.

[1]David H. Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [2]Computational and Systems Biology Graduate Program, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [3]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [4]McGovern Institute for Brain Research, Department of Brain and Cognitive Sciences, Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [5]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [6]Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts, USA. [7]Whitehead Institute for Biomedical Research, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [8]These authors contributed equally to this work. Correspondence should be addressed to F.Z. (zhang@broadinstitute.org) or P.A.S. (sharppa@mit.edu).

Received 3 February; accepted 27 March; published online 20 April 2014; doi:10.1038/nbt.2889

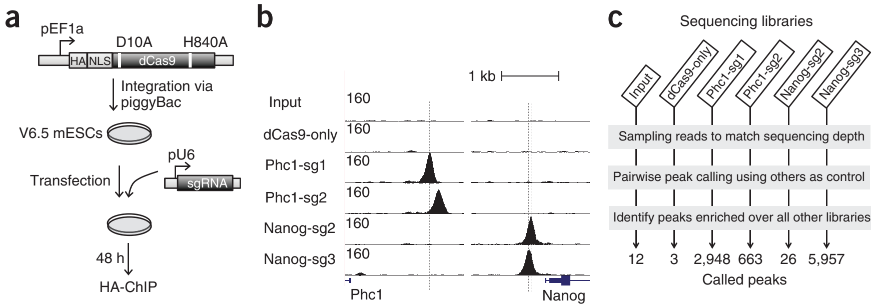


**Figure 1** Genome-wide *in vivo* binding of dCas9-sgRNA. (**a**) Schematic of dCas9 ChIP. EF1a promoter-driven, HA-tagged dCas9 with a nuclear localization signal (NLS) is integrated into the genome of mESCs by the piggyBac system. Plasmids containing U6 promoter-driven sgRNAs were transfected into these cells, and ChIP was carried out 2 d later with an antibody specific for HA. (**b**) ChIP signals (normalized read counts) around on-target sites. Vertical dotted lines indicate designed target sites (the region complementary to the sgRNA). (**c**) Strategy used for peak calling. Reads were sampled from each library, and for each library, peaks were called using all the other five libraries as a control. Only peaks called over all the other five libraries were retained. The numbers at the bottom indicate the numbers of peaks called for each library using these criteria.

© 2014 Nature America, Inc. All rights reserved.

For each sgRNA, we observed ~100-fold enrichment for dCas9 at the on-target site compared to flanking regions, and the spatial resolution is sufficient to distinguish between two binding sites separated by 22 base pairs (bp) (Nanog-sg2 and Nanog-sg3) (**Fig. 1b**).

Using the standard ChIP-seq peak-calling procedure MACS[34] and comparing immunoprecipitated material to input (whole-cell extract) DNA, we identified 2,000 to 20,000 peaks in each sequencing library (**Supplementary Fig. 1a**). The library from cells expressing dCas9 but not transfected with sgRNAs (dCas9-only ChIP) had 2,115 peaks. Most (77%) of the peaks detected in the dCas9-only ChIP were also detected in libraries prepared from dCas9-sgRNA immunoprecipitations (**Supplementary Fig. 1b**). The peaks in the dCas9-only ChIP were enriched in open chromatin regions (**Supplementary Fig. 2a**), and 41% contained GG/CC-rich motifs that closely resemble CTCF binding motifs (**Supplementary Fig. 2b–d**). These dCas9-only ChIP peaks could either represent 'sampling' by dCas9 of accessible sites containing NGG[33] or transcription-dependent artifacts as previously reported for GFP ChIP in yeast[35].

To identify sgRNA-dependent dCas9 binding sites, we matched sequencing depth by randomly sampling an equal number of reads from all six libraries (including input) and then performed pair-wise peak calling with MACS using each of the other five libraries as the control; we retained only peaks that were enriched over all the other five libraries (**Fig. 1c**). Using this approach, only three peaks were specific for dCas9-only ChIP. The number of sgRNA-specific peaks varied substantially; for example, there were nearly 6,000 peaks for Nanog-sg3 but only 26 peaks for Nanog-sg2 (**Fig. 1c**). Many of the off-target peaks showed high binding levels, as defined by the peak height relative to on-target peaks after subtracting dCas9-only reads at that site. For example, there were 91 off-target peaks with more than 50% of the binding level of the on-target site for Nanog-sg3 (**Supplementary Table 1**). These results suggest that there are substantial numbers of off-target binding sites, and the majority of the dCas9-sgRNA complexes bind outside the designed target site.

## A 5-nucleotide seed for dCas9 binding

Sequence motifs enriched within 50 bp of peak summits were identified using MEME-ChIP[36]. The top motif found for each ChIP library matched the PAM-proximal region of the transfected sgRNA plus the PAM NGG (**Fig. 2a** and **Supplementary Fig. 3**). For three of the four sgRNAs, only PAM-proximal positions 1 to 5 in the target DNA showed a preference of base match to the guide (**Fig. 2a**). We therefore define positions 1–5 as the 'seed' region of the sgRNA. For Nanog-sg2, the guide match extends to about 10–12 bases to the 5′ end, possibly due to the presence of multiple U's in the seed that lowers the thermodynamic stability of the sgRNA-DNA interaction. For Nanog-sg3 and Phc1-sg2, an exact match to the 5-nucleotide seed followed by NGG (seed+NGG) within 50 bp of peak summits explained 96% and 97% of the peaks, respectively. When the 50 nucleotides flanking peak summits were shuffled, preserving dinucleotide frequency, ≤5.7% of the shuffled sequences contained seed+NGG (**Fig. 2a**) for all four sgRNAs. Moreover, the seed+NGG sequences were highly enriched at the center of the peak (**Fig. 2a**, right), suggesting these sequences are directly bound by sgRNA-guided dCas9.

We found that seed+NGG is sufficient for Cas9 binding *in vivo* and *in vitro*. For example, there were 92 peaks in the Nanog-sg3 sample containing only seed+NGG matches, that is, mismatches at all the other 15 positions. The strongest peak containing only seed+NGG showed 52% binding activity relative to the on-target site (**Fig. 2b**). *In vitro* gel shift assays confirmed specific binding to seed+NGG–only substrates but with lower affinity than to the on-target site (**Fig. 2c**,**d**).

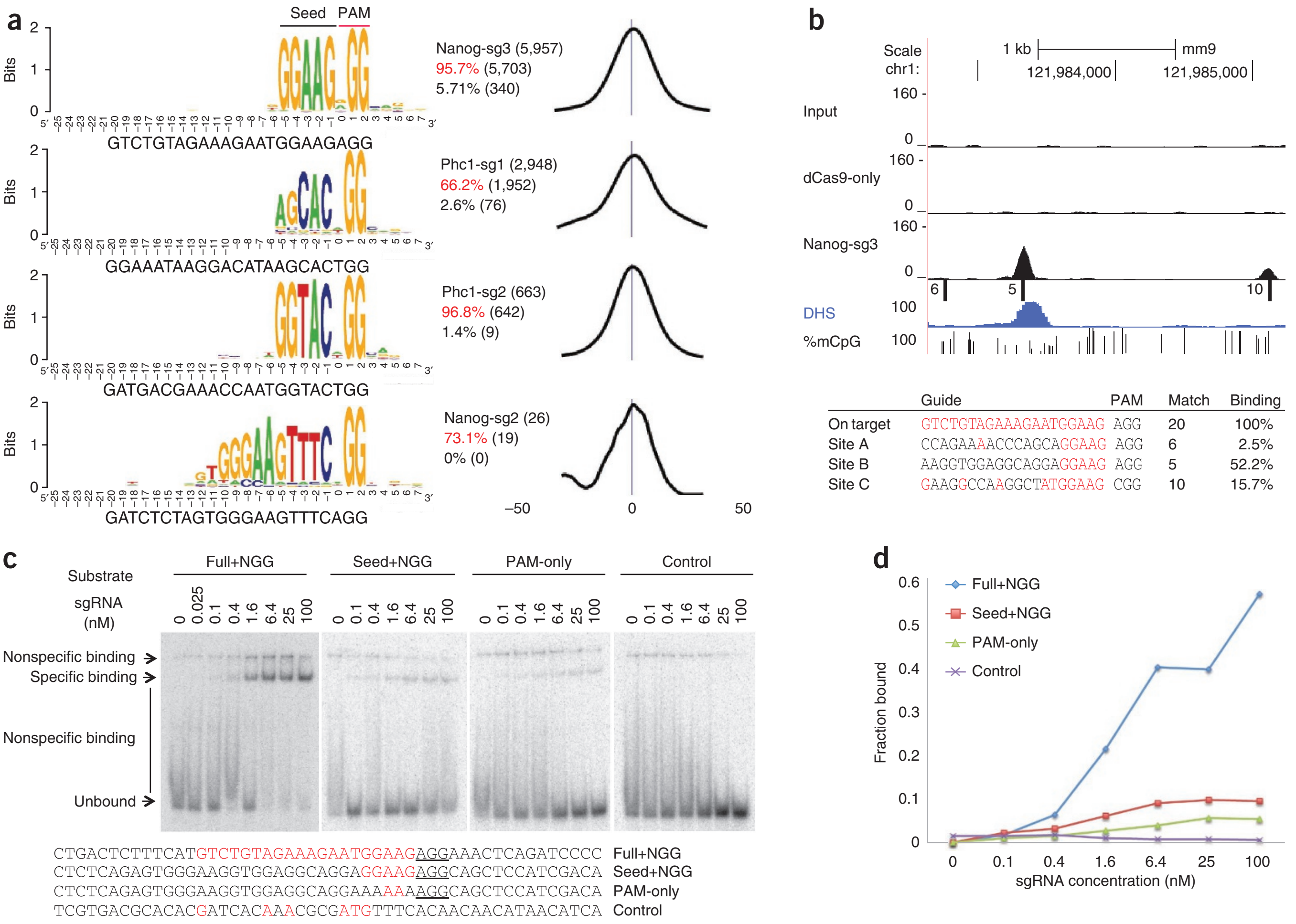
The peak motif analysis (**Supplementary Fig. 3**) revealed no enrichment of binding at seed sites followed by NAG, an alternative PAM previously reported to function in Cas9-mediated cleavage[9,16,27]. For example, of all 996 (33%) Phc1-sg1 ChIP peaks without seed+NGG sites, only 18 had seed+NAG within 50 bp of the peak summit, which is no more than expected by chance (**Supplementary Fig. 4**). ChIP-seq in human HEK293FT cells transfected with dCas9 and the same sgRNAs used in a previous study[9] in which NAG cleavage was reported, also did not detect binding at those NAG off-target sites (**Supplementary Fig. 5**). *In vitro* we observed a more than tenfold decrease in affinity when NGG was mutated to NAG in the on-target substrate (**Supplementary Fig. 6**, and see **Supplementary Table 2** for substrate sequences). Our *in vivo* and *in vitro* binding data are consistent with previous *in vitro* cleavage data showing that NAG or other variants rarely function as PAMs under enzyme-limiting conditions[27].

## Chromatin accessibility is a major determinant of binding

There are hundreds of thousands of seed+NGG sites in the genome for each sgRNA—for example, 621,651 for Nanog-sg3. To understand why only a small fraction of sites (<1%) were bound, we first looked for a correlation between the number of base matches to the 20-nucleotide guide region and the binding levels of ChIP peaks. Overall, the correlation was very weak (Pearson correlation coefficient $r = 0.03$, 0.12, 0.15 and 0.55 for Nanog-sg3, Phc1-sg2, Phc1-sg1 and Nanog-sg2, respectively (**Fig. 3a** and **Supplementary Fig. 7**).

We next applied a linear regression model of a set of sequence (mono- and di-nucleotide frequency), structural (melting temperature, DNA energy and flexibility[37]) and epigenetic (chromatin accessibility as assayed by DNase I hypersensitivity (DHS)[38] and DNA CpG methylation[39]) features around the seed+NGG sites for each sgRNA (Online Methods). We found that chromatin accessibility (DHS) is the strongest indicator of binding *in vivo*, explaining up to 19% of the variation in binding when considering all individual seed+NGG sites

© 2014 Nature America, Inc. All rights reserved.



| | Guide | PAM | Match | Binding |
|---|---|---|---|---|
| On target | GTCTGTAGAAAGAATGGAAG | AGG | 20 | 100% |
| Site A | CCAGAAAACCCAGCAGGAAG | AGG | 6 | 2.5% |
| Site B | AAGGTGGAGGCAGGAGGAAG | AGG | 5 | 52.2% |
| Site C | GAAGGCCAAGGCTATGGAAG | CGG | 10 | 15.7% |

```
CTGACTCTTTCATGTCTGTAGAAAGAATGGAAGAGGAAACTCAGATCCCC  Full+NGG
CTCTCAGAGTGGGAAGGTGGAGGCAGGAGGAAGAGGCAGCTCCATCGACA  Seed+NGG
CTCTCAGAGTGGGAAGGTGGAGGCAGGAAAAAAAGGCAGCTCCATCGACA  PAM-only
TCGTGACGCACACGATCACAAACGCGATGTTTCACAACAACATAACATCA  Control
```

**Figure 2** A 5-nucleotide seed for dCas9 binding. (**a**) Most peaks are associated with seed+NGG matches. The sequences with best match to the sgRNA followed by NGG within 50 bp of peak summits were aligned to generate the sequence logo using WebLogo[46]. The text to the right of the logos indicates the total number of peaks (top line), percentage and number of peaks with exact 5-nucleotide seed+NGG match within 50 bp of peak summits (middle line, in red), or when the 100-nucleotide sequences were shuffled while maintaining dinucleotide frequency (bottom line). The distribution of the exact seed+NGG match relative to the peak summit was shown on the right (the numbers indicate nucleotide positions). (**b**) Example of binding at seed+NGG–only sites. On the top are six tracks: input, dCas9-only immunoprecipitation and Nanog-sg3 immunoprecipitation read density, seed+NGG sites (position indicated by bars, named as A/B/C, and the numbers to the left indicates the number of matches to the guide), DNase I hypersensitivity read density (DHS) and percent of methylated alleles at CpG sites. Below are the target sequences, PAM, number of matches to the sgRNA and relative binding at each site. Guide-matched bases are in red. Genomic coordinates are based on UCSC mm9 genome. (**c**) Gel shift assay for 50-bp double-stranded DNA substrates with sequences matching the Nanog-sg3 on-target site ("Full+NGG") and a seed+NGG only off-target site ("Seed+NGG", site B in **Fig. 2b**). "PAM-only" is the "Seed+NGG" substrate with a mutated seed. The negative control substrate ("Control") was designed to contain no NGG or NAG. Complete substrate sequences are shown at the bottom, with PAM underlined and guide-matched bases in red. (**d**) The quantification of the gels in **c**. Shown is the percentage of the specific binding band relative to the entire lane at each sgRNA concentration.

in the genome (**Fig. 3b**). The difference in the number of seed+NGG sites in DHS peaks (i.e., accessible seed+NGG sites) explained 92% of the variation in the number of dCas9 peaks among the four sgRNAs (**Fig. 3c**, $n = 4$, $P < 0.05$, $F$-test). Although this is based on a limited set of sgRNAs, it suggests that it might be possible to predict the approximate number of off-target peaks based on the seed sequence in cell types where chromatin accessibility data are available.

Previous data suggested that Cas9 cleavage activity is not affected by DNA CpG methylation[9]. However, for the 17% of seed+NGG sites in the genome that contain CpG dinucleotides within the 20-mer guide match and NGG, CpG methylation became the strongest predictor of dCas9 binding and negatively correlated with binding (**Fig. 3d** and **Supplementary Fig. 8a**,**b**). In a regression model, adding CpG methylation to DHS for sites containing CpGs almost doubled the amount of variation explained (**Supplementary Fig. 8c**). Our data suggest that CpG methylation likely reflects an aspect of chromatin accessibility not fully captured by DHS or that, when combined with extensive mismatches, CpG methylation may impede binding.

The correlation with chromatin accessibility suggested that dCas9 off-target binding might preferentially occur at active genes. For Nanog-sg3, 70% of the off-target sites were associated with genes, including 18% in promoter regions (<2 kb upstream of the gene transcription start site), 6% near enhancer regions and 46% within genes (**Fig. 3e**). For example, an off-target peak that co-localized with the *Dusp19* gene transcription start site and a DHS peak showed 74% binding relative to the on-target site although it had only 7 base matches to Nanog-sg3 (**Fig. 3f**).

## Seed sequences influence sgRNA abundance and specificity

The Nanog-sg2 sgRNA had substantially fewer off-target binding sites than predicted by accessible seed+NGG sites (**Fig. 3c**). Although the

© 2014 Nature America, Inc. All rights reserved.

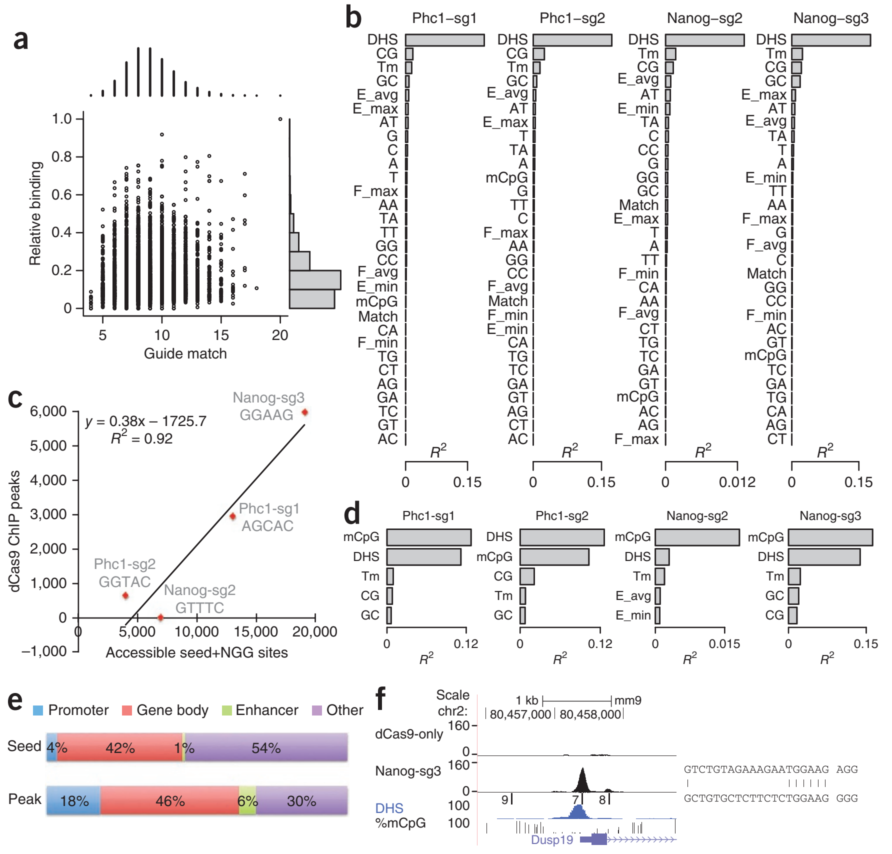


**Figure 3** Chromatin accessibility is a major determinant of binding *in vivo*. (**a**) Scatter (center) and histogram (top and right) plots of the number of matches to the sgRNA guide region (*x*-axis) and binding relative to the on-target site (*y*-axis) for all Nanog-sg3 peaks. Relative binding levels (0 to 1) are divided into ten equal bins and the number of peaks in each bin is shown on the right of the scatter plot. (**b**) Ranking of features based on $R^2$, the percent of variation in binding explained by each feature in a linear regression model (using *R*, one feature a time). DHS: DNase I hypersensitivity read density; Tm: melting temperature; match: number of bases that match the sgRNA; E(F)_min/max/avg: minimum, maximum and average tetranucleotide energy (flexibility) score within the guide+NGG region; A/C/G/T or their combination indicates mono- and di-nucleotide frequency in the guide+NGG region; mCpG: average fraction of methylated CpG in the guide+NGG region. (**c**) Scatter plot and linear regression between the number of dCas9 ChIP peaks and the number of accessible seed+NGG sites (i.e., sites overlapping DHS peaks). (**d**) Same as for **b**, but only plotting the top five features after regression was done using sites containing CpG dinucleotides. (**e**) Off-target peaks are preferentially associated with genes, for Nanog-sg3. Shown is the percentage of Nanog-sg3 seed+NGG sites (top) or ChIP peaks (bottom) that fall in each region category. (**f**) Example of off-target binding at the *Dusp19* promoter. Tracks are the same as **Figure 2b**. On the right is the alignment of the off-target site with seven matches (bottom) to the Nanog-sg3 guide sequence (top). Genomic coordinates are based on UCSC mm9 genome.

same amount of sgRNA plasmids were transfected, the abundance of Nanog-sg2 was more than sevenfold lower than the other three sgRNAs as determined by northern blot analysis (**Fig. 4a**). The same pattern of sgRNA abundance was observed when cells were transfected with sgRNA expression plasmids without co-transfecting dCas9, although all four sgRNAs showed substantially decreased levels of abundance, consistent with previous reports that Cas9 stabilizes sgRNA in cells[13].

To test if sgRNA abundance influences the number of off-target sites bound, we repeated the ChIP experiments after transfection with various amounts of sgRNA plasmids. Northern blot analysis confirmed the decrease in sgRNA when less plasmid was transfected (**Fig. 4a**), and we identified fewer peaks with less plasmid (**Fig. 4b**). When the level of Nanog-sg3 was reduced to a similar level as that of Nanog-sg2 (**Fig. 4a**, comparing lane 13 to lanes 16 and 17), the number of peaks for Nanog-sg3 was still much higher than for Nanog-sg2, presumably due to the presence of more accessible Nanog-sg3 seed+NGG sites in the genome (**Fig. 3c**). When 0.02 μg plasmid was transfected, Nanog-sg3 RNA was barely detected (lane 14); the 122 peaks identified in this library showed little overlap (9%) with our previous Nanog-sg3 ChIP, suggesting these were mostly non-specific signals (data not shown).

A comparison of the seed regions of the four sgRNAs suggested that UUU in the seed of Nanog-sg2 might be responsible for decreased sgRNA abundance and increased specificity, consistent with a recent observation that U in PAM-proximal positions 1–4 leads

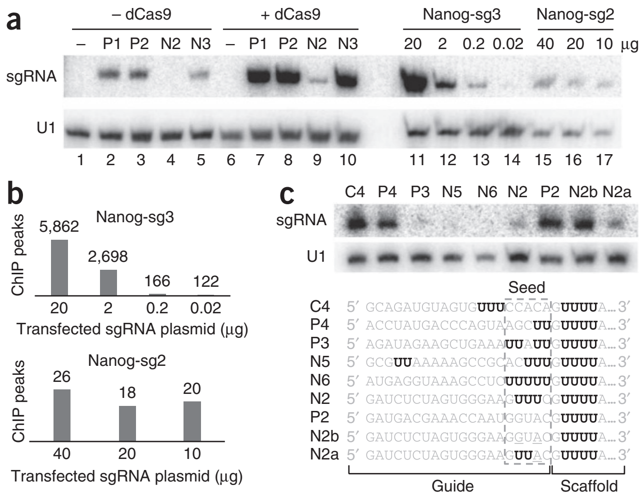


**Figure 4** Seed sequences influence sgRNA abundance and specificity. (**a**) Northern blot analysis showing the abundance of sgRNAs. Lanes 1–10: from cells transfected with dCas9 (lanes 6–10) or without dCas9 (lanes 1–5), and with either no sgRNA (lanes 1 and 6) or one of the four sgRNAs (P1: Phc1-sg1; P2: Phc1-sg2; N2: Nanog-sg2; N3: Nanog-sg3). Lanes 11–14: Nanog-sg3 abundance from dCas9-mESCs transfected with 20, 2, 0.2 or 0.02 μg Nanog-sg3 plasmid. Lanes 15–17: Nanog-sg2 abundance from dCas9-mESCs transfected with 40, 20 or 10 μg Nanog-sg2 plasmid. (**b**) The number of ChIP peaks detected from cells transfected with decreasing amount of sgRNA plasmids. (**c**) U-rich seed limits sgRNA abundance. Northern blot analysis from dCas9 cells transfected with the sgRNAs listed below. Consecutive U's are highlighted in bold black.

© 2014 Nature America, Inc. All rights reserved.

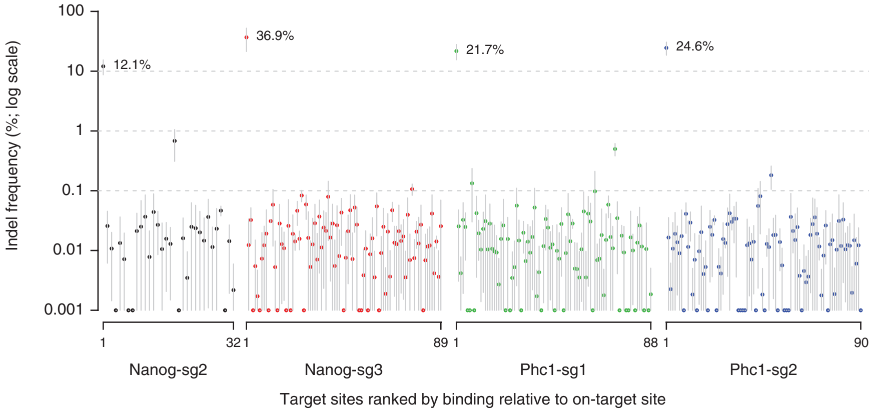


**Figure 5** Indel frequencies at on-target sites and 295 off-target sites. For each sgRNA, selected target sites (**Supplementary Table 3**) were ranked by decreasing ChIP binding relative to on-target site. Dots and gray bars indicate the mean and s.d. of indel frequency from three biological replicates, respectively. The *y*-axis was truncated at 0.001% for visualization at log scale. The indel frequencies for the four on-target sites are labeled with percentages.

to low gene-knockout efficacy[14]. Indeed, two mutations (U to G and U to A) in the Nanog-sg2 seed region that converted the seed (GUUUC) to the same sequence as the Phc1-sg2 seed (GGUAC), led to higher levels of sgRNA (sgRNA N2b in **Fig. 4c**). Considering the presence of GUUUUA adjacent to the seed and because sgRNAs are transcribed by RNA polymerase III, which is terminated by U-rich sequences[40,41], we speculate that, together with the downstream U-rich region, multiple U's in the seed might induce termination of sgRNA transcription. Consistent with this, three sgRNAs with seeds UUAUU, ACUUU and UUUUU also showed very low abundance (**Fig. 4c**, sgRNA P3, N5 and N6). When GUUUC was placed upstream of the seed (i.e., away from GUUUUA in the sgRNA), the sgRNA was well expressed (sgRNA C4 in **Fig. 4c**).

## One of 295 off-target sites is mutated above background

To test if dCas9 binding correlates with Cas9 nuclease–induced mutation, we examined the indel frequencies of the four on-target sites and 295 selected off-target sites by targeted PCR and sequencing[9]. These sites were selected to cover a broad range of binding levels and numbers of mismatches to the sgRNA: we ranked all peaks by binding (background-subtracted read counts) and, for each binding level, selected a peak with the fewest mismatches and another peak with most mismatches to the guide.

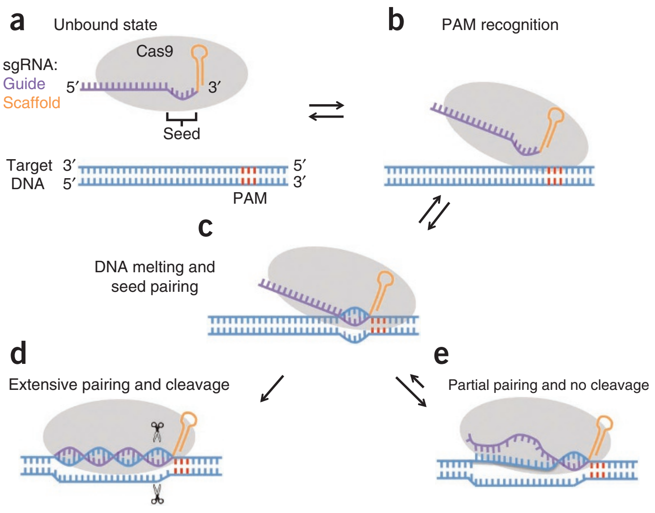

a
Unbound state
Cas9
sgRNA:
Guide
Scaffold
5′
3′
Seed
Target DNA
3′
5′
5′
3′
PAM
b
PAM recognition
c
DNA melting and seed pairing
d
Extensive pairing and cleavage
e
Partial pairing and no cleavage

**Figure 6** A model for Cas9 target binding and cleavage. (**a**) In the unbound state, Cas9 is loaded with sgRNA but not bound to DNA. The PAM region in the DNA is colored in red. (**b**) Recognition of the PAM by Cas9. (**c**) Cas9 melts the DNA target near the PAM to allow seed pairing. (**d**) If base pairing can be propagated to PAM-distal regions, the two Cas9 nuclease domains may be able to 'clamp' the target DNA and cleave it. (**e**) If only partial pairing occurs, there is no cleavage and Cas9 remains bound to the target.

We determined the indel frequency of the 299 selected binding sites in wild-type mESCs transfected with active Cas9 and each of the four sgRNAs, for three independent biological replicates (**Supplementary Table 3**). The level of Cas9 protein transiently expressed in the cells was 2.6-fold higher than in cells with stably integrated dCas9 used for ChIP (**Supplementary Fig. 9a**, comparing lane 1 to lane 8). The same ChIP and peak-calling procedures in cells transiently transfected with dCas9 identified 2.7 times more Nanog-sg3 peaks (16,119 versus 5,957 in dCas9 stable cell lines), including 96% (85) of the 89 peaks selected for indel analysis. The amount of Cas9 or dCas9 plasmids we used for transfection was similar to levels used for genome editing applications by others in the field (**Supplementary Fig. 9b**).

Using our previously validated model[9], the background indel frequencies due to sequencing errors were determined for each individual target using two biological replicates transfected with only Cas9 but no sgRNA (control). Importantly the control samples showed no evidence of targeted mutations by Cas9 (note that background indels in the absence of Cas9 might also occur). We manually reviewed sequencing alignments of all loci with indel frequencies >0.03%. We found that 12–37% of sequencing reads from the on-target sites contained indels. One off-target site, which was from Nanog-sg2, was mutated at a frequency of 0.7% (**Fig. 5**). There was no detectable correlation between binding and indel frequency (sites in **Fig. 5** are ranked by decreasing binding from left to right for each sgRNA). The selected sites include 7 of the top 10 (including all the top 6) and 36 of the top 50 Nanog-sg3 binding sites with the strongest ChIP signals, and 4 of the 8 Nanog-sg3 off-target binding sites that had fewer than four mismatches to the sgRNA; none of these off-target sites showed cleavage significantly above the background level.

## DISCUSSION

We have shown that dCas9 binding is more promiscuous than previously thought. The low binding specificity is explained by the limited requirement for an accessible match to a 5-nucleotide seed followed by an NGG PAM. The position of the seed region next to PAM was consistent with previous observations that base pairing near PAM is critical for targeting[6,8,16,33], but the seeds we identified for three of the four sgRNAs tested here are shorter than those previously reported; seed lengths of 8–13 nucleotides have been described as required for cleavage by Cas9 (refs. 6,8,16,33).

The seed sequence influences the specificity of Cas9-sgRNA binding in several ways. First, seed composition determines the frequency of a seed+NGG site in the genome. Second, seed composition determines the likelihood of a seed+NGG site occurring in open chromatin. Third, seed composition affects sgRNA abundance, probably at the level of transcription, and thus the effective concentration of

© 2014 Nature America, Inc. All rights reserved.

the Cas9-sgRNA complex. Lastly, seed composition may also affect loading into Cas9 and again tune the level of functional Cas9 (ref. 14). Through all four mechanisms, U-rich seeds are likely to increase target specificity.

Our results suggest that applications based on dCas9 or dCas9-effector fusions, such as transcription modulation, imaging and epigenome editing, could be complicated by substantial off-target binding. Previous studies suggest that several sgRNAs targeting the same gene are frequently necessary for gene activation[22–24]; this could potentially reduce off-target effects owing to the requirement of co-targeting. However, the use of multiple sgRNAs increases the number of potential off-target binding sites, which might complicate interpretation. Although we only detected indels at a low frequency (0.7%) above background for one off-target binding site among 295 selected sites, 295 is a small fraction of all possible binding sites and may not be representative of the complete off-target mutation profile of each sgRNA. This is an important consideration as low frequencies of indels could complicate certain CRISPR-Cas9 applications, such as genome-wide screening that involves selective growth[14,15]. Therefore, to minimize the likelihood of false-positive screening hits resulting from off-targeting, we recommend using multiple-guide RNAs to target each gene and the concordance among multiple guides to interpret screening results. We further note that although binding sites with NAG PAMs are not enriched in the ChIP data, a previous study has shown that NAG-flanked genomic loci can contribute to off-target indel mutations. Therefore, unbiased and more sensitive detection of genome-wide mutations will be needed to determine Cas9 cutting specificity.

The observation that most of the sites bound by Cas9 do not seem to be cleaved substantially is reminiscent of the eukaryotic Argonaute-microRNA system, in which most target mRNAs bearing partial microRNA matches are bound without cleavage and only a few targets with extensive pairing are cleaved[42]. We propose a two-state model (**Fig. 6**) similar to the Argonaute-microRNA system, in which pairing of a short seed region triggers binding after PAM recognition and subsequent DNA unwinding. In this model, targets with only seed complementarity remain bound by Cas9 without cleavage; only those with extensive pairing undergo efficient cleavage. This suggests a conformation change between binding and cleavage as observed for Argonaute-microRNA complexes[42,43]. While this paper was under review, a pair of Cas9 structural studies were published[44,45], including a crystal structure of dCas9 in complex with sgRNA and target DNA, which not only supports our observation of a PAM-proximal 5-nucleotide seed but also suggests a large conformation change during the inactive-active state transition[45].

## METHODS

Methods and any associated references are available in the online version of the paper.

**Accession codes.** GEO: GSE54745. SRA: SRP038774.

*Note: Any Supplementary Information and Source Data files are available in the online version of the paper.*

ACKNOWLEDGMENTS

We would like to thank J. Zamudio and T. Kelly for optimizing the ChIP protocol, and the entire Sharp lab for support and discussion. We also thank the Core Facility in the Swanson Biotechnology Center at the David H. Koch Institute for Integrative Cancer Research at MIT for their assistance with high-throughput sequencing. This work was supported by United States Public Health Service grants RO1-GM34277, R01-CA133404 from the National Institutes of Health, and PO1-CA42063 from the National Cancer Institute to P.A.S., and partially by Cancer Center Support (core) grant P30-CA14051 from the National Cancer Institute. F.Z. is supported by an US National Institutes of Health Director's Pioneer Award (1DP1-MH100706), the Keck, McKnight, Poitras, Merkin, Vallee, Damon Runyon, Searle Scholars, Klingenstein, and Simons Foundations, Bob Metcalfe, and Jane Pauley. X.W. is a Howard Hughes Medical Institute International Student Research Fellow. S.C. is a Damon Runyon Fellow (DRG-2117-12). P.D.H. is a James Mills Pierce Fellow. D.A.S. is an US National Science Foundation pre-doctoral fellow.

### AUTHOR CONTRIBUTIONS

X.W., F.Z. and P.A.S. designed experiments; X.W. and A.J.K. performed most experiments; D.A.S. performed targeted indel sequencing; A.W.C. and D.B.D. cloned the piggyBac dCas9 and sgRNA expressing vectors; A.C.C. generated the dCas9 stable cell line; P.D.H., A.E.T. and S.K. purified Cas9; P.D.H. contributed to *in vitro* binding assay; S.C. contributed to ChIP experiments with transient transfection. X.W., F.Z. and P.A.S. wrote the manuscript with help from all other authors. R.J., F.Z. and P.A.S. supervised the research.

### COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. van der Oost, J., Jore, M.M., Westra, E.R., Lundgren, M. & Brouns, S.J.J. CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends Biochem. Sci.* **34**, 401–407 (2009).
2. Deveau, H., Garneau, J.E. & Moineau, S. CRISPR/Cas system and its role in phage-bacteria interactions. *Annu. Rev. Microbiol.* **64**, 475–493 (2010).
3. Horvath, P. & Barrangou, R. CRISPR/Cas, the immune system of bacteria and archaea. *Science* **327**, 167–170 (2010).
4. Terns, M.P. & Terns, R.M. CRISPR-based adaptive immune systems. *Curr. Opin. Microbiol.* **14**, 321–327 (2011).
5. Marraffini, L.A. & Sontheimer, E.J. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat. Rev. Genet.* **11**, 181–190 (2010).
6. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
7. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
8. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
9. Hsu, P.D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31**, 827–832 (2013).
10. Mojica, F.J.M., Díez-Villaseñor, C., García-Martínez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**, 733–740 (2009).
11. Mali, P., Esvelt, K.M. & Church, G.M. Cas9 as a versatile tool for engineering biology. *Nat. Methods* **10**, 957–963 (2013).
12. Gasiunas, G. & Siksnys, V. RNA-dependent DNA endonuclease Cas9 of the CRISPR system: Holy Grail of genome editing? *Trends Microbiol.* **21**, 562–567 (2013).
13. Jinek, M. *et al.* RNA-programmed genome editing in human cells. *eLife* **2**, e00471 (2013).
14. Wang, T., Wei, J.J., Sabatini, D.M. & Lander, E.S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80–84 (2014).
15. Shalem, O. *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84–87 (2014).
16. Jiang, W., Bikard, D., Cox, D., Zhang, F. & Marraffini, L.A. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat. Biotechnol.* **31**, 233–239 (2013).
17. Wu, Y. *et al.* Correction of a Genetic Disease in Mouse via Use of CRISPR-Cas9. *Cell Stem Cell* **13**, 659–662 (2013).
18. Schwank, G. *et al.* Functional repair of CFTR by CRISPR/Cas9 in intestinal stem cell organoids of cystic fibrosis patients. *Cell Stem Cell* **13**, 653–658 (2013).
19. Dickinson, D.J., Ward, J.D., Reiner, D.J. & Goldstein, B. Engineering the *Caenorhabditis elegans* genome using Cas9-triggered homologous recombination. *Nat. Methods* **10**, 1028–1034 (2013).
20. Qi, L.S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173–1183 (2013).
21. Gilbert, L.A. *et al.* CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**, 442–451 (2013).
22. Cheng, A.W. *et al.* Multiplexed activation of endogenous genes by CRISPR-on, an RNA-guided transcriptional activator system. *Cell Res.* **23**, 1163–1171 (2013).
23. Perez-Pinera, P. *et al.* RNA-guided gene activation by CRISPR-Cas9-based transcription factors. *Nat. Methods* **10**, 973–976 (2013).
24. Mali, P. *et al.* CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat. Biotechnol.* **31**, 833–838 (2013).
25. Maeder, M.L. *et al.* CRISPR RNA-guided activation of endogenous human genes. *Nat. Methods* **10**, 977–979 (2013).

26. Chen, B. *et al.* Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell* **155**, 1479–1491 (2013).
27. Pattanayak, V. *et al.* High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat. Biotechnol.* **31**, 839–843 (2013).
28. Carroll, D. Staying on target with CRISPR-Cas. *Nat. Biotechnol.* **31**, 807–809 (2013).
29. Cradick, T.J., Fine, E.J., Antico, C.J. & Bao, G. CRISPR/Cas9 systems targeting β-globin and CCR5 genes have substantial off-target activity. *Nucleic Acids Res.* **41**, 9584–9592 (2013).
30. Chiu, H., Schwartz, H.T., Antoshechkin, I. & Sternberg, P.W. Transgene-free genome editing in *Caenorhabditis elegans* using CRISPR-Cas. *Genetics* **195**, 1167–1171 (2013).
31. Cho, S.W. *et al.* Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Res.* **24**, 132–141 (2014).
32. Fu, Y. *et al.* High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.* **31**, 822–826 (2013).
33. Sternberg, S.H., Redding, S., Jinek, M., Greene, E.C. & Doudna, J.A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507**, 62–67 (2014).
34. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
35. Teytelman, L., Thurtle, D.M., Rine, J. & van Oudenaarden, A. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proc. Natl. Acad. Sci. USA* **110**, 18602–18607 (2013).
36. Machanick, P. & Bailey, T.L. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27**, 1696–1697 (2011).
37. Packer, M.J., Dauncey, M.P. & Hunter, C.A. Sequence-dependent DNA structure: tetranucleotide conformational maps. *J. Mol. Biol.* **295**, 85–103 (2000).
38. Stamatoyannopoulos, J.A. *et al.* An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol.* **13**, 418 (2012).
39. Stadler, M.B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**, 490–495 (2011).
40. Orioli, A. *et al.* Widespread occurrence of non-canonical transcription termination by human RNA polymerase III. *Nucleic Acids Res.* **39**, 5499–5512 (2011).
41. Nielsen, S., Yuzenkova, Y. & Zenkin, N. Mechanism of eukaryotic RNA polymerase III transcription termination. *Science* **340**, 1577–1580 (2013).
42. Bartel, D.P. MicroRNAs: target recognition and regulatory functions. *Cell* **136**, 215–233 (2009).
43. Jinek, M. & Doudna, J.A. A three-dimensional view of the molecular machinery of RNA interference. *Nature* **457**, 405–412 (2009).
44. Jinek, M. *et al.* Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science* **343**, 1247997 (2014).
45. Nishimasu, H. *et al.* Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156**, 935–949 (2014).
46. Crooks, G.E., Hon, G., Chandonia, J.-M. & Brenner, S.E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).

© 2014 Nature America, Inc. All rights reserved.

npg

## ONLINE METHODS

**Oligonucleotides.** All oligonucleotides used in this study were purchased from Integrated DNA Technologies. Sequences are listed in **Supplementary Table 2**.

**Cloning.** A two-step fusion PCR was used to amplify Cas9 nickase open reading frame (ORF) from pX335 vector (Addgene: 42335) and incorporate H840A mutation to create a nuclease-deficient Cas9 (dCas9). This PCR product was inserted into the Gateway donor backbone pCR8/GW/TOPO to create pAC84 (Addgene: 48218). The dCas9 ORF in pAC84 was then transferred to a piggyBac-based destination vector pAC150 (PB-Lox-HygroR-Lox-4xHSInsulators-EF1a-DEST) by LR Clonase reaction (Invitrogen) to create pAC159 (PB-LHL-4xHS-EF1a-dCas9). The sgRNA expression cassette was amplified by PCR from pX335 vector and cloned into a piggyBac vector pAC158 (PB-neo-4xHSInsulators) to create pAC103 (PBneo-sgExpression). sgRNA was then cloned into BbsI-digested pAC103 by oligo cloning method, as described previously[8]. Cas9 transient transfection constructs consisted of CBh-driven WT-Cas9 or Cas9-D10AH840A (dCas9) containing a C-terminal HA-tag.

**Cell culture.** V6.5 (mESCs were cultured in DMEM supplemented with 15% FBS, penicillin and streptomycin, L-glutamine, nonessential amino acids and leukemia inhibitory factor (LIF). For generation of cells stably integrating dCas9, cells were transfected in a 6-well plate and selected using Hygromycin B at 100 µg/ml 24 h after transfection, which was increased to 150 µg/ml 48 h after transfection. Cells were split onto 10-cm plates and single clones were isolated, expanded and used for all experiments described. HEK293FT cells were cultured as previously described[9]. All transfections were done with Lipofectamine 2000 (Invitrogen).

**ChIP.** Three million cells were seeded on to 10-cm plates on day 1, transfected with sgRNAs plasmids (or together with HA-dCas9 plasmids) on day 2, transferred to 15-cm plates on day 3; and cross-linking was done on day 4 with ~50 million cells. Cross-linking was done by adding 2 ml (i.e., 0.1 of the volume of the cell media) 37% formaldehyde to the plate, incubating at room temperature for 15 min, and quenched by adding 1 ml 2.5 M glycine. Cells were rinsed twice with cold PBS and scraped to collect in cold PBS. Cells were centrifuged at 1,350*g* for 5 min at 4 °C and washed again in cold PBS. Cells were flash frozen in a dry ice/ethanol mix and stored at −80 °C. The cell pellet was resuspended in 5 ml cold lysis buffer 1 (50 mM HEPES-KOH pH 7.5, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NP-40, 0.25% Triton X-100, 1× Roche complete protease inhibitors), rotated at 4 °C for 10 min followed by centrifugation at 1,350*g* for 5 min at 4 °C. The pellet was resuspended in 5 ml lysis buffer 2 (10 mM Tris-Cl pH 8, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 1× Roche complete protease inhibitors), rotated at 4 °C for 10 min followed by centrifugation at 1,350*g* for 5 min at 4 °C. The nuclear pellet was resuspended in 2 ml sonication buffer (20 mM Tris-Cl pH 8, 150 mM NaCl, 2 mM EDTA, 0.1% SDS, 1% Triton X-100, 1× Roche complete protease inhibitors) and sonicated (60 min total time, 30 s on, 30 s off, in 6 rounds of 10 min) in a Bioruptor (Diagenode). The lysate was centrifuged in Eppendorf tubes in a microfuge at 4 °C at maximum speed for 20 min. Supernatant was collected, and 50 µl of this was saved as input. Protein G Dynabeads were conjugated to 5 µg rabbit anti-rat antibody (Thermo) in 0.1 M Na-Phosphate pH 8 buffer at 4 °C with rotation followed by conjugation to 5 µg HA antibody (Roche 3F10, #11867431001). Beads were resuspended in 50 µl sonication buffer and added to samples to immunoprecipitate overnight. The next day, beads were washed twice in sonication buffer, once in sonication supplemented with 500 mM NaCl, once in LiCl buffer (10 mM Tris-Cl pH 8, 250 mM LiCl, 1 mM EDTA, 1% NP-40) and once in TE + 50 mM NaCl. Each wash was accomplished with rotation at 4 °C for 5 min. Chromatin was eluted at 65 °C for 15 min in elution buffer (50 mM Tris-Cl pH 8, 10 mM EDTA, 1% SDS). Input was combined with elution buffer and both input and immunoprecipitation cross-links were reversed at 65 °C overnight. RNA was digested with RNase A at 0.2 mg/ml final concentration (Sigma) at 37 °C for 2 h and protein was digested with proteinase K at 0.2 mg/ml final concentration (Life Technologies) at 55 °C for 45 min. DNA was extracted with phenol:chloroform:isoamyl alcohol (Life Technologies) and precipitated with ethanol. Barcoded libraries were prepared and sequenced on Illumina HiSeq2000.

**ChIP-seq data analysis.** Reads were de-multiplexed and mapped to mouse genome mm9 using bowtie[47], requiring unique mapping with at most two mismatches (-n 2 -m 1–best–strata). Mapped reads were collapsed and the same number of reads (about 9 million) was randomly sampled from each library to match sequencing depth. Peaks were called using MACS[34] with default settings. For each sample, the other samples are each used as a control and only peaks called over all five controls are defined as target sites. To quantify relative binding strength, reads were first extended at the 3′ end to the average fragment length (*d*) estimated by MACS, and then the number of fragments (extended reads) overlapping with the seed+NGG region is counted and normalized by subtracting counts from dCas9-only control. If multiple seed+NGG match sites were found, the one with the highest relative binding was assigned to the peak.

**Analysis on determinants of binding.** mESC DNase Hypersensitivity data (bigwig file and narrow peak file) were downloaded from UCSC genome browser hosting the mouse ENCODE project[38]. DNA CpG methylation data were downloaded from GEO data set GSE30202. Melting temperature (Tm) was calculated using the *oligotm* program in *primer3* version 2.3.6. DNA stability and flexibility were calculated using a table of tetranucleotide scores derived from X-ray crystal structures in a previous study[37]. The linear regression was done by using the *lm* function in *R*, one feature a time to calculate the $R^2$ value for each feature.

**Northern blot analysis.** Total RNA was isolated using TRIzol (Life Technologies) and 5 µg of total RNA was loaded on 8% denaturing PAGE. Northern blot analysis was done as previously described[9], using a probe targeting the scaffold shared by all sgRNAs.

**Protein purification.** Human codon-optimized Cas9 (Addgene plasmid 42230) was subcloned into a custom pET-based expression vector with an N-terminal hexahistidine (6xHis) tag followed by a SUMO protease cleavage site. The fusion construct was used to transform *Escherichia coli* Rosetta 2(DE3) competent cells (Millipore), which were then grown in LB media to $OD_{600}$ 0.6, and induced with 0.2 mM IPTG for 16 h at room temperature. Cells were pelleted, resuspended and washed with Milli-Q H2O supplemented with 0.2 mM PMSF, and lysed with lysis buffer (20 mM Trizma base, 500 mM NaCl, 0.1% NP-40, 2 mM DTT, 10 mM imidazole). The lysis buffer was supplemented with protease inhibitor cocktail (Roche) immediately before use. Whole lysate was sonicated at 40% amplitude (Biologics Inc., 2s on, 4s off) before ultracentrifugation (30,000 r.p.m. for 45 min). The clarified lysate was applied to cOmplete His-tag purification columns (Roche), washed with wash buffer 1 (20 mM Trizma base, 500 mM NaCl, 0.1% NP-40, 2 mM DTT, 10% glycerol, 10 mM imidazole) and wash buffer 2 (20 mM Trizma base, 250 mM NaCl, 0.1% NP-40, 2 mM DTT, 10% glycerol, 50 mM imidazole). The 6xHis affinity tag was released by SUMO protease cleavage and bound protein was eluted with a linear gradient of 150 mM–500 mM imidazole. Eluted protein was concentrated with Amicon centrifugal filter units with Ultracel membrane (Millipore) and stored at −80 °C.

***In vitro* transcription.** A T7 promoter forward oligo was annealed to an sgRNA template oligo by heating to 95 °C for 3 min in 1× T4 DNA ligase buffer and then cooled at room temperature for 30 min. The annealed product was used as a template and transcribed with MEGAshortscript T7 Kit (Life Technologies). RNAs were purified by MEGAclear Kit (Life Technologies) and frozen at −80 °C.

**Gel shift assay.** Single-stranded DNA oligos of 50 nucleotides were purchased from IDT and PAGE purified. Double-stranded substrates were generated by mixing 100 pmol each strand in water (10 µl total), heating to 95 °C for 3 min and cooled to room temperature. The substrates were then 5′ end labeled with [γ-32P]-ATP using T4 PNK (New England Biolabs) for 30 min at 37 °C, and free ATP removed by G-25 column (GE Healthcare). For each reaction, 100 nM Cas9 was mixed with a 1:4 dilution series of sgRNA (from 0 to 100 nM) in 1× NEBuffer 3 at 37 °C for 10 min, and then about 0.5 nM labeled substrate oligos were added and incubated for 5 min at 37 °C in a 10 µl reaction. Reactions were stopped on ice and 1/2 volume of 50% glycerol added.

© 2014 Nature America, Inc. All rights reserved.

doi:10.1038/nbt.2889

Samples were loaded on to 12% native PAGE and run at 300 V for 2 h at room temperature. Gels were visualized by phosphorimaging. Gel quantification is done with ImageJ. The fraction bound shown in **Figure 2c** was calculated as the ratio of intensity from the specific binding band to the total intensity of the entire lane.

**Targeted sequencing and indel detection.** For biological replicate 1, cells were seeded in 6-well plates (300,000 cells per well), transfected with 2 μg sgRNA plasmid, 2 μg Cas9 plasmid, using 10 μl Lipofectamine 2000 reagent per sample for 3 h. For replicate 2 and 3, 50% more plasmids were used. DNA was extracted, and selected target sites were PCR amplified, normalized and pooled in equimolar proportions. Pooled libraries were denatured, diluted to a 14-pM concentration and sequenced using the Illumina MiSeq Personal Sequencer (Illumina). Sequencing data were demultiplexed using paired barcodes, mapped to reference amplicons and analyzed for indels, as described previously[9].

47. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).

© 2014 Nature America, Inc. All rights reserved.

doi:10.1038/nbt.2889