Mapping the genomic landscape of CRISPR–Cas9 cleavage

Peter Cameron[1,4], Chris K Fuller[1,4], Paul D Donohoue[1], Brittnee N Jones[1,3], Matthew S Thompson[1], Matthew M Carter[1], Scott Gradia[1], Bastien Vidal[1], Elizabeth Garner[1], Euan M Slorach[1], Elaine Lau[1], Lynda M Banh[1], Alexandra M Lied[1], Leslie S Edwards[1], Alexander H Settle[1], Daniel Capurso[1], Victor Llaca[2], Stéphane Deschamps[2], Mark Cigan[2,3], Joshua K Young[2] & Andrew P May[1,3]

RNA-guided CRISPR–Cas9 endonucleases are widely used for genome engineering, but our understanding of Cas9 specificity remains incomplete. Here, we developed a biochemical method (SITE-Seq), using Cas9 programmed with single-guide RNAs (sgRNAs), to identify the sequence of cut sites within genomic DNA. Cells edited with the same Cas9–sgRNA complexes are then assayed for mutations at each cut site using amplicon sequencing. We used SITE-Seq to examine Cas9 specificity with sgRNAs targeting the human genome. The number of sites identified depended on sgRNA sequence and nuclease concentration. Sites identified at lower concentrations showed a higher propensity for off-target mutations in cells. The list of off-target sites showing activity in cells was influenced by sgRNP delivery, cell type and duration of exposure to the nuclease. Collectively, our results underscore the utility of combining comprehensive biochemical identification of off-target sites with independent cell-based measurements of activity at those sites when assessing nuclease activity and specificity.

The RNA-guided endonuclease Cas9, derived from the CRISPR–Cas microbial immune system, has emerged as a potent genome engineering tool[1–3]. Ribonucleoprotein (RNP) complexes formed between a single guide RNA (sgRNA) and Cas9 (sgRNP) recognize DNA through sequence-specific interactions with two DNA regions, the protospacer and the protospacer adjacent motif (PAM)[4,5]. Protospacers are bound via base pairing between target DNA and a 20-nucleotide complementary sequence at the 5′ end of the sgRNA, the spacer; while PAM binding is facilitated by direct interactions between Cas9 amino acid residues in the C-terminal PAM interacting domain and the target DNA[6,7]. Cas9 tolerates both protospacer and PAM recognition mismatches that may result in off-target nuclease activity[8–16].

Cellular repair of double-strand breaks (DSBs) may result in mutagenic insertions or deletions (indels), or even in larger chromosomal rearrangements[11–14]. Therefore, genome-wide methods to detect off-target cleavage by sgRNPs are essential, especially as CRISPR–Cas9 is used for a variety of biotechnology applications. For example, in human therapeutic applications of genome editing, a map of possible off-target cleavage events could aid in avoiding promiscuous guides that may result in poor clinical outcomes. For applications such as crop improvement, a means to track off-target mutations could assist in mutation removal by segregation during subsequent crosses.

To detect off-target nucleolytic activity, multiple experimental methods have recently been developed, including several that are genome wide[11–18]. However, each of these techniques has limitations. Methods that use synthetic variant libraries are intrinsically biased by library design[15,16]. Genome-wide techniques such as GUIDE-seq and HTGTS rely on cellular events such as the integration of donor sequences or chromosomal translocations. These methods may be confounded by site- and cell-line-dependent differences in DNA repair as well as by interactions between editing events and cell division[11,12]. Digenome-seq aims to detect Cas9 cut sites by sequencing genomic DNA but requires high read depth to do so[13,14]. As a result, Digenome-seq may lack the sensitivity to detect the full spectrum of possible Cas9–sgRNA cut sites and is not suitable for screening large numbers of guide RNAs.

To resolve these issues, we have developed a biochemical method that uses the selective enrichment and identification of tagged genomic DNA ends by sequencing (SITE-Seq) to identify Cas9 cleavage sites in purified genomic DNA. Since SITE-Seq is a biochemical assay, genomic DNA can be digested with a range of sgRNP concentrations, from limiting to saturating, thus permitting the recovery of both high- and low-cleavage-sensitivity off-target sites. This can then be used to guide careful and comprehensive examination of possible off-target sites in cells, measuring both mutation frequency and functional cellular consequence. Finally, SITE-Seq produces sequencing libraries that are highly enriched for sgRNP cleavage fragments, enabling specificity

[1]Caribou Biosciences, Berkeley, California, USA. [2]DuPont Pioneer, Johnston, Iowa, USA. [3]Present addresses: Omicia, Inc., Oakland, California, USA (B.N.J.); Genus Research, DeForest, Wisconsin, USA (M.C.) and Chan Zuckerberg Biohub, San Francisco, California, USA (A.P.M.). [4]These authors contributed equally to this work. Correspondence should be addressed to J.K.Y. (josh.young@pioneer.com) or A.P.M. (apmay1@gmail.com).

RECEIVED 15 MARCH; ACCEPTED 6 APRIL; PUBLISHED ONLINE 1 MAY 2017; DOI:10.1038/NMETH.4284

profiling with minimal read depth, which is critical for SITE-Seq's implementation as a high-throughput guide selection tool.

We used SITE-Seq to profile the landscape of biochemical cleavage for a panel of sgRNPs, and then we examined whether the cut sites identified were modified in cells (hereafter referred to as cellular off-targets).

RESULTS

SITE-Seq workflow

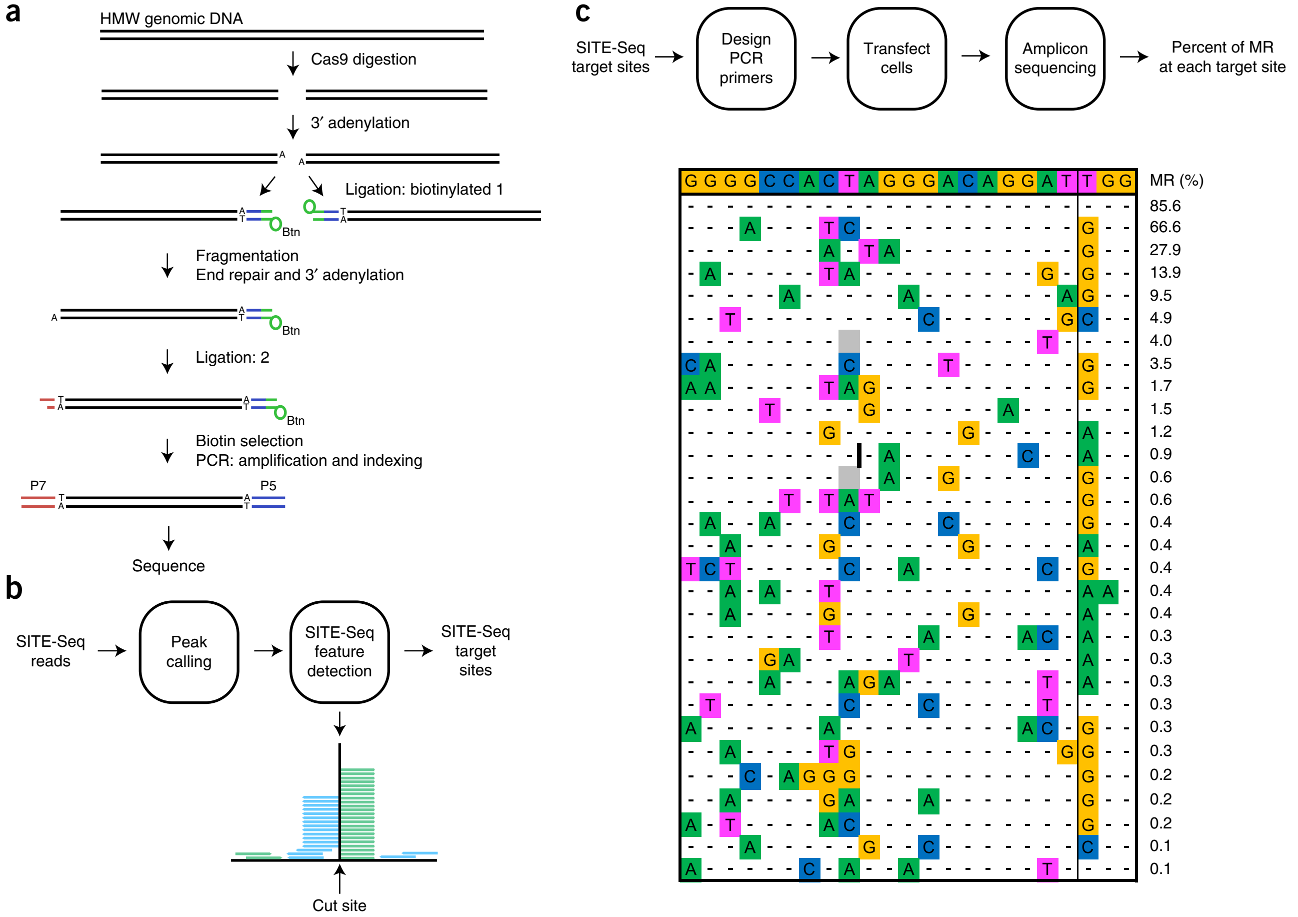
SITE-Seq is a multistep biochemical procedure whereby cleaved genomic DNA is selectively tagged, enriched, sequenced, and then mapped to a corresponding reference genome (**Fig. 1a** and **Supplementary Protocol**). In the resulting alignment, sites cleaved by the sgRNP yield sequence read pileups that terminate at the cut site, ~3 nucleotide proximal to the PAM, producing a distinct signature that can be detected computationally (**Fig. 1b**). It should be noted that this signature is similar to that observed with Digenome-seq[13,14].

The SITE-Seq method was initially tested on human embryonic kidney (HEK293) gDNA digested with Cas9 sgRNP at 64 nM targeting the *AAVS1* locus[19]. From this library, we found a total of 771 biochemical cleavage sites (i.e., SITE-Seq target sites). Most sites displayed similarity to the on-target sequence and occupied a position that placed the NGG PAM three nucleotides 3′ of the cleavage site, as expected (**Supplementary Table 1**). We next examined SITE-Seq target sites in a cellular environment by transfecting *AAVS1* sgRNA into HEK293 cells stably expressing Cas9–GFP (HEK293-Cas9–GFP) and measuring off-target editing 3 d later (**Fig. 1c**). Of the 771 SITE-Seq target sites, a subset of 68 sites–containing a wide range of nucleotide substitutions relative to the on-target sequence–was selected for evaluation.

Overall, we found 29 cellular off-targets with mutation frequencies ranging from 66.6% to 0.1% (compared with the on-target, which yielded 85.6%) (**Fig. 1c**). Most sgRNA-to-protospacer mismatches were nucleotide to nucleotide, although two sgRNA bulges and one target DNA bulge were also observed. Taken together, these data suggest that SITE-Seq enables us to uncover the full landscape of sgRNP biochemical cleavage sites and then pinpoint the subset of those targets that accumulate off-target mutations in a cellular context.

Concentration-dependent recovery of SITE-Seq target sites

We designed SITE-Seq with the underlying assumption that sgRNP concentration could be modulated to recover genomic



```
G G G G C C A C T A G G G A C A G G A T | T G G   MR (%)
- - - - - - - - - - - - - - - - - - - - | - - -   85.6
- - - A - - - T C - - - - - - - - - - - | G - -   66.6
- - - - - - - A - T A - - - - - - - - - | G - -   27.9
- A - - - - - T A - - - - - - - - - G - | G - -   13.9
- - - - - A - - - - - A - - - - - - - A | G - -   9.5
- - T - - - - - - - - - C - - - - - - G | C - -   4.9
- - - - - - - -   - - - - - - - - - T - | - - -   4.0
C A - - - - - - C - - - - T - - - - - - | G - -   3.5
A A - - - - - T A G - - - - - - - - - - | G - -   1.7
- - - - T - - - - G - - - - - - A - - - | - - -   1.5
- - - - - - - G - - - - - - G - - - - - | A - -   1.2
- - - - - - - - - | - A - - - - - - C - - | A - -   0.9
- - - - - - - -   - A - - G - - - - - - | G - -   0.6
- - - - - T - T A T - - - - - - - - - - | G - -   0.6
- A - - A - - - C - - - - C - - - - - - | G - -   0.4
- - A - - - - G - - - - - - G - - - - - | A - -   0.4
T C T - - - - - C - - A - - - - - - C - | G - -   0.4
- - A - A - - T - - - - - - - - - - - - | A A -   0.4
- - A - - - - G - - - - - - G - - - - - | A - -   0.4
- - - - - - - T - - - - A - - - - A C - | A - -   0.3
- - - - G A - - - - - T - - - - - - - - | A - -   0.3
- - - - A - - - A G A - - - - - - - T - | A - -   0.3
- T - - - - - - C - - - C - - - - - T - | - - -   0.3
A - - - - - - A - - - - - - - - - A C - | G - -   0.3
- - A - - - - T G - - - - - - - - - - G | G - -   0.3
- - - C - A G G G - - - - - - - - - - - | G - -   0.2
- - A - - - - G A - - - A - - - - - - - | G - -   0.2
A - T - - - - A C - - - - - - - - - - - | G - -   0.2
- - - A - - - - - G - - C - - - - - - - | C - -   0.1
A - - - - - C - A - - A - - - - - - T - | - - -   0.1
```

Figure 1 | SITE-Seq workflow. (**a**) Schematic depicting SITE-Seq workflow, starting with high molecular weight (HMW) genomic DNA and ending with indexed SITE-Seq libraries prepared for sequencing. Adapter 1 contains an Illumina-compatible sequence (blue) and an outer unique SITE-Seq recovery sequence (green). Adapter 2 (red) contains an Illumina-compatible sequence. (**b**) Schematic depicting SITE-Seq analytical pipeline. SITE-Seq features are detected in sequencing libraries to recover SITE-Seq target sites. (**c**) Schematic depicting cell-based validation pipeline (top). SITE-Seq targets found at 64 nM sgRNP (targeting the *AAVS1* gene), the percent of mutant reads (MR) are shown to the right of each site. The on-target sequence is shown at the top of the alignment, mismatched nucleotides are highlighted in color, putative sgRNA bulges (gray) and target DNA bulges (black line) are shown.

target sites with both higher and lower cleavage sensitivities. To examine this, we performed SITE-Seq using eight distinct sgRNAs across a range of sgRNP concentrations (0.25–1,024 nM). Target sites were selected because they had been investigated previously with other off-target analysis methods (*VEGFA*, *FANCF*)[11–14], or they were shown to be active in cell-based experiments (data not shown). Across all eight sgRNPs tested, the number of SITE-Seq target sites recovered increased as a function of nuclease concentration, with a small number of sites being recovered at the lowest concentration and hundreds to thousands of sites being recovered at the highest concentration (**Fig. 2** and **Supplementary Tables 2–9**). Of note, the on-target site was recovered with the lowest concentration of nuclease that permitted recovery of target sites, most often at 0.25 nM sgRNP (**Supplementary Tables 2–9**). We also found a *VEGFA* off-target sequence that was repeated thousands of times throughout the genome (**Supplementary Table 10**). Our pipeline segregated sites with this sequence (and a subset of sites closely resembling this sequence), and these sites did not contribute to the main analyses (**Fig. 2**).

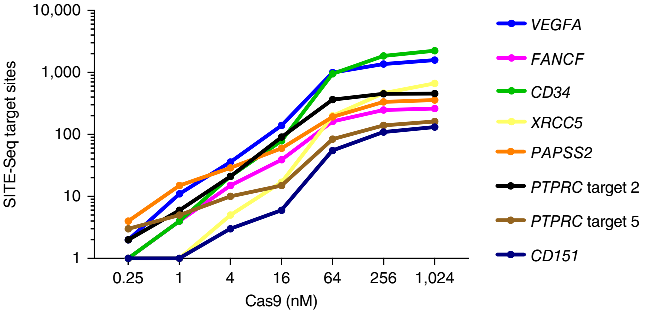


Figure 2 | SITE-Seq target sites recovered with 0.25–1,024 nM Cas9. For all guides, the number of SITE-Seq target sites recovered as a function of sgRNP concentration are shown; HEK293 cells were the source for genomic DNA (*n* = 1 for each data point).

Next, we examined how sgRNP concentration affects specificity by constructing sequence logos from the subset of SITE-Seq target sites recovered at each sgRNP concentration (**Supplementary Fig. 1**). Positional specificity as a function of sgRNP concentration appeared to diverge in a protospacer-sequence-dependent manner. For some guides (*FANCF* and *PTPRC* target 2), specificity at the 3′ end of the protospacer was preserved at all concentrations, consistent with the 'seed sequence' model of Cas9 guide RNA specificity[4]; whereas for other guides (i.e., *CD34*, *PAPSS2*) this trend was less clear. These data are consistent with those reported by the GUIDE-seq method[11], where the positions of sequence mismatches found in off-target sites were highly variant within and across guides and located throughout the length of the protospacer.

Biochemical validation of SITE-Seq target sites

Next, we provided further evidence that SITE-Seq was recovering genuine sgRNP off-targets by performing an independent biochemical cleavage assay. We selected 101 *VEGFA* SITE-Seq target sites demonstrating a wide range of cleavage sensitivities; and then, using amplicons as substrates, we performed biochemical digestion experiments with increasing concentrations of sgRNP (up to 1,024 nM). Nearly all off-target sites recovered with lower sgRNP concentrations in SITE-Seq were digested in our biochemical cleavage assay (5/5 sites recovered with ≥1 nM, 14/14 sites recovered with ≥4 nM, and 38/39 sites recovered with ≥ 16 nM). Sites recovered only with higher concentrations of sgRNP in SITE-Seq were less likely to show cleavage, although the majority were nonetheless digested (28/36 recovered with ≥64 nM, 2/3 recovered with ≥256 nM, and 0/3 recovered with 1,024 nM) (**Supplementary Table 11**). Overall, SITE-Seq results were predictive of the biochemical activity observed with our amplicon cleavage assay.

Cell-based validation of SITE-Seq target sites

Editing activity at target sites identified using SITE-Seq was then measured in cells by transfecting sgRNAs into HEK293-Cas9–GFP cells and quantifying indel formation at the Cas9 cleavage site 3 d later using targeted amplicon sequencing (see Online Methods for more details). For each sgRNA, we selected ~100–400 sites showing a range of sgRNP biochemical cleavage sensitivities for evaluation (**Supplementary Tables 12–19**). To generate our validation test set, sites were grouped according to the lowest concentration of sgRNP that enabled their recovery, then randomly selected from each cohort. In this manner, we sampled from sites showing a range of biochemical cleavage sensitivities, including those sites recovered at most of the concentrations of sgRNP tested (i.e., sites with high cleavage sensitivity) and sites only identified with higher concentrations of sgRNP. Our test set also included larger fractions of the high cleavage sensitivity sites, since these were comparatively rarer.

Across sgRNAs, only a subset of SITE-Seq target sites were confirmed as cellular off-targets (**Fig. 3**), which was potentially on account of limits in sgRNP concentration levels achieved with our sgRNP delivery protocol. Most off-target mismatches were nucleotide to nucleotide, although we also observed bulges in the sgRNA relative to the target DNA or bulges in the target DNA (**Fig. 3a**). sgRNA bulges appeared common with certain guides (*XRCC5*) and rare with others (*VEGFA*), and we found sequences with either two or three classes of mismatches. Importantly, we found that two guides (*CD151* and *PTPRC* target 2) showed high specificity and presented at most one cellular off-target after screening most SITE-Seq target sites.

SITE-Seq libraries generated from gDNA digested at lower concentrations of sgRNP contained fewer sites, but a large percentage were confirmed as cellular off-targets. In contrast, SITE-Seq libraries made from gDNA digested at higher concentrations of sgRNP contained many sites, yet most did not show off-target editing in cells (**Fig. 3b** and **Supplementary Fig. 2**). Importantly, all cellular off-targets in our validation test set were recovered when gDNA was digested at higher nuclease concentrations. Indeed, 62/63 cellular off-targets in our validation test set were recovered when gDNA was digested at 64 nM sgRNP. Overall, this suggests that (i) biochemical cleavage sensitivity is a strong predictor of cellular off-target editing, and (ii) SITE-Seq with higher concentrations of sgRNP may recover all relevant cellular off-targets, at least for certain delivery strategies.

Off-target editing is increased with extended sgRNP treatment

In developing SITE-Seq, we hypothesized that the subset of biochemical off-targets that was prone to editing in cells (as well as the frequencies of indels) would be dependent on cellular context, such as sgRNP delivery conditions. Accordingly, we next tested whether extending the duration of sgRNP expression could increase off-target editing. We selected three sgRNAs (*AAVS1*, *VEGFA* and *FANCF*) and delivered sgRNP either transiently (as a preassembled complex or as sgRNA transfected into HEK293-Cas9–GFP cells) or stably (by generating polyclonal cell lines). SITE-Seq target sites were examined for editing 3 d after transient transfection and between 6 to 28 d after stable expression (**Fig. 4** and **Supplementary Tables 20–22**).

Across guides, direct delivery of the preassembled sgRNP produced the least off-target activity, consistent with previous reports[20,21], while prolonged expression produced a time-dependent increase in off-target editing, albeit for a subset of sites. For *AAVS1*

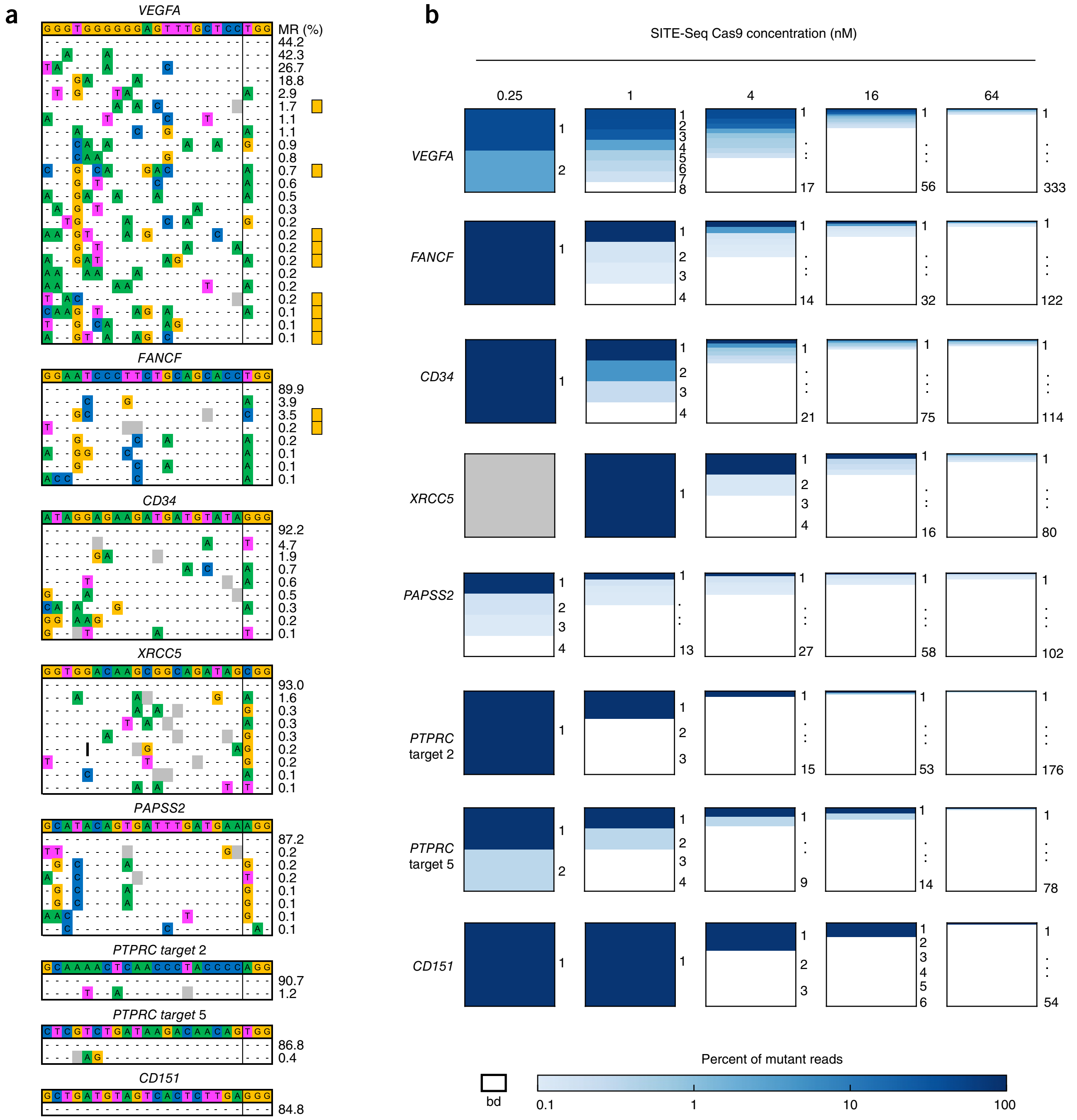


Figure 3 | SITE-Seq predicts cell-based activity. (**a**) SITE-Seq target sites that accumulated >0.1% cellular off-target mutations are shown. The percent of mutant reads is shown to the right of each site. The on-target sequence is shown at the top of each alignment, directly underneath the target site gene name. Mismatched nucleotides are highlighted in color; putative sgRNA bulges (gray) and target DNA bulges (black line) are shown. For *VEGFA* and *FANCF*, cellular off-targets not recovered with other experimental methods are marked with an orange square. (**b**) Heatmap showing relationship between biochemical cleavage sensitivity and cell-based editing. For each sgRNA, SITE-Seq target sites tested in cells are shown, and sites recovered at a given sgRNP concentration are grouped together in boxes (the number of sites is listed to the right of each box). For a given box, if fewer sites are present, the total area for each site is greater. On-target sites are always site number 1; and gray boxes signify that no SITE-Seq target sites were recovered. The color-intensity scales with fraction of mutant reads up to 100%; and only sites with >0.1% indels are colored; bd, below detection.

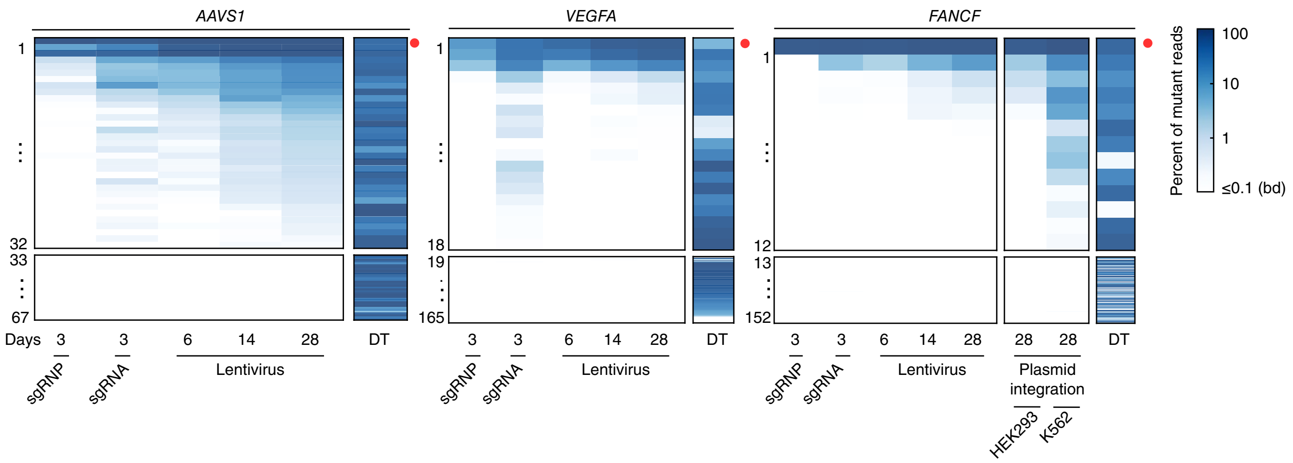


Figure 4 | Off-target editing varies with sgRNP delivery method, duration of treatment and cell type. sgRNP was delivered transiently by either first preassembling biochemically then transfecting into HEK293 cells ('sgRNP') or by transfecting sgRNA into cells stably expressing Cas9 ('sgRNA'). sgRNP was delivered stably by generating polyclonal cell lines expressing both Cas9 and sgRNA by either transducing HEK293-Cas9–GFP cells with lentivirus directing expression of sgRNA ('Lentivirus') or transfecting and integrating linearized plasmid DNA expressing both Cas9 and sgRNA ('plasmid integration'). Indel frequency as a function of sgRNP delivery method, duration of treatment and cell type is shown by heatmap. The color intensity scales with fraction of mutant reads up to 100%; bd, below detection. Sites with >0.1% indels across any of the conditions are shown in the top box. Sites are numbered to the left, and on-target sites are marked with a red dot. The sgRNP delivery method and duration of treatment is shown at the bottom, and the column labeled "DT" shows data from sgRNAs directly targeting each SITE-Seq target site.

off-target sites, the increase in editing over time was the most pronounced, whereas editing at *VEGFA* and *FANCF* off-target sites did not appear to increase significantly from day 14 to day 28. We also observed that stable expression of *FANCF* sgRNP generated more off-target activity in K562 cells than it did in HEK293 cells. Surprisingly, transfection of *VEGFA*-targeting sgRNA into HEK293-Cas9–GFP cells generated less on-target editing, yet also gave rise to a new subset of cellular off-targets relative to the prolonged treatment condition. Of note, sgRNAs were cotransfected with exogenous 'carrier' plasmid DNA, and we found insertion of DNA sequences containing homology to this plasmid at most of the cellular off-targets in question (data not shown). Thus, it may be that inclusion of exogenous DNA altered DNA repair, amplifying low-frequency editing events, as has been previously observed with linear single- and double-stranded DNA[22]. Taken together, these data demonstrate that further extending the duration of sgRNP treatment, altering cell type, or even including exogenous DNA can alter off-target mutation frequencies and reveal additional SITE-Seq target sites as bona fide cellular off-targets.

Effect of accessibility on off-target editing

Many SITE-Seq target sites are not edited in cells even after persistent sgRNP expression, possibly because these sites are less accessible, impeded by structural elements of eukaryotic chromatin[23]. We examined accessibility at each SITE-Seq target site as reported by either (i) DNase-I activity[24] or (ii) direct targeting of sgRNPs with the expectation that sites with reduced accessibility would also show reductions in on-target editing. sgRNAs were transfected into HEK293-Cas9–GFP cells, and editing was measured after 2 d. At a given SITE-Seq target site, we observed no obvious correlation between off-target or on-target activity and DNaseI hypersensitivity (**Fig. 4** and **Supplementary Fig. 3**), which suggested that genomic accessibility was not a major determinant of off-target editing activity in our experiments.

Comparison of SITE-Seq with other off-target analysis methods and *in silico* prediction

The sgRNAs chosen for targeting *VEGFA* and *FANCF* have been previously characterized using other genome-wide off-target identification methods[11,12,14]. Relative to each other, only partially overlapping subsets of off-target sites were recovered by those studies, perhaps because of differences in source DNA or biases in the methods. In comparison, SITE-Seq recovered virtually all the sites identified by GUIDE-seq, HTGTS and Digenome-seq (**Fig. 5** and **Supplementary Tables 2**,**5**,**23**), as well as 11 new cellular off-targets, despite only testing a subset of SITE-Seq targets in cells (**Fig. 3a**). We additionally compared SITE-Seq with two commonly used *in silico* off-target prediction programs, CCtop and Cas-OFFinder[25,26]. Across guides, both CCtop and Cas-OFFinder only recovered a subset of SITE-Seq target sites (**Supplementary Tables 2–9**,**24**), missed some that were found as cellular off-targets, and did not predict which sites would have cellular activity. Taken together, these data suggest that off-target prediction programs remain an important area for future development, and that experimental determination of cut sites remains essential in understanding nuclease specificity.

SITE-Seq-based screening of guide specificity

The experiments in this study suggest that SITE-Seq with lower concentrations of sgRNP will produce a manageable list of sites to examine in cells (for some guides), and that this list will contain most or all cellular off-targets in the context of certain sgRNP delivery conditions. Indeed, SITE-Seq at 4 nM sgRNP (targeting *FANCF* or *VEGFA*) identified virtually all cellular off-targets in our validation test set that were generated after transfection of preassembled sgRNP or 14 d of stable sgRNP expression (data not shown). Accordingly, we next screened 83 sgRNAs tiling the epidermal growth factor receptor (*EGFR*) to highlight the guides that show low biochemical off-target activity and high cellular on-target activity (**Supplementary Fig. 4a**). Specifically, we

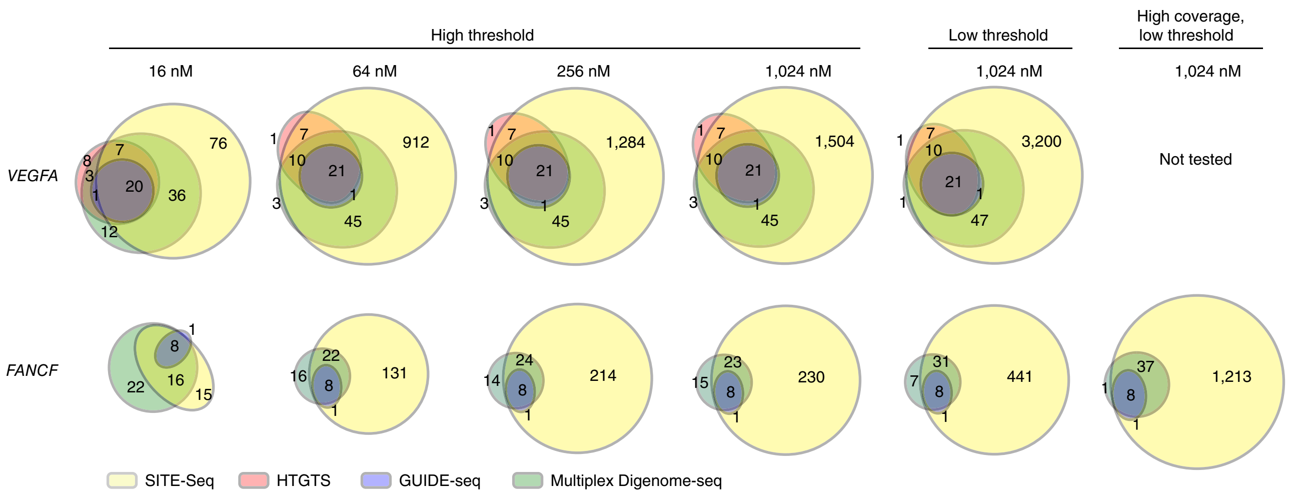


Figure 5 | Comparison of SITE-Seq with other off-target analysis methods. Venn diagrams showing the number of overlapping sites captured by SITE-Seq, HTGTS, GUIDE-seq and multiplex Digenome-seq[11,12,14]. The sgRNP target site is shown on the left, the concentration of sgRNP used in SITE-Seq is shown above the Venn diagrams, and the analysis pipeline and coverage is shown at the top. High-threshold analysis recovered sites with >10 reads that terminated at the same position and contained statistically significant target motifs ±2 nt from the termination point. Low-threshold analysis recovered sites with >5 reads that terminated at the same position and contained statistically significant target motifs ±30 nt from the termination point.

assembled sgRNPs and (i) performed SITE-Seq at 4 nM then (ii) delivered each as a preassembled complex into cells and evaluated the frequency of EGFR knockdown by flow cytometry.

SITE-Seq libraries made from control guides (*AAVS1*, *VEGFA*) recovered all cellular off-targets previously found after transfection of preassembled sgRNP, supporting the rationale behind our screen (**Supplementary Fig. 4b**). SITE-Seq analysis of the *EGFR* guides yielded a wide range of specificities, with some guides displaying no off-targets and the most promiscuous guide generating 1,644 SITE-Seq target sites. Next, we plotted on-target activity with the number of SITE-Seq target sites recovered (**Supplementary Fig. 4c**). Through this analysis, we identified eight sgRNPs that exhibited >70% EGFR knockdown and < 30 SITE-Seq target sites (five of these guides showed no off-target cleavage in any known genomic coding sequence). Thus, SITE-Seq in conjunction with functional assays may be used to highlight promising sgRNPs for detailed examination and product development.

DISCUSSION

While CRISPR–Cas9 has been widely deployed in a research setting, for further development of biotechnology products, there remains a critical need to develop methods that build toward a thorough understanding of enzyme specificity and activity. The method presented here, SITE-Seq, enables the detailed biochemical mapping of Cas9 cleavage sites within genomic DNA. These sites can then be verified for off-target editing in cells using standard, well-established high-sensitivity methods such as amplicon sequencing. In contrast to GUIDE-seq and HTGTS, SITE-Seq is not reliant on cellular events such as DNA repair—which will likely be influenced by Cas9 delivery method, cell type, and target site[11,12], as we observed in our cell-based experiments. Digenome-seq, similar to SITE-Seq, finds off-target sites by searching for read pileups with identical termination sites; but it does not enrich for cleaved fragments and may therefore require impractically high sequencing coverage to screen off-target sites comprehensively[13,14].

We observed that only a small fraction of the total SITE-Seq target sites recovered were of high biochemical sensitivity (i.e., cleaved with lower concentrations of sgRNP). Sites in this category were highly enriched in the collection of targets that presented off-target editing in cells after transient transfection. We also found that even after 4 weeks of persistent sgRNP expression, sites with lower biochemical sensitivity (i.e., recovered only with ≥ 256 nM sgRNP in SITE-Seq) did not present as cellular off-targets. However, in the large collection of sites recovered with ≤ 64 nM sgRNP, we uncovered additional off-target editing after either prolonging expression of sgRNP, altering our delivery method and/or changing cell type. This suggests that off-target editing is multifactorial and indicates that a biochemical approach to identify candidate cellular off-targets, followed by detailed cell-based experiments, may be the only way to obtain a complete picture of specificity for a given genome editing application. Moreover, SITE-Seq can also be applied to genomic DNA purified from a patient, where nucleotide and structural variants may impact off-target activity. Ultimately, we envision that SITE-Seq, in conjunction with cell-based validation, will constitute a robust pipeline for selecting sgRNPs and delivery methods with maximized activity and specificity.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

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

ACKNOWLEDGMENTS

The authors thank R. Haurwitz, S. Sternberg, B. McClung, and the members of Caribou Biosciences for providing helpful comments on the manuscript.

AUTHOR CONTRIBUTIONS

J.K.Y., V.L., S.D., M.C., A.P.M. and P.D.D. designed critical proof-of-concept experiments to initially develop the SITE-Seq protocol. P.C. performed all SITE-Seq experiments in this study; B.N.J. performed amplicon digestion experiments;

M.S.T. and L.M.B. performed sequencing for cell-based validation experiments. P.C., E.M.S., B.V., E.G. and E.L. performed cell-based validation experiments. S.G., A.M.L. and L.S.E. generated sgRNA and other reagents. P.C., C.K.F., P.D.D., M.M.C., D.C., M.C., B.N.J., A.H.S., J.K.Y. and A.P.M. analyzed the data. P.C., C.K.F., P.D.D., M.C., J.K.Y. and A.P.M. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Hsu, P.D., Lander, E.S. & Zhang, F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **157**, 1262–1278 (2014).
2. Doudna, J.A. & Charpentier, E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science* **346**, 1258096 (2014).
3. Sternberg, S.H. & Doudna, J.A. Expanding the biologist's toolkit with CRISPR-Cas9. *Mol. Cell* **58**, 568–574 (2015).
4. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
5. Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V. Cas9–crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc. Natl. Acad. Sci. USA* **109**, E2579–E2586 (2012).
6. Anders, C., Niewoehner, O., Duerst, A. & Jinek, M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* **513**, 569–573 (2014).
7. Nishimasu, H. *et al.* Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156**, 935–949 (2014).
8. Hsu, P.D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31**, 827–832 (2013).
9. Pattanayak, V. *et al.* High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat. Biotechnol.* **31**, 839–843 (2013).
10. Fu, Y. *et al.* High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.* **31**, 822–826 (2013).
11. Tsai, S.Q. *et al.* GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat. Biotechnol.* **33**, 187–197 (2015).
12. Frock, R.L. *et al.* Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases. *Nat. Biotechnol.* **33**, 179–186 (2015).
13. Kim, D. *et al.* Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nat. Methods* **12**, 237–243, 1 p following 243 (2015).
14. Kim, D., Kim, S., Kim, S., Park, J. & Kim, J.-S. Genome-wide target specificities of CRISPR-Cas9 nucleases revealed by multiplex Digenome-seq. *Genome Res.* **26**, 406–415 (2016).
15. Fu, B.X.H., St Onge, R.P., Fire, A.Z. & Smith, J.D. Distinct patterns of Cas9 mismatch tolerance *in vitro* and *in vivo*. *Nucleic Acids Res.* **44**, 5365–5377 (2016).
16. Fu, B.X.H., Hansen, L.L., Artiles, K.L., Nonet, M.L. & Fire, A.Z. Landscape of target:guide homology effects on Cas9-mediated cleavage. *Nucleic Acids Res.* **42**, 13778–13787 (2014).
17. Wang, X. *et al.* Unbiased detection of off-target cleavage by CRISPR-Cas9 and TALENs using integrase-defective lentiviral vectors. *Nat. Biotechnol.* **33**, 175–178 (2015).
18. Crosetto, N. *et al.* Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nat. Methods* **10**, 361–365 (2013).
19. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
20. Kim, S., Kim, D., Cho, S.W., Kim, J. & Kim, J.-S. Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome Res.* **24**, 1012–1019 (2014).
21. Liang, X. *et al.* Rapid and highly efficient mammalian cell engineering via Cas9 protein transfection. *J. Biotechnol.* **208**, 44–53 (2015).
22. Richardson, C.D., Ray, G.J., Bray, N.L. & Corn, J.E. Non-homologous DNA increases gene disruption efficiency by altering DNA repair outcomes. *Nat. Commun.* **7**, 12463 (2016).
23. Horlbeck, M.A. *et al.* Nucleosomes impede Cas9 access to DNA *in vivo* and *in vitro*. *eLife* **5**, e12677 (2016).
24. Boyle, A.P., Guinney, J., Crawford, G.E. & Furey, T.S. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* **24**, 2537–2538 (2008).
25. Stemmer, M., Thumberger, T., Del Sol Keyer, M., Wittbrodt, J. & Mateo, J.L. CCTop: an intuitive, flexible and reliable CRISPR/Cas9 target prediction tool. *PLoS One* **10**, e0124633 (2015).
26. Bae, S., Park, J. & Kim, J.-S. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics* **30**, 1473–1475 (2014).

ONLINE METHODS

A step-by-step protocol for SITE-Seq is available as a **Supplementary Protocol** and at *Protocol Exchange*[27].

Cas9 and sgRNAs. Recombinant *Spy* Cas9 and biochemically transcribed sgRNAs were generated as previously described[28]. Oligonucleotides used in the generation of sgRNA templates can be found in **Supplementary Table 25**. The sgRNA DNA template was removed before the SITE-Seq procedure by digesting the reaction with RNase-free DNase I (New England BioLabs, NEB). For all experiments except cell-based validation with biochemically assembled RNP, sgRNAs were purified using the GeneJET RNA purification kit (Thermo Fisher) according to the manufacturer's instructions and quantified by Nanodrop spectrophotometry measuring absorbance at 260 nm. For cell-based validation with biochemically assembled sgRNPs, sgRNAs were used without purification. For generation of polyclonal cell lines stably expressing sgRNP, human U6 promoter-driven sgRNAs were delivered as linearized plasmids or Lentiviral vectors.

SITE-Seq method. *Genomic DNA digestion.* High molecular weight (HMW) genomic DNA (gDNA) was purified from HEK293 cells using the DNeasy Blood & Tissue kit (Qiagen) according to the manufacturer's instructions. sgRNPs were assembled biochemically to digest the genomic DNA. Specifically, sgRNA was denatured by heating at 95 °C for 2 min then allowed to cool at room temperature for ~5 min. For the serial dilution experiments, recombinant *Spy* Cas9 was incubated with 15-fold molar excess sgRNA at 37 °C for 10 min, then mixed with 7.5 µg HMW gDNA at 37 °C for 4 h in a 50 µL final reaction volume (for the *EGFR* tiling experiment, sgRNP was assembled with 3 µL of *in vitro* transcribed sgRNA, which should be in molar excess of Cas9 across guides). sgRNP assembly and gDNA digestion proceeded in reaction buffer (20 mM HEPES, pH 7.4, 100 mM KCl, 5 mM $MgCl_2$, 1 mM DTT, 5% glycerol). The digestion reaction was terminated by incubating with Proteinase K at 0.2 µg/µL (Denville Scientific) and 0.8 µg/µL RNase A (Sigma) at 37 °C for 20 min then at 55 °C for 20 min. Digested HMW gDNA was purified with SPRIselect (Beckman Coulter) according to manufacturer's instructions at a 1:1 bead:volume ratio and eluted in 50 µL molecular-biology-grade water.

Adaptor set 1 ligation. 42 µL of purified, digested gDNA was 3′ adenylated with the NEBNext dA-tailing module (NEB) according to the manufacturer's instructions. Adenylated gDNA was then ligated to biotinylated, Illumina-compatible adaptor set 1 (see **Supplementary Table 26** for oligo information). In brief, the adaptor set 1 was formed by incubating 25 µM AS1A and 25 µM AS1B oligos (**Supplementary Table 26**) in annealing buffer (10 mM Tris, pH 8.0, 50 mM NaCl, 1 mM EDTA) at 95 °C for 2 min then allowing the reaction to cool to room temperature for ~45 min. Next, 38 µL of 3′ adenylated gDNA was incubated with 500 nM adaptor set 1 and 5 µL NEB quick ligase in 1× T4 DNA ligase buffer (NEB) for 30 min at 25 °C then for ~16 h at 16 °C; then the gDNA was purified with SPRIselect (Beckman Coulter) according to manufacturer's instructions at a 1:1 bead:volume ratio and eluted in 50 µL molecular-biology-grade water.

Fragmentation and adaptor set 2 ligation. 40 µL of adaptor-ligated gDNA was fragmented by incubating with 5 µL NEBNext dsDNA Fragmentase (NEB) in 1× Fragmentase buffer (v2) at 37 °C for 12 min in a 50 µL total reaction volume. The reaction was quenched by adding 12.5 µL 0.5 M EDTA, and fragmented gDNA was purified with SPRIselect (Beckman Coulter) according to manufacturer's instructions at a 1.5:1 (sgRNP serial dilution experiment) or 1:1 (*EGFR* tiling experiment) bead:volume ratio and eluted in 50 µL molecular-biology-grade water. The fragmented gDNA was end repaired and 3′ adenylated with the NEBNext End-Repair module (NEB) according to the manufacturer's instructions. The DNA was then ligated to a second adaptor set with either the NEBNext Ultra Ligation module (*VEGFA*) or NEBNext quick ligation module (all other guides) according to the manufacturer's instructions (NEB). The second set of adapters contained an outer Illumina-compatible sequence, a stretch of 5–7 random nucleotides (see AS2B-AS2D, **Supplementary Table 26**) and an inner sequence to enable double-stranded adaptor formation.

Biotin–streptavidin affinity purification, library PCR, and sequencing. The dual adaptor-ligated DNA was purified using 50 µL of M-280 Dynabead (Thermo Fisher) according to the manufacturer's instructions. DNA-bound beads were then used as a template for an initial round of PCR amplification using Q5 Hot-Start High Fidelity DNA Polymerase (NEB) and 16 cycles of PCR with primers RF and RR (see **Supplementary Table 26** for sequences). The library was then diluted 1:50 in molecular-biology-grade water and amplified a second time using index primers IF and IR (see **Supplementary Table 26**) in a 12-cycle PCR reaction with Q5 Hot-Start High Fidelity DNA Polymerase (NEB). The indexed library was then purified with SPRIselect (Beckman Coulter) according to manufacturer's instructions at a 0.9:1 bead:volume ratio. Indexed SITE-Seq libraries were sequenced on a MiSeq, HiSeq (2 × 151 paired-end reads) or NextSeq (151 single-end reads) (Illumina).

SITE-Seq analysis and statistics. Sequencing reads were aligned to the reference human genome (hg38) using the Bowtie2 aligner[29] with default settings, and sequencing coverage for all samples (serial dilution experiment) fell within ~0.62–2.46 million paired-end reads. The alignments were converted to the BAM file format, indexed, and sorted using SAMtools[30]. Preliminary identification of peaks in the aligned reads was performed using MACS2 with a qvalue of 0.05 and no model[31]. To identify the subset of peaks containing the sharp discontinuity in aligned reads expected on account of double-stranded DNA cleavage by the sgRNP, an algorithm sensitive to the peak composition was applied. Putative Cas9 cleavage sites were called using minimum thresholds of either five or ten reads with edges aligned to the same nucleotide. By using unbiased motif finding to identify likely true positives, as discussed below, the estimated false discovery rate (FDR) ranged from 5% to 25% depending on the specific sgRNA and peak detection threshold used.

To minimize the number of false positives caused by spurious peaks or background double-stranded breaks, the collection of putative cleavage sites for each guide was searched for over-represented sequences matching the combined protospacer and PAM motif for the on-target. An unbiased search for 23-mer motifs was performed using genomic sequence either ± 30 nt (low threshold) or ± 2 nt (high threshold) around each discontinuity with MEME[32]. The input set comprised the unique loci identified across all eight sgRNP concentrations assayed. For each guide, the top motif discovered was highly over-represented in the set (E-values below 10^{-4}) and matched the on-target protospacer and PAM.

doi:10.1038/nmeth.4284

Other motifs discovered include several dozen centromeric loci for each guide containing AATGG repeats, consistent with hotspots in double-stranded DNA breaks[33].

The **Supplementary Software** SITE-Seq_core_feature_calling_functions.py contains functions written in the Python programming language to identify SITE-Seq features from sequence alignment files. It requires Python 2.7 and the Pysam interface to the Samtools software package. The function find_initial_read_pileups() returns a list of all peaks present in the aligned reads of a BAM sequence file. This list is used as an input to the call_site_seq_features() function, which identifies valid SITE-Seq loci based on the precise shape of the read pileup. These functions can be embedded inside general-purpose routines for aligning and processing sequence data depending on the needs of an individual laboratory.

Biochemical digestion of amplicons containing SITE-Seq target sites. Double-stranded DNA containing target sites for use in biochemical Cas9 cleavage assays was produced using PCR amplification of the target region from HEK293 genomic DNA. PCR reactions were carried out using Q5 Hot Start High-Fidelity 2× Master Mix (NEB) as per the manufacturer's recommendation. 200 ng of human gDNA was mixed with primers at a final concentration of 500 nM each in a total reaction volume of 100 µL. PCR was performed under the following conditions: 98 °C for 2 min; 35 cycles of 20 s at 98 °C, 20 s at 60 °C, 20 s at 72 °C; and a final extension at 72 °C for 2 min. The quality and quantity of the amplified products were analyzed using the Fragment Analyzer system (Advanced Analytical Technologies) and the DNF-473 Standard-Sensitivity NGS Fragment Analysis Kit (Advanced Analytical Technologies) according to the manufacturer's protocols.

Cas9 was serially diluted two-fold in reaction buffer (20 mM HEPES, 100 mM KCl, 5 mM $MgCl_2$, 1 mM DTT, and 5% glycerol at pH 7.4). The final concentrations of Cas9 in the reaction were: 64 nM, 256 nM, and 1,024 nM. Next, the sgRNA was incubated for 2 min at 95 °C and allowed to equilibrate to room temperature. Then, the sgRNA was added to the Cas9 dilutions at a final concentration of 2,048 nM and allowed to complex with Cas9 for 10 min at 37 °C. The cleavage reactions were initiated by the addition of target DNA at a final concentration of 15 nM. Samples were mixed and centrifuged briefly before incubation at 37 °C for 15 min. The cleavage reactions were terminated by the addition of Proteinase K (Denville Scientific) at a final concentration of 0.2 µg/µL and 0.8 µg/µl RNase A Solution (Sigma-Aldrich) and incubated for 20 min at 37 °C and 20 min at 55 °C. Reactions were analyzed using the Fragment Analyzer system (Advanced Analytical Technologies) and the DNF-910 dsDNA 910 Reagent Kit, 35–1,500 bp (Advanced Analytical Technologies) according to the manufacturer's protocols. The Fragment Analyzer system provided the quantity (ng) of each peak in the reaction. This quantity was used to calculate the fraction of amplicon cleaved using the formula "Fraction cleaved = (Frag1 + Frag2)/(Frag1 + Frag2 + Parent)."

Cell-based experiments. HEK293, K562 and HeLa cells were obtained from ATCC, and the stable Cas9-expressing line (HEK293-Cas9–GFP)[28] was generated from the HEK293 line. HEK cells are listed in the database of commonly misidentified cell lines[34], and they were chosen in this study because previous experiments showed that they support robust gene editing[28]. The cell lines used here were authenticated by STR profiling, and PCR screening confirmed that they were negative for mycoplasma contamination. HEK293, K562 and HeLa cells were cultured in DMEM, IMDM or EMEM medium, respectively, and supplemented with 10% fetal bovine serum (FBS) and antibiotics and antimycotics.

Transient transfections. For validation of SITE-Seq target sites, 100 ng of sgRNA was transfected into HEK293-Cas9–GFP cells by combining the sgRNA with 200 ng pUC-18 plasmid DNA and 0.3 µL TransIT-X2 (Mirus Bio) in a total volume of 50 µL DMEM and incubating at room temperature for 30 min. For experiments with modified sgRNAs directly targeting SITE-Seq target sites (**Fig. 4**), 500 ng of sgRNA was transfected into HEK293-Cas9–GFP cells by combining the sgRNA with 0.4 µL TransIT-X2 in a total volume of 50 µL DMEM and incubating at room temperature for 30 min. Transfection mix was combined with either 1 × 10^5 (validation of SITE-Seq targets sites) or 7.5 × 10^4 (direct targeting of SITE-Seq target sites) HEK293-Cas9–GFP cells in 100 µL culture medium and plated into wells of a collagen-I-coated 96-well cell culture plate. For sgRNP delivery, 2.2–2.5 µL of *in vitro* transcribed sgRNAs was incubated at 95 °C for 2 min, cooled to room temperature for 5 min, then incubated with 20 pmol of Cas9 in 5 µL total volume (reaction buffer of 20 mM HEPES, pH 7.4, 100 mM KCl, 5 mM $MgCl_2$, 5% glycerol) at 37 °C for 10 min to assemble as an sgRNP. sgRNPs were delivered into HEK293 and HeLa cells via nucleofection (Lonza).

Generation of lines stably expressing sgRNP. For lentivirus transduction experiments, HEK293 cells were plated at a density of 5 × 10^6 cells per 10 cm dish 24 h before transfection. Cells were transfected with 2 µg custom lentiviral sgRNA expression vectors and 10 µg Ready-to-use Packaging Plasmid Mix (Cellecta) using TransIT-293 (Mirus Bio) as per manufacturers' instructions. Viral supernatants were harvested 48 h post-transfection. Supernatants were filtered using 0.45 µM PVDF filter membranes, supplemented with 5 µg/mL Polybrene and used to infect HEK293-Cas9–GFP expressing cells. Stably expressing cells were selected by adding puromycin (Gibco) 48 h after transfection and maintained until harvesting. For testing stable expression of *FANCF* sgRNP in HEK293 and K562, plasmid constructs directing expression of Cas9 and *FANCF* sgRNA were linearized before transfection. Next, 5 × 10^6 HEK293 cells were transfected with 40 µg linearized plasmid DNA using 20 µl TransIT-X2, and 5 × 10^6 K562 cells were transfected with 14 µg linearized plasmid DNA using the Neon transfection system (Thermo Fisher). Stably expressing cells were selected by adding puromycin (Gibco) 48 h after transfection and maintained until harvesting.

FACS analysis of EGFR knockdown. FACS was performed 4 d after nucleofection of HeLa cells with *EGFR*-targeting sgRNPs. 100 M cells were detached with TrypLE Express (Gibco), stained with 5 µL APC anti-human EGFR (Clone AY13, Sony Biotechnology) in 100 µL total volume and analyzed using CytoFLEX Flow Cytometer (Beckman Coulter Life Sciences) and FlowJo (FlowJo, LLC).

Amplicon PCR and sequencing. For sgRNA and sgRNP transient transfections, cells were harvested 2–3 d after transfection. For polyclonal cell lines generated from puromycin selection,

doi:10.1038/nmeth.4284

cells were harvested 6, 14 and 28 days after the initial transduction/transfection. For all experiments, gDNA was lysed using QuickExtract DNA Extraction Solution (Epicentre) as per manufacturer's instructions. SITE-Seq target sites were amplified in a two-step PCR reaction. In brief, 3.75–8 μL (corresponding to ~2,000–8,000 cells, depending on the sample) of cell lysate was used as a template for PCR amplification with Q5 Hot-Start High Fidelity DNA Polymerase (NEB) and ~100–400 unique primer pairs containing an internal locus-specific region and an outer Illumina-compatible adaptor sequence (**Supplementary Tables 12–19**). A second PCR reaction targeting the outer-adaptor sequence was performed to append unique indices to each amplicon (primers IF and IR in **Supplementary Table 26**). SITE-Seq target sites were sequenced on a MiSeq system with 2 × 151 paired-end reads and version 2 chemistry (Illumina). Depth of coverage was ~5,000–25,000 reads/amplicon.

Off-target editing analysis. For each site, indel frequencies were calculated by subtracting control reference cells from transfected cells. Sites with <500 total reads or >0.2% mutation frequencies calculated in the control reference condition (usually due to sequencing errors or polymorphisms unrelated to genome editing located near the cut site) were discarded from analysis. Indels were counted as mutant if they occurred within 10 nt of the putative Cas9 cut site. For all heatmap analyses, sites were only tallied if they accumulated >0.1% mutant reads relative to control condition; and visual inspection confirmed that indels were abutting the Cas9 cut site in the transfected condition and were not present in the control condition.

Data availability statement. The data that support the findings of this study are available from the corresponding author upon request as well as under the BioProject ID PRJNA329375.

27. Cameron, P. *et al.* SITE-Seq: a genome-wide method to measure Cas9 cleavage. *Protocol Exchange* http://dx.doi.org/10.1038/protex.2017.043 (2017).
28. Briner, A.E. *et al.* Guide RNA functional modules direct Cas9 activity and orthogonality. *Mol. Cell* **56**, 333–339 (2014).
29. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
30. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
31. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
32. Bailey, T.L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proc. Int. Conf. Intell. Syst. Mol. Biol.* (eds. Altman, R. *et al.*) 28–36 (AAAI Press, 1994).
33. Blitzblau, H.G., Bell, G.W., Rodriguez, J., Bell, S.P. & Hochwagen, A. Mapping of meiotic single-stranded DNA reveals double-stranded-break hotspots near centromeres and telomeres. *Curr. Biol.* **17**, 2003–2012 (2007).
34. Capes-Davis, A. *et al.* Check your cultures! A list of cross-contaminated or misidentified cell lines. *Int. J. Cancer* **127**, 1–8 (2010).

doi:10.1038/nmeth.4284