CIRCLE-seq: a highly sensitive *in vitro* screen for genome-wide CRISPR–Cas9 nuclease off-targets

Shengdar Q Tsai[1–4,6], Nhu T Nguyen[1–3], Jose Malagon-Lopez[1–5], Ved V Topkar[1–3], Martin J Aryee[1–5] & J Keith Joung[1–4]

Sensitive detection of off-target effects is important for translating CRISPR–Cas9 nucleases into human therapeutics. *In vitro* biochemical methods for finding off-targets offer the potential advantages of greater reproducibility and scalability while avoiding limitations associated with strategies that require the culture and manipulation of living cells. Here we describe circularization for *in vitro* reporting of cleavage effects by sequencing (CIRCLE-seq), a highly sensitive, sequencing-efficient *in vitro* screening strategy that outperforms existing cell-based or biochemical approaches for identifying CRISPR–Cas9 genome-wide off-target mutations. In contrast to previously described *in vitro* methods, we show that CIRCLE-seq can be practiced using widely accessible next-generation sequencing technology and does not require reference genome sequences. Importantly, CIRCLE-seq can be used to identify off-target mutations associated with cell-type-specific single-nucleotide polymorphisms, demonstrating the feasibility and importance of generating personalized specificity profiles. CIRCLE-seq provides an accessible, rapid, and comprehensive method for identifying genome-wide off-target mutations of CRISPR–Cas9.

CRISPR–Cas9 nucleases can be easily programmed to create targeted double-stranded breaks (DSBs)[1–5], and the simplicity of this process has driven widespread adoption of this genome-editing technology[6–12]. Cas9-induced DSBs can be repaired by cellular DNA repair pathways, resulting in targeted sequence alterations in the genomes of living cells and organisms[13–17]. Efficient cleavage by the commonly used *Streptococcus pyogenes* Cas9 (SpCas9) requires 17–20 nt of complementarity between a Cas9-associated single guide RNA (sgRNA) and a target site (protospacer)[5,18–21] adjacent to a 5′-NGG protospacer adjacent motif (PAM)[5,22,23].

For clinical translation of CRISPR–Cas9, defining the frequencies and locations of unintended nuclease-induced off-target mutations is important[10,24–26]. Although cell-based methods for genome-wide off-target identification have been described[27–31], these can miss off-target mutations that occur with frequencies below ~0.1% in nuclease-treated cell populations. Furthermore, requirements for efficient cellular transfection limit the feasibility, scalability, and reproducibility of these methods, particularly with nontransformed cell types that are used for many research and therapeutic applications.

By contrast, *in vitro* strategies for defining nuclease-induced off-target DSBs have potential advantages over cell-based approaches. Assays using purified components improve reproducibility, bypass the need for efficient cellular transduction or transfection, and avoid cell fitness effects. Importantly, concentrations of active nuclease can be raised to high levels *in vitro*, potentially enabling identification of sequences that may be rarely cleaved in cells. An *in vitro* method for characterizing Cas9 cleavage specificity using partially randomized DNA libraries biased toward specific target DNA sites has previously been described, but a limitation of this approach is that many identified sites do not actually occur in the human genome[32].

To our knowledge, only a single *in vitro* genome-wide off-target identification method, Digenome-seq[33], has been described in the literature. This approach relies on nuclease cleavage of genomic DNA, sequencing adaptor ligation to all free ends (nuclease- and non-nuclease-induced), high-throughput sequencing, and bioinformatic identification of nuclease-cleaved sites exhibiting signature uniform mapping end positions. However, Digenome-seq analysis requires a large number of reads (~400 million), and the high background of random genomic DNA reads makes it challenging to identify low-frequency nuclease-induced cleavage events.

Here we describe CIRCLE-seq, an *in vitro* screen for identifying genome-wide off-target cleavage sites of CRISPR–Cas9 that virtually eliminates the high background of random reads observed with Digenome-seq. This improvement enables not only substantially more sensitive off-target site detection, but it also allows researchers to deploy CIRCLE-seq more easily using widely accessible benchtop next-generation sequencing platforms. For most

[1]Molecular Pathology Unit, Massachusetts General Hospital, Charlestown, Massachusetts, USA. [2]Center for Cancer Research, Massachusetts General Hospital, Charlestown, Massachusetts, USA. [3]Center for Computational and Integrative Biology, Massachusetts General Hospital, Charlestown, Massachusetts, USA. [4]Department of Pathology, Harvard Medical School, Boston, Massachusetts, USA. [5]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA. [6]Present address: Department of Hematology, St. Jude Children's Research Hospital, Memphis, Tennessee, USA. Correspondence should be addressed to J.K.J. (jjoung@mgh.harvard.edu) or S.Q.T. (shengdar.tsai@stjude.org).

RECEIVED 15 MARCH; ACCEPTED 6 APRIL; PUBLISHED ONLINE 1 MAY 2017; DOI:10.1038/NMETH.4278

Cas9–guide RNA complexes tested, CIRCLE-seq can identify all off-target sites in human genomic DNA found by GUIDE-seq[29] and high-throughput gene translocation sequencing (HTGTS[30]), two of the most sensitive previously described cell-based methods. Importantly, CIRCLE-seq also identifies new bona fide off-target sites that occur in human cells. We also show that CIRCLE-seq can identify off-target sites in the absence of a reference genome, opening the door to off-target profiling in organisms lacking full genomic sequence or outbred populations with considerable sequence heterogeneity. Lastly, we demonstrate how CIRCLE-seq can be used to identify off-target cleavage sites that are enhanced or diminished by cell-type-specific single-nucleotide polymorphisms (SNPs), demonstrating the feasibility and importance of defining personalized off-target profiles.

RESULTS

Overview and optimization of CIRCLE-seq

To reduce background genomic DNA reads that occur with Digenome-seq and enhance detection of desired Cas9-nuclease-cleaved genomic DNA fragments, we sought to selectively sequence Cas9-cleaved genomic DNA. We designed restriction-enzyme-independent strategies to generate and enzymatically select for the conversion of randomly sheared DNA into one of two different types of covalently closed DNA structures: attachment of stem-loops to linear DNA ends (**Supplementary Fig. 1**) or circularization of linear fragments (**Fig. 1**; **Supplementary Figs. 2** and **3**). Subsequent nuclease-induced cleavage of either population of covalently closed DNA molecules at on- and off-target sites would release free DNA ends required for subsequent adaptor-ligation and high-throughput sequencing. Comparison of these two approaches demonstrated that circularization was at least an order of magnitude more effective enriching for Cas9-nuclease cleaved genomic DNA fragments than attaching stem loops to linear DNA (**Supplementary Fig. 4a**). Importantly, nearly all sites identified, starting from linear DNA fragments with hairpin ends, were also identified from circularized DNA; and read counts were strongly correlated between both methods, suggesting that circularization does not bias the range or frequency of identified off-target sites (**Supplementary Fig. 4b**). We named the circularization method CIRCLE-seq, and optimization and characterization of its technical reproducibility are described in the **Supplementary Note 1** (**Supplementary Figs. 3** and **5**). In contrast to other genome-wide nuclease off-target discovery methods, CIRCLE-seq uniquely enables sequencing of both sides of a single cleavage site in one DNA molecule using paired-end sequencing.

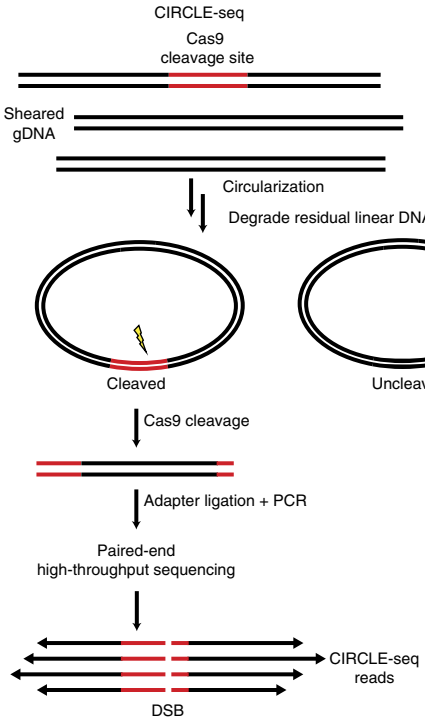

Figure 1 | Overview of CIRCLE-seq. Genomic DNA is sheared and circularized by intramolecular ligation. Undesired linear DNA molecules are degraded away by exonuclease treatment. Circular DNA molecules containing a Cas9 cleavage site (red) can subsequently be linearized with Cas9—releasing newly cleaved DNA ends for adaptor ligation, PCR amplification, and paired-end high-throughput sequencing. Each pair of reads generated by Cas9 cleavage contains complete sequence information for a single off-target site.

CIRCLE-seq enables sensitive *in vitro* detection of CRISPR–Cas9 off-target sites

To test the sensitivity of CIRCLE-seq, we compared off-target results of an sgRNA targeting the human *HBB* gene with off-target profiles of the same sgRNA generated by the most recent and accurate version of Digenome-seq[34]. Using genomic DNA from human K562 cells, CIRCLE-seq identified not only 26 of 29 off-target sites previously seen with Digenome-seq, but also 156 new sites (**Supplementary Fig. 6a**). For the three sites found by Digenome-seq but not by CIRCLE-seq, we observed supporting reads in the CIRCLE-seq data (**Supplementary Fig. 6b**), demonstrating that these sites were simply undersampled in these particular experiments. Of 156 new sites called by CIRCLE-seq but not Digenome-seq, we found that 29 also showed evidence of cleavage in the original Digenome-seq data[33] (**Supplementary Fig. 6c**); Digenome-seq likely failed to call these sites because of stringent informatics scoring criteria required to contend with the abundant genome-wide background reads generated by this method. By contrast, we found that such background reads were rare with CIRCLE-seq (**Supplementary Fig. 6d**). Indeed, we estimate the enrichment factor of CIRCLE-seq for nuclease-cleaved sequence reads to random background reads is ~180,000-fold higher than Digenome-seq based on examination of an on-target site with the two methods and adjustment for sequencing depth (**Supplementary Fig. 6d**). Start mapping positions of bidirectional CIRCLE-seq reads are consistent with the expected cleavage site of SpCas9 (3 bp before the PAM sequence) (**Supplementary Fig. 6e**); this demonstrates CIRCLE-seq's ability to map cleavage positions with nucleotide-level precision. Taken together, these results demonstrate that CIRCLE-seq possesses higher signal-to-noise ratios relative to those of Digenome-seq and using approximately 100-fold fewer sequencing reads; this likely accounts for its greater sensitivity for identifying genome-wide off-target sites.

Comparisons of CIRCLE-seq with cell-based off-target methods

To compare CIRCLE-seq with GUIDE-seq[29], we assessed six different sgRNAs targeted to nonrepetitive sequences that had been previously characterized by GUIDE-seq across two different

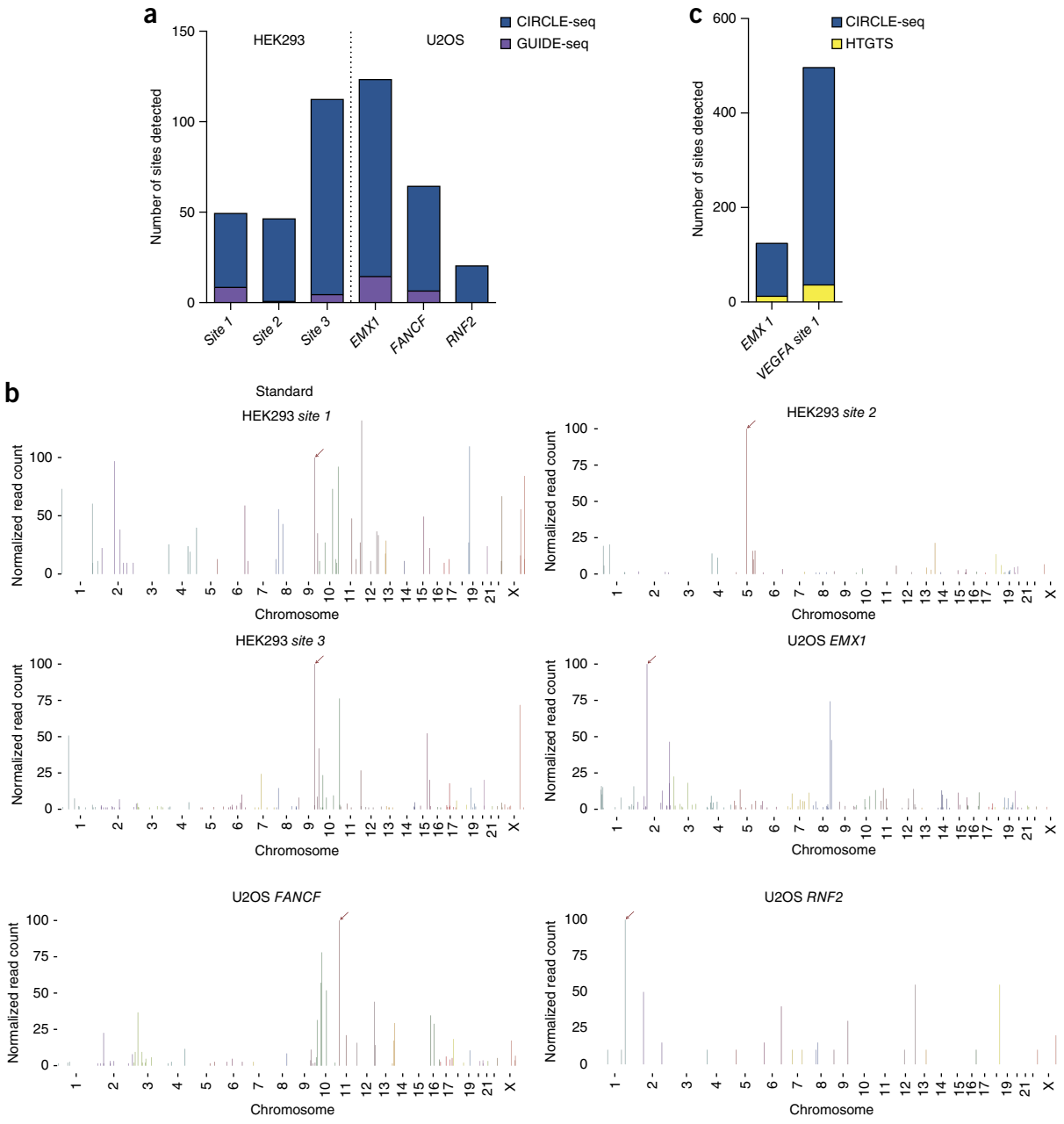

Figure 2 | Comparisons of CIRCLE-seq with cell-based GUIDE-seq and HTGTS methods. (**a**) Histogram showing the number of sites identified exclusively by CIRCLE-seq (blue) and by both CIRCLE-seq and GUIDE-seq (magenta) for sgRNAs designed toward standard nonrepetitive target sites. Numbers are tabulated from single CIRCLE-seq experiments. (**b**) Manhattan plots of CIRCLE-seq detected off-target sites, with bar heights representing CIRCLE-seq read count (normalized to site with highest read count) and organized by chromosomal position. Arrow indicates the intended target. (**c**) Histogram showing the number of sites detected exclusively by CIRCLE-seq (blue) or by both CIRCLE-seq and HTGTS (yellow). Numbers are tabulated from single CIRCLE-seq experiments performed on genomic DNA isolated from U2OS cells.

human cell lines. For these six different sgRNAs, CIRCLE-seq identified variable numbers of genome-wide off-target cleavage sites ranging in number from as few as 21 to as many as 124 (**Fig. 2**, **Supplementary Note 2**, and **Supplementary Table 2**) with up to six mismatches relative to the on-target site (**Supplementary Fig. 7**). For four of the six sgRNAs, CIRCLE-seq detected all off-target sites identified by GUIDE-seq (**Supplementary Fig. 8**); and for two other sgRNAs, it detected all but one off-target site for each (**Supplementary Fig. 8**). Closer examination of the CIRCLE-seq data for these experiments revealed supporting reads for these two sites but not of a sufficient number to exceed the statistical threshold for detection. In addition, these two undetected sites had been at the lower boundary of detection in our GUIDE-seq experiments. Taken together, these findings again suggest that these two off-target sites would be detected with modestly increased CIRCLE-seq sequencing depth. Importantly, for all six sgRNAs, CIRCLE-seq identified many more off-target sites than were previously found by GUIDE-seq, including sites for a sgRNA targeted to the *RNF2* gene, for which we had previously been unable to identify off-target sites. We obtained

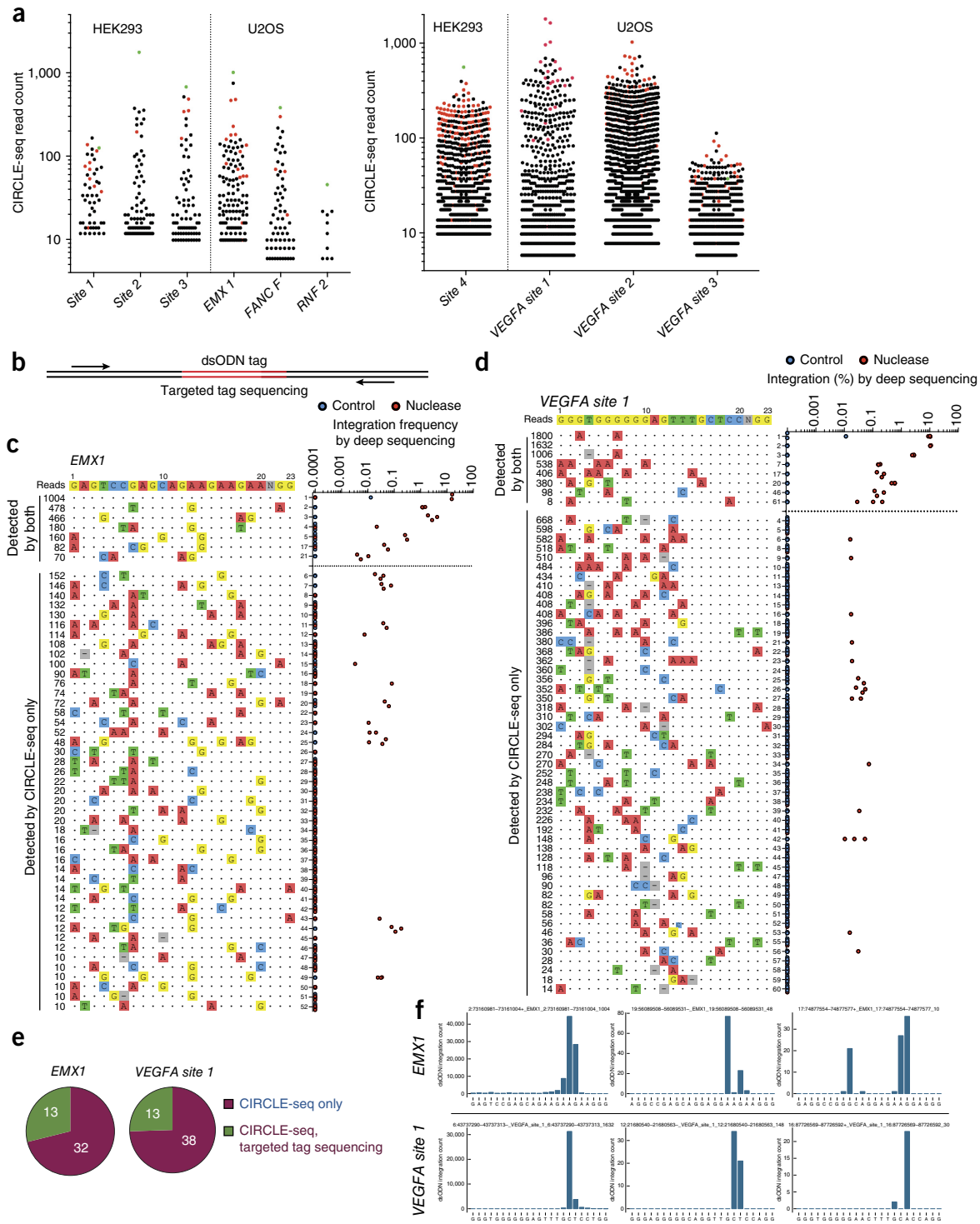

Figure 3 | CIRCLE-seq detects off-target sites that are cleaved in human cells. (**a**) Stem–leaf plot of CIRCLE-seq read counts for ten sgRNAs previously analyzed by GUIDE-seq. The on-target site is shown as a green dot, and off-target sites detected by GUIDE-seq are shown as red dots. (**b**) Schematic overview of the targeted tag sequencing approach. Primers are designed to amplify genomic regions flanking nuclease-induced DSBs from genomic DNA of cells treated with nuclease and double-stranded oligodeoxynucleotide (dsODN) tag. (**c**,**d**) Targeted tag integration frequencies at control off-target sites detected by both CIRCLE-seq and GUIDE-seq (upper part of panel) and off-target sites detected by CIRCLE-seq but not GUIDE-seq (lower part of panel) for sgRNAs targeted to *EMX1* and *VEGFA* site 1. Off-target sites are ordered top to bottom by CIRCLE-seq read count with mismatches to the intended target sequence indicated by colored nucleotides. Tag integration frequencies observed for control (blue) and nuclease-treated (red) cells are plotted on a log scale. (**e**) Pie charts showing fractions of CIRCLE-seq sites analyzed that are also detected by targeted tag sequencing. (**f**) Plots of integration positions observed by targeted tag sequencing. PAM bases are the last 3 nt from the right. Integrations occur at positions proximal to the location of the predicted DSB (3 bp from the PAM).

similar results using CIRCLE-seq to profile four additional sgRNAs targeted to repetitive sequences that we had previously characterized by GUIDE-seq (**Supplementary Note 1**, **Supplementary Note 2**, **Supplementary Figs. 7** and **8**).

We next compared CIRCLE-seq profiles of two sgRNAs targeted to *EMX1* and *VEGFA* that had previously been characterized by the cell-based HTGTS method. These CIRCLE-seq experiments identified 48 of the 51 off-target sites (94%) previously identified by HTGTS (**Supplementary Fig. 9**). Among the three HTGTS sites not detected by CIRCLE-seq, two were found when additional experimental replicates were performed; and the third had a low HTGTS score (**Supplementary Table 3**), suggesting that these sites would be detected with greater sequencing depth. Importantly, CIRCLE-seq also found a much greater number of off-target sites than were previously identified by HTGTS (**Fig. 2c**).

Off-target sites are mutated in human cells

An important question is whether novel off-target cleavage sites identified *in vitro* by CIRCLE-seq (and not by GUIDE-seq or HTGTS) are mutated in cells by Cas9–sgRNA complexes. Many off-target sites detected by both CIRCLE-seq and GUIDE-seq have high numbers of mapping CIRCLE-seq sequencing read counts (**Fig. 3a**), which strongly suggests that GUIDE-seq primarily detects off-target sites that are among the most efficiently cleaved *in vitro*. By contrast, many off-target sites found only by CIRCLE-seq have lower CIRCLE-seq read counts (**Fig. 3a**), suggesting that these might be missed by GUIDE-seq (**Supplementary Fig. 10**). In this case, we would anticipate difficulty in validating these sites in cells using standard targeted amplicon sequencing, because the error rate of next-generation sequencing places a floor for indel mutation detection of approximately 0.1%. Thus, to determine whether novel off-target sites detected only by CIRCLE-seq (but not in our original GUIDE-seq experiments) might be cleaved in human cells, we reasoned that we could instead perform high-depth targeted amplicon sequencing using genomic DNA obtained from cell-based GUIDE-seq experiments and look for tag integration as evidence of off-target cleavage (**Fig. 3b**). This strategy sidesteps the problem of the indel error rate associated with deep sequencing because tag integration occurs with a negligible background frequency, thereby permitting detection of lower frequency sites.

Using targeted tag integration sequencing, we examined 98 off-target sites found by CIRCLE-seq (but not by GUIDE-seq) for SpCas9 and sgRNAs targeted to *EMX1* and *VEGFA* (site 1). We chose sites that exhibited a range of CIRCLE-seq read counts and numbers of mismatches relative to the on-target site (**Supplementary Table 4**). As positive controls, we also selected a smaller set of off-target sites with variable numbers of CIRCLE-seq read counts that were found by both CIRCLE-seq and GUIDE-seq (**Supplementary Table 4**). Targeted amplicon sequencing revealed detection of the dsODN at all of the control off-target sites with frequencies that correlated well with GUIDE-seq read counts (**Fig. 3c,d**). Notably, we also detected double-stranded oligodeoxynucleotide (dsODN) integration at 26 of the 96 novel off-target sites identified only by CIRCLE-seq analysis (**Fig. 3c–e**), with frequencies in the low range (0.003 – 0.2%), as expected. The locations of all tag integrations map to the expected cleavage positions 3 bp away from a PAM sequence, consistent with these sites representing bona fide off-target cleavage sites (**Fig. 3f**). To further characterize the mutations observed at the CIRCLE-seq sites analyzed by targeted tag sequencing, we performed somatic variant calling. Using both the somatic variant caller MuTect and visual verification of sites with integrated genome viewer (IGV), we did not find evidence of single-nucleotide variants in our targeted tag sequencing experiments (data not shown).

Reference-genome-independent off-target site discovery

Because each pair of CIRCLE-seq reads yields sequences from both sides of a single CRISPR–Cas9 nuclease cleavage site, we reasoned that our method could identify off-target sites even without a reference genome sequence. We developed a mapping-independent off-target site discovery algorithm that merges CIRCLE-seq paired-end reads and directly searches for off-target cleavage sites resembling the on-target site (see Online Methods). Using this algorithm, we identified, on average, ~99.5% of CIRCLE-seq sites detected by our standard reference-based mapping algorithm and with more than ten CIRCLE-seq reads (**Supplementary Fig. 11**). Thus, CIRCLE-seq could potentially be used in a reference-independent fashion to identify off-target cleavage sites for organisms whose genome sequences are less well characterized and/or show high genetic variability (e.g., noninbred species in the wild).

Association of off-target sites with SNPs

With its higher throughput, an *in vitro* method such as CIRCLE-seq provides researchers with the opportunity to define patient-specific off-target profiles for any given Cas9–sgRNA nuclease. A previously published study described an interesting example of a single SNP influencing off-target cleavage[35]. To more broadly test whether genetic differences can influence nuclease-induced off-target cleavage, we performed CIRCLE-seq experiments on human K562 genomic DNA with six sgRNAs we had already assessed on human HEK293 and U2OS genomic DNAs (three sgRNAs on HEK293s and three on U2OS). Although many off-target sites for these sgRNAs showed well-correlated CIRCLE-seq read counts on DNA from both cell types tested, we also observed 55 sites that were preferentially cleaved in one cell type or the other (**Fig. 4a**). Further examination revealed that eight of these off-target cleavage sites harbored nonreference single-nucleotide polymorphisms (SNPs) that might account for these cell-type-specific differences in cleavage efficiencies (**Fig. 4b** and **Supplementary Table 5**). Interestingly, these SNPs were located in regions of protospacer complementarity as well as the PAM.

We next sought to estimate how frequently SNPs might impact off-target cleavage efficiency. To do this, we examined the genotypes of 2,504 individuals from the 1000 Genomes Project[36] at all 1,247 off-target sites we detected with CIRCLE-seq for the six sgRNAs (targeted to standard nonrepetitive sequences). We found, on average, genetic variation in ~2.5% of these off-target sites at the individual level (**Fig. 4c**). At a population level, we found that superpopulations contained genetic variation in an average of ~20% of these off-target sites (**Fig. 4c**). In addition, 50% of these off-target sites contained genetic variation for at least one individual sequenced in the 1000 Genomes Project (**Fig. 4c**). These frequencies are consistent with the expectation that, given existence of ~100 million validated human SNPs in the

most recent version of the dbSNP database[37], one might expect to find an SNP in ~69% of SpCas9 off-target sites in the human genome. As expected, the range of mismatches observed at the off-target sites we examined was increased when considering diverse individual genotypes from the 1000 genomes project (**Fig. 4d**). Interestingly, approximately 9% of variant off-target-site haplotypes

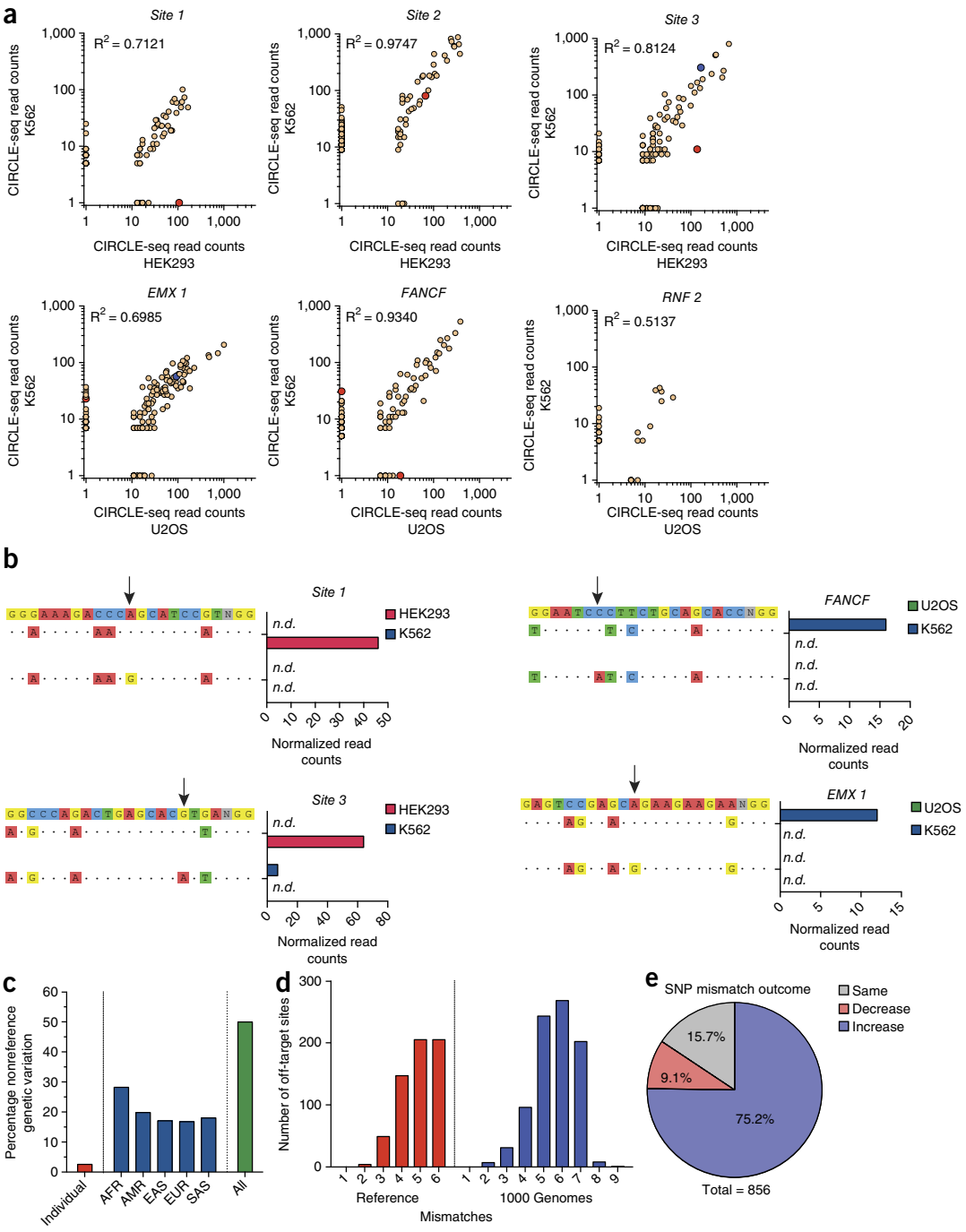

Figure 4 | Using CIRCLE-seq to assess the impacts of personalized SNPs on off-target site analysis. (**a**) Scatterplots of CIRCLE-seq read counts from experiments performed on genomic DNA from two different cell types. Sites with nonreference genetic variation in only one cell type are highlighted in red, while those with nonreference variation in both cell types are highlighted in blue. (**b**) Examples of allele-specific CIRCLE-seq read counts at off-target sites with nonreference genetic variation. Mismatches to the intended target sequence are indicated with colored nucleotides, while matching bases are indicated with a dot. The base position harboring the differential genetic change between cell types is indicated with a small arrow. n.d., not detected. (**c**) Proportion of CIRCLE-seq off-target sites where nonreference genetic variation was identified in genotyped individuals from the 1000 Genomes Project (AFR, African; AMR, Ad Mixed American; EAS, East Asian; EUR, European; SAS, Southeast Asian superpopulations) and a combined population average. (**d**) Histogram showing distribution of CIRCLE-seq off-target sites by numbers of mismatches in reference human genome sequence (red) and in 1000 Genomes Project data-derived off-target site haplotypes (blue). (**e**) Proportion of 1000 Genomes Project-derived haplotypes with increased (blue), decreased (red), or same (gray) numbers of mismatches in off-target sites identified by CIRCLE-seq.

are more closely matched to the intended sgRNA target sequence than the corresponding off-target site in the reference genome, suggesting that individuals with these particular genetic variations would have a higher risk of off-target cleavage at those sites (**Fig. 4e**). Taken together, these results highlight the importance of considering individual genetic variation and illustrate how CIRCLE-seq might be used in future studies to produce personalized genome-wide off-target profiles.

DISCUSSION

CIRCLE-seq has a substantially reduced rate of observed background reads, enabling it to sensitively identify off-target sites using a small fraction (~1.7%) of the total sequencing reads used with existing *in vitro* methods. Requiring only 4–5 million reads, the method is accessible to most labs and might be amenable to automation and scaling, particularly if the relatively large amount of genomic DNA required for each CIRCLE-seq reaction (25 µg; **Supplementary Protocol**) can be reduced.

CIRCLE-seq might enable the production of larger data sets for training of more accurate predictive algorithms for Cas9 off-target determination. This will require that the method be cost effectively automated and scaled to identify off-target cleavage sites and relative *in vitro* cleavage efficiencies for thousands of sgRNAs. Furthermore, analysis of CIRCLE-seq data coupled with large-scale cell-based off-target profiles obtained in ENCODE-characterized cell lines[38] with methods such as GUIDE-seq may make it possible to better understand the impact of chromatin accessibility and epigenetic modifications on the ability of nucleases to induce cellular DSBs. For example, an initial analysis we performed using publicly available DNase-seq data sets for U2OS cells indicated that CIRCLE-seq sites were significantly more likely to be detected by GUIDE-seq or targeted tag sequencing if located in DNaseI hypersensitive regions (P = 0.000017, odds ratio = 2.04[1.49–2.82]). However, because this analysis was only performed with a small number of sites and on only a single human cancer cell line, additional studies will be needed to further define this relationship.

For both routine and therapeutic applications of genome editing, because CIRCLE-seq is more sensitive than cell-based genome-wide off-target detection methods, we envision that CIRCLE-seq might be used as an initial screen to identify potential off-target sites that can then be verified with an orthogonal approach in nuclease-modified cells in culture or *in vivo*. In this study, we used targeted sequencing to search for GUIDE-seq dsODN tags to validate low-frequency sites (<0.1%) that would be challenging to identify by standard amplicon sequencing on account of the indel error rate associated with next-generation sequencing (typically ~0.1%). However, this approach is limited to cells that can be transfected with the GUIDE-seq dsODN tag. Thus, an important area for future work will be the development of alternate methods to more sensitively measure off-target mutagenesis in cells below the error rate of current high-throughput sequencing technologies. Currently, the true false-positive rate for CIRCLE-seq may be challenging to estimate because of limited detection sensitivity due to high error rates of current technologies for orthogonal validation.

Our CIRCLE-seq results provide greater support for the concept that human genetic variation can affect off-target cleavage[35]. Our findings illustrate the importance of considering individual genotypes when evaluating off-target risk and suggest that safety assessments of nucleases might ultimately include patient-specific genome-wide activity profiling. Alternatively, CIRCLE-seq performed on genomic DNA isolated from a large number of genetically diverse individuals may provide an effective strategy to define the vast majority of common and SNP-specific off-target effects for any given nuclease.

Finally, CIRCLE-seq could play an important role in further improvement of CRISPR–Cas nucleases. We recently described an engineered high-fidelity variant, SpCas9-HF1 (ref. 39), which for sgRNAs targeted to nonrepetitive sites routinely failed to show evidence of off-target effects as judged by GUIDE-seq. A key question for future studies is whether CIRCLE-seq can identify low-frequency SpCas9-HF1 off-target mutations that might fall below the detection limit of GUIDE-seq. If such off-target sites were found, we envision that beyond providing an important potential assay for therapeutic applications, CIRCLE-seq would also play a major role in enabling the continued refinement of approaches to improve CRISPR–Cas9 genome-wide specificities.

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

This work was supported by a National Institutes of Health (NIH) Director's Pioneer Award (DP1 GM105378), NIH R35 GM118158, and NIH R01 GM107427 (to J.K.J.); and the Jim and Ann Orr Research Scholar Award (to J.K.J.).

AUTHOR CONTRIBUTIONS

S.Q.T. and J.K.J. conceived of and designed experiments. S.Q.T. and N.T.N. performed all experiments. S.Q.T., J.M.-L., V.V.T., and M.J.A. wrote the CIRCLE-seq analysis pipeline and analyzed CIRCLE-seq data. S.Q.T. and J.K.J. wrote the manuscript with input from all authors.

COMPETING FINANCIAL INTERESTS

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

1. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
2. Hwang, W.Y. *et al.* Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nat. Biotechnol.* **31**, 227–229 (2013).
3. Jinek, M. *et al.* RNA-programmed genome editing in human cells. *eLife* **2**, e00471 (2013).
4. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
5. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
6. Doudna, J.A. & Charpentier, E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science* **346**, 1258096 (2014).
7. Bolukbasi, M.F., Gupta, A. & Wolfe, S.A. Creating and evaluating accurate CRISPR-Cas9 scalpels for genomic surgery. *Nat. Methods* **13**, 41–50 (2016).
8. Mali, P., Esvelt, K.M. & Church, G.M. Cas9 as a versatile tool for engineering biology. *Nat. Methods* **10**, 957–963 (2013).
9. Sander, J.D. & Joung, J.K. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat. Biotechnol.* **32**, 347–355 (2014).
10. Maeder, M.L. & Gersbach, C.A. Genome-editing technologies for gene and cell therapy. *Mol. Ther.* **24**, 430–446 (2016).

11. Lin, J. & Musunuru, K. Genome engineering tools for building cellular models of disease. *FEBS J.* **283**, 3222–3231 (2016).
12. Hsu, P.D., Lander, E.S. & Zhang, F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **157**, 1262–1278 (2014).
13. Brandsma, I. & Gent, D.C. Pathway choice in DNA double strand break repair: observations of a balancing act. *Genome Integr.* **3**, 9 (2012).
14. Symington, L.S. & Gautier, J. Double-strand break end resection and repair pathway choice. *Annu. Rev. Genet.* **45**, 247–271 (2011).
15. Kass, E.M. & Jasin, M. Collaboration and competition between DNA double-strand break repair pathways. *FEBS Lett.* **584**, 3703–3708 (2010).
16. Wyman, C. & Kanaar, R. DNA double-strand break repair: all's well that ends well. *Annu. Rev. Genet.* **40**, 363–383 (2006).
17. Rouet, P., Smih, F. & Jasin, M. Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. *Mol. Cell. Biol.* **14**, 8096–8106 (1994).
18. Sternberg, S.H., LaFrance, B., Kaplan, M. & Doudna, J.A. Conformational control of DNA target cleavage by CRISPR-Cas9. *Nature* **527**, 110–113 (2015).
19. Kiani, S. *et al.* Cas9 gRNA engineering for genome editing, activation and repression. *Nat. Methods* **12**, 1051–1054 (2015).
20. Dahlman, J.E. *et al.* Orthogonal gene knockout and activation with a catalytically active Cas9 nuclease. *Nat. Biotechnol.* **33**, 1159–1161 (2015).
21. Fu, Y., Sander, J.D., Reyon, D., Cascio, V.M. & Joung, J.K. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat. Biotechnol.* **32**, 279–284 (2014).
22. 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).
23. Shah, S.A., Erdmann, S., Mojica, F.J.M. & Garrett, R.A. Protospacer recognition motifs: mixed identities and functional diversity. *RNA Biol.* **10**, 891–899 (2013).
24. Tsai, S.Q. & Joung, J.K. Defining and improving the genome-wide specificities of CRISPR–Cas9 nucleases. *Nat. Rev. Genet.* **17**, 300–312 (2016).
25. Gori, J.L. *et al.* Delivery and specificity of CRISPR-Cas9 genome editing technologies for human gene therapy. *Hum. Gene Ther.* **26**, 443–451 (2015).
26. Cox, D.B.T., Platt, R.J. & Zhang, F. Therapeutic genome editing: prospects and challenges. *Nat. Med.* **21**, 121–131 (2015).
27. Gabriel, R. *et al.* An unbiased genome-wide analysis of zinc-finger nuclease specificity. *Nat. Biotechnol.* **29**, 816–823 (2011).
28. Ran, F.A. *et al.* *In vivo* genome editing using *Staphylococcus aureus* Cas9. *Nature* **520**, 186–191 (2015).
29. 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).
30. Frock, R.L. *et al.* Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases. *Nat. Biotechnol.* **33**, 179–186 (2015).
31. Crosetto, N. *et al.* Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nat. Methods* **10**, 361–365 (2013).
32. Pattanayak, V. *et al.* High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat. Biotechnol.* **31**, 839–843 (2013).
33. 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).
34. 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).
35. Yang, L. *et al.* Targeted and genome-wide sequencing reveal single nucleotide variations impacting specificity of Cas9 in human stem cells. *Nat. Commun.* **5**, 5507 (2014).
36. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
37. Sherry, S.T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
38. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
39. Kleinstiver, B.P. *et al.* High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490–495 (2016).

ONLINE METHODS

A step-by-step protocol is available as a **Supplementary Protocol** and as an open resource in *Protocol Exchange*[40].

Cell culture and transfection. Cell culture experiments were performed on human U2OS (gift from T. Cathomen, University of Freiburg), HEK293 (Thermo-Fisher), K562, and PGP1 fibroblast cells (gift from G. Church, Harvard Medical School). U2OS and HEK293 cells were cultured in Advanced DMEM (Life Technologies) supplemented with 10% FBS, 2 mM GlutaMax (Life Technologies) and penicillin–streptomycin at 37 °C with 5% CO_2. K562 cells were cultured in RPMI 1640 (Life Technologies) supplemented with 10% FBS, 2 mM GlutaMax and penicillin–streptomycin at 37 °C with 5% CO_2. Human PGP1 fibroblasts were cultured in Eagle's DMEM (ATCC) with 10% FBS, 2 mM GlutaMax and penicillin–streptomycin at 37 °C with 5% CO_2. For CIRCLE-seq experiments, genomic DNA was isolated using Gentra Puregene Tissue Kit (Qiagen) and quantified by Qubit (Thermo Fisher). For targeted tag-integration deep-sequencing experiments, U2OS cells (program DN-100), HEK293 cells (program CM-137), and K562 cells (program FF-120) were transfected in 20 µl Solution s.e.m. (Lonza) on a Lonza Nucleofector 4-D according to the manufacturer's instructions. In U2OS cells, 500 ng of pCAG-Cas9 (pSQT817), 250 ng of sgRNA encoding plasmids, and 100 pmol of GUIDE-seq end-protected dsODN were cotransfected. Genomic DNA for targeted tag integration sequencing was harvested approximately 72 h post-transfection using the Agencourt DNAdvanced Genomic DNA Isolation Kit (Beckman Coulter Genomics).

***In vitro* transcription of sgRNAs.** Annealed oligonucleotides containing sgRNA target sites were cloned into plasmid NW59 containing a T7 RNA polymerase promoter site. The sgRNA expression plasmid was linearized with HindIII restriction enzyme (NEB) and purified with MinElute PCR Purification Kit (Qiagen). The linearized plasmid was used as DNA template for *in vitro* transcription of the sgRNA using MEGAshortscript Kit according to the manufacturer's instructions (Thermo-Fisher).

CIRCLE-seq library preparation. For the experiments with sgRNAs previously evaluated by GUIDE-seq, CIRCLE-seq experiments were performed on genomic DNA from the same cells in which they were evaluated by GUIDE-seq (either U2OS or HEK293 cells). Purified genomic DNA was sheared with a Covaris S200 instrument to an average length of 300 bp, end repaired, A tailed, and ligated to uracil-containing stem-loop adaptor oSQT1288 5′-P-CGGTGGACCGATGATCUATCGGTCCACC G*T-3′, where * indicates phosphorothioate linkage. Adaptor-ligated DNA was treated with a mixture of Lambda Exonuclease (NEB) and *E. coli* Exonuclease I (NEB), then with USER enzyme (NEB) and T4 polynucleotide kinase (NEB). DNA was circularized at 5 ng/µl concentration with T4 DNA ligase, and treated with Plasmid-Safe ATP-dependent DNase (Epicentre) to degrade remaining linear DNA molecules. *In vitro* cleavage reactions were performed in a 100 µl volume with Cas9 nuclease buffer (NEB), 90 nM SpCas9 protein (NEB), 90 nM *in vitro* transcribed sgRNA, and 250 ng of Plasmid-Safe-treated circularized DNA. A concentration of SpCas9 and gRNA was chosen that could cleave an amplicon containing the intended target sequence to near completion (**Supplementary Fig. 12**). Digested products were A tailed, ligated with a hairpin adaptor, treated with USER enzyme (NEB), and amplified by PCR using Kapa HiFi polymerase (Kapa Biosystems). Completed libraries were quantified by droplet digital PCR (Bio-Rad) and sequenced with 150 bp paired-end reads on an Illumina MiSeq instrument. A detailed user protocol for CIRCLE-seq library construction is provided (**Supplementary Protocol**).

Targeted deep sequencing. U2OS cells were transfected with Cas9 and sgRNA expression plasmids in addition to the GUIDE-seq dsODN as described above. Off-targets sites identified by CIRCLE-seq were amplified from the isolated U2OS genomic DNA using Phusion Hot Start Flex DNA polymerase (New England Biolabs) with primers listed in **Supplementary Table 6**. Triplicates of PCR products were generated from each transfection condition with 100 ng of genomic DNA as the input for each PCR. PCR products were normalized in concentration, pooled into different libraries corresponding to different transfection conditions, and purified with Ampure XP magnetic beads (Agencourt). Illumina Tru-seq deep-sequencing libraries were constructed using 500 ng of each pooled sample (KAPA Biosystems), quantified by real-time PCR (KAPA Biosystems), and sequenced on an Illumina MiSeq instrument.

Analysis of targeted tag integration sequencing data for single-nucleotide variants was performed using MuTect2 via GATK version 3.7 and Picard 2.9.0. We visually verified the sites called by Mutect2 in IGV.

CIRCLE-seq data analysis. Paired-end reads were merged and then mapped using bwa[41] mem with default parameters. The start mapping positions of reads that map in the expected orientation with mapping quality ≥50 were tabulated, and genomic intervals enriched in nuclease-treated samples were identified. The interval and 20 bp of flanking reference sequence on either side were searched for potential nuclease-induced off-target sites with an edit distance of less than or equal to six, allowing for gaps.

samtools[42,43] mpileup was used to discover nonreference genetic variation in identified off-target sites. Positions with average quality score greater than 20 were considered possible variants and confirmed by visual inspection (**Supplementary Table 3**).

Reference-independent discovery of off-target cleavage sites was performed by reverse complementing the sequence of one read of a pair and concatenating it with the other. An interval of starting 20 bp on either side of the junction was directly searched for potential off-target cleavage sites with edit distance of ≤6 allowing for gaps and read counts corresponding to identified sites were tabulated.

CIRCLE-seq open-source analysis software. To enable the broad use of CIRCLE-seq for genome-wide detection of nuclease off-target sites, we developed a freely available, open-source Python package circleseq for the analysis of CIRCLE-seq experimental data. Provided with a simple sample manifest, the circleseq software performs full end-to-end analysis of CIRCLE-seq sequencing data with a single command and returns tables of candidate off-target cleavage site positions, as well as visual alignments of off-target sequences. Source code and running instructions will be made freely available online (https://github.com/tsailabSJ/circleseq).

Digenome-seq data analysis. Read counts of mapping positions in a narrow window (±3 bp) around cleavage sites identified by CIRCLE-seq were tabulated from original Digenome-seq sequencing alignments.

Chromatin accessibility analysis. To determine if sites identified by both CIRCLE-Seq and GUIDE-seq/targeted tag sequencing were associated with chromatin accessibility, we used DNase-seq data for U2OS cells (GEO sample GSM2341641). We ran a Cochran–Mantel–Haenzel test stratified by the mismatch number (3:5), where the response is whether or not the CIRCLE-Seq site was also called by GUIDE-Seq, and the predictor is the categorical variable for chromatin accessibility as inferred from DNaseI hypersensitivity ('open' or 'closed'). We considered all the sites found by CIRCLE-Seq in U2OS cells whose mismatch with the on-target site ranged from 3 to 5 (as we did not have enough sites for other class of mismatches).

Statistics. An empirical read count distribution was used to determine statistical enrichment of CIRCLE-seq read counts. For analysis of Digenome-seq data, significant evidence of cleavage at a 0.01 significance level was evaluated by fitting a negative binomial distribution, and statistically significant sites by this criteria were included in **Figure 2b**.

Software availability. Freely available, open-source software for analysis of CIRCLE-seq data can be obtained at: https://github.com/tsailabSJ/circleseq.

Data availability statement. Data supporting this work are available in supporting figure tables and supplementary information. High-throughput sequencing information associated with this study will be made available through NCBI SRA accession number SRP103697. Source data files for **Figures 2–4** are available online.

40. Tsai, S.Q. & Joung, J.K. CIRCLE-seq protocol. *Protocol Exchange*. http://dx.doi.org/10.1038/protex.2017.047 (2017).
41. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
42. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
43. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

doi:10.1038/nmeth.4278