Massively Targeted Evaluation of Therapeutic CRISPR Off-Targets in Cells

Xiaoguang Pan
University of Chinese Academy of Sciences
Kunli Qu
Aarhus University
Hao Yuan
BGI-Shenzhen, Shenzhen, China
Xi Xiang
Aarhus University
Christian Anthon
University of Copenhagen
Liubov Pashkova
Copenhagen University https://orcid.org/0000-0002-7195-9329
Xue Liang
BGI-Qingdao
Peng Han
Beijing Genomics Institute https://orcid.org/0000-0002-3405-087X
Giulia Corsi
Center for non-coding RNA in Technology and Health, IVH, University of Copenhagen
https://orcid.org/0000-0001-5932-0664
Fengping Xu
BGI Research
Ping Liu
MGI, BGI-Shenzhen
Jiayan Zhong
University of Chinese Academy of Sciences
Yan Zhou
Aarhus University
Tao Ma
Beijing Genomics Institute
Hui Jiang
Junnian Liu
BGI-Shenzhen
Jian Wang

China National Genebank
Niels Jessen
Department of Biomedicine, Aarhus University https://orcid.org/0000-0001-5613-7274
Lars Bolund
Danish Regenerative Engineering Alliance for Medicine (DREAM), Department of Biomedicine, Aarhus University
Huanming Yang
BGI-Shenzhen
Xun Xu
BGI-Shenzhen, Shenzhen 518083, China https://orcid.org/0000-0002-5338-5173
George Church
Harvard University
Jan Gorodkin
University of Copenhagen
Lin Lin
Department of Biomedicine, Aarhus University https://orcid.org/0000-0002-7546-4948
Yonglun Luo (alun@biomed.au.dk)
Aarhus University https://orcid.org/0000-0002-0007-7759

Article

Keywords: Gene Editing, Genome Engineering, Gene Therapy, High-throughput Methods, Off-targets

Posted Date: March 17th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1427273/v1

Version of Record: A version of this preprint was published at Nature Communications on July 13th, 2022. See the published version at https://doi.org/10.1038/s41467-022-31543-6.

Abstract

Sensitive and high-throughput evaluation of CRISPR RNA-guided nucleases (RGNs) off-targets (OTs) in cells are essential for safety assessment and advancing RGN-based gene therapies. Here we apply a modified library approach, SURRO-seq, for simultaneously evaluating thousands of off-target sites for therapeutic RGNs in cells. The SURRO-seq captures RGN-induced indels in barcoded surrogate off-target sites in cells by a pool of lentiviral vectors and targeted deep sequencing. We first evaluate 170 previously investigated OTs from 11 RGNs with SURRO-seq in HEK293T cells. SURRO-seq captures nearly 100% OTs found by T7E1, most (70%) GUIDEseq-identified OTs, and approximately half (51%) OTs found by CIRCLE-seq. We then apply SURRO-seq to evaluate 7140 OTs from 110 therapeutic RGNs. 105 RGNs are found to introduce significantly detectable indels in 753 OTs, of which 180 OTs have an indel frequency above 3%. Notably, significantly detectable off-target indels are identified in 37 cancer genes including tumor suppressor gene exons. We deep sequence 23 endogenous genome loci in five human cell lines and further validate that SURRO-seq can sensitively capture off-targets with indel frequency below 0.1%. High fidelity RGNs have substantially reduced indel rates introduced at OTs but significantly detectable indels are still found in some OTs. Analyses of OTs with significantly detectable indels indicates that thermodynamically stable wobble base pair (rG•dT) and free binding energy strongly affect RGN specificity. Our study emphasizes the necessity of carefully selecting high fidelity RGNs and evaluating therapeutic RGN to minimize inevitable off-targets.

Introduction

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) RNA-guided nucleases (RGNs) has been used in therapy of several inherited human diseases [1–4]. Major efforts have focused on improving RGN editing efficiency via stabilization of the small guide RNA (sgRNA) thermodynamics [5], modification of the RGNs [6–8], utilization of homology-independent mediated targeted integration (HITI) [9] and optimization of RGN delivery [10,11]. The inevitably adverse effects caused by unspecific RGN editing of cancer genes is a major concern for the clinical application of RGN-based therapies. Improvement of RGN specificity and development of methods for identifying and evaluating the potential off-targets (OT) introduced by RGNs are equally essential to advance RGN-based gene therapy. A number of experimental RGN OT identification/quantification methods have been developed (**Supplementary Data 1**), which can be grouped into three categories (**Supplementary Figure S1**). Category One contains genome-wide cell-free biochemical methods which relies on the capture of RGN-induced OT cleavage on either naked DNA or fixed chromatin fibers by sequencing. Examples are CIRCLE-seq (cell-free) [12], Digenome-seq (cell-free) [13], SITE-seq (cell-free) [14], BLISS (ex vivo) [15] and DIG-seq (ex vivo) [16]. Category Two contains methods depending on genome-wide in-cell capturing of RGN-induced off-target cleavage by sequencing, such as GUIDE-seq and IDLV-capture relying on insertion of double strand DNA and IDLV vector to the DNA double strand breaks respectively [17,18], HTGTS and PEM-seq relying on translocation between on-target and off-targets [19,20], and DISCOVER-seq relying on immunoprecipitation of DNA repair protein MRE11 to capture the DNA double strand break (DSB) sites [21]. While cell-free biochemical methods are rapid, conventional

and not depending on reference genomes, they inevitably capture many pseudo off-target sites. In-cell methods (e.g., GUIDE-seq) capture the bona fide RGN off-targets more faithfully as compared to cell-free methods. However, spontaneous DSBs lead to capturing pseudo off-targets independent of RGNs [17]. To complement this, Category Three is composed of targeted in-cell RGN OT validation methods, such as T7E1, targeted deep sequencing, TIDE and CUT-PCR [22, 23]. However, current targeted in-cell RGN off-target evaluation methods are greatly limited by their scales. Only a few sites can be evaluated for each RGN in a single study due to their high labor and time cost.

Here we introduce and apply SURRO-seq, a high-throughput method for targeted in-cell capture of RGN off-targets based on a pool of lentivirus vectors encoding gRNA and barcoded surrogate off-target sites, to targeted evaluate therapeutic RGN off-target in cells. SURRO-seq exhibits high sentificity and accuracy compared to GUIDE-seq and CIRCLE-seq by evaluating 170 previously investigated OTs from 11 RGNs in HEK293T cells. We then applied SURRO-seq to evaluate 7140 OTs from 110 therapeutic RGNs and identify 753 OTs showing significantly detectable indels. 37 OTs with significantly detectable indels are found in cancer genes, highlight the clinical significance and great need of pre-assessing RGN OTs with SURRO-seq. The SURRO-seq identified OTs were further validated by targeted deep sequencing of five RGN-edited human cell lines. Analyses of OTs with high indel frequencies revealed that mismatch types leading to thermodynamically stable wobble base pair strongly increase RGN OT effect. We further perform benchmark analyses of latest RGN OT prediction tools with SURRO-seq OT data. The energy-based predictors, which incorporate gRNA and DNA binding energies, give the best performance.

Results

Design of SURRO-seq

Libraries of surrogate vectors have been used in many studies to massively capture on-target efficiencies[24, 25]. We and others show that single surrogate episomal vector (or genomic integration site) has been used as a sensitive method to measure RGN off-target activity. But the method is only applicable for evaluating a limited number of RGN OTs[7, 26]. Previously, we introduced an optimized high-throughput approach for targeted in-cell evaluation of on-target RGN efficiency using a pool of lentiviral surrogate vectors[27]. Here we introduce site specific barcoding and repurpose the method for high throughput targeted evaluation of RGN off-targets (OTs) in cells. For a given RGN, protospacer sequences of all OTs are very similar and only differ for 1–5 nucleotides (nt). Following RGN editing, deletion indels could erase nucleotides that differ between OTs, making it impossible to uniquely assign the deletion indels to the OTs (**Supplementary Figure S2**). To overcome the indel split problem, we introduced a 10-nt barcoding strategy to distinguish indels reads in the ON and OT sites (**Supplementary Figure S2**). As showed in Fig. 1, SURRO-seq contains nine major steps (**see Supplementary Note 1 for extended description of the method**) with three modifications compared to our previous on-target method CRISPRon[27]: (1) The surrogate site contains a 10-nt barcode preceding the 27-nt surrogate OT site, which contains the OT protospacer (20 nt), protospacer adjacent motif (PAM, NGG), 4-nt PAM downstream

sequences; (2) Barcode-guided split of indel reads (Supplementary Figure S2, S3); and (3) Fishers' exact test of OTs with significant indels [a. Two-fold higher indel frequency in the SpCas9 cells as compared to the wild type cells (MOCK); b. Fishers exact test and Benjamini and Hochberg (BH)-adjusted p-value less than 0.05] (**Supplementary Note 2**).

Validation of previously evaluated RGN OTs with SURRO-seq

First, we sought to assess if SURRO-seq can capture RGN OTs previously evaluated by other methods. We generated a small library (LibA) containing 170 OTs from 11 RGNs (Fig. 2a). These 11 RGNs and 170 OTs had been detected by T7E1[28, 29], GUIDE-seq[17] and/or CIRCLE-seq[12]. We transduced SpCas9-overexpressing (SpCas9) and wildtype (MOCK) HEK293T cells with LibA (MOI = 0.3, and 4000-fold coverage of LibA, see methods). Eight days after LibA transduction, indel frequencies introduced in the surrogate OTs were quantified by targeted deep sequencing. Analyses of LibA data (**Supplementary Figure S3-5**, Fig. 2b, **Supplementary Data 2**) showed that SURRO-seq can validate nearly 100% of the T7E1-detected OTs (22 out of 23, Fig. 2b), most (104 out of 149) of GUIDE-seq-captured OTs (Fig. 2c), and approximately half (78 out of 153) CIRCLE-seq-captured OTs (Fig. 2d). We sought to compare the false discovery rate (FDR) of OT identification between GUIDE-seq, CIRCLE-seq and SURRO-seq, reasoning that bona fide RGN OTs should be captured by at least two independent methods. Analyses of the 141 OTs from 5 RGNs (Fig. 2e) showed that there is only one SURRO-seq-identified OT that cannot be captured by the other two methods, while there are 40 and 17 OTs respectively captured by CIRCLE-seq and GUIDE-seq only. The existence of false positive RGN OT identified by GUIDE-seq and CIRCLE-seq has been documented and acknowledged earlier [12, 13, 17] (**Supplementary Figure S1**). The high accuracy of detecting bona fide RGN OTs by SURRO-seq can be explained by SURRO-seq's dependency on direct indel quantification. SURRO-seq thus offers a complementary in-cell method for targeted validation of RGN OTs identified by genome-wide screening approaches.

Large-scale evaluation of therapeutic RGNs OTs with SURRO-seq

To investigate if SURRO-seq can be used for high throughput targeted in-cell evaluation of RGN OTs, we selected 110 RGNs targeting 21 human genes that have been used in preclinical gene therapy studies (**Supplementary Data 3**). Given that most RGNs can only tolerate a few (1–3) mismatches between the sgRNA spacer and the protospacer sequences[30–32], we blasted the human reference genome with the 20-nt spacer of each RGN, and retrieved all potential OTs with up to 4 mismatches. In total, 8150 OTs from these 110 RGNs were selected and synthesized, cloned into the SURRO-seq vector and packaged into lentivirus, hereafter referred to as library B (LibB) (Fig. 3a). We transduced SpCas9 and wildtype (MOCK) HEK293T cells with LibB (MOI = 0.3, 4000-fold coverage) and analyzed indel frequencies from cells eight days after transduction by deep sequencing.

We first analyzed the on-target gRNA efficiency of these 110 RGNs. SURRO-seq successfully captures on-target efficiencies for all 110 RGNs (Fig. 3b). The SpCas9 protein is overexpressed in the HEK293T cells by doxycycline addition. Consistent with our previous observation[27], most (n = 96) RGNs exhibited high on-target activity (indel frequencies% (IF%) > 80%, Fig. 3b). A few RGNs (n = 14) had relatively low efficiency (IF% < 80%), and these were also significantly ($p < 0.0001$) lower in GC content compared to highly efficient RGNs (IF% > 80%) (**Supplementary Figure S6**). Next, we analyzed indel frequencies in the OTs introduced by RGNs. Surrogate OT sites with low sequencing quality (total clean reads < 32 for both MOCK and SpCas9), low synthetic quality (IF% in MOCK > 4%, **Supplementary Figure S7**), or possibly affecting cell proliferation (**Supplementary Figure S8**, fold change in enrichment or depletion (MOCK vs. SpCas9) > 2) were filtered out. After filtering, 7140 OTs were retained for downstream analyses. Quantification of indel frequencies in each OT in SpCas9 and MOCK cells showed that there were not significantly detectable indels for most of these OTs (n = 6387, hereafter referred to as NSOT). Significantly detectable indels (fold-change (FC of IF% SpCas9 / IF% MOCK) > 2, adj. p-value < 0.05) were identified for 753 OTs in SpCas9 cells compared to MOCK (Fig. 3c, **Supplementary Data 3**). However, the indel frequency of most Sig. OTs were less than 3% (573 out of 753, Fig. 3c and 3d). We further divided the Sig. OTs into two groups: based on IF% in SpCas9 < 3% (Low Indel Sig. OTs, hereafter referred to as LIOT) and IF% >= 3% (High Indel Sig. OTs, hereafter referred to as HIOT). Notably, most HIOTs contain 1–3 mismatches (**Supplementary Figures S9**). Our results demonstrate that the SURRO-seq can be used for high throughput targeted evaluation of RGN-induced indels at surrogate off target sites in cells.

To investigate where were these LIOTs and HIOTs located in the genome and in genes, we annotated their genomic locations according to the presence in intergenic region (IGR) or in genes (2kb upstream, 5' untranslated region, exon, intron, 3' untranslated region, 2kb downstream; **Supplementary Data 3**). Despite that most of the Sig. OTs are found in intron and IGRs, there are still a substantial number of Sig. OTs (nr. of LIOT = 200, nr. of HIOT = 87) found in gene exons and/or regulatory regions that might affect gene expression (Fig. 3d, **Supplementary Figure S10**). Notably, one RGN11153 (spacer sequences, CTGCTGCTGCTGCTGCTGGA), which was proposed for Huntington's Disease therapy by targeting the CAG expansion tract[33], exhibit great off-target effect (nr. of LIOT = 43, nr. of HIOT = 35). Two HIOTs of RGN11153 are found in cancer genes *ZFHX3* (exon) and *SOHLH2* (intron). The zinc-finger homeobox 3 (*ZFHX3*) is a tumor suppressor gene and knockout of *ZFHX3* in mouse leads to development of neoplastic lesions. Loss of function mutations in *ZFHX3* are frequently detected in human cancers i.e. high-grade human prostate cancers [34], endometrial cancers[35], urothelial bladder carcinoma[36], lung and brain tumors[37]. This finding emphasizes that carefully evaluating if therapeutic RGNs introduced any off-target indels in cancer genes is need. HIOTs were also found in another two cancer genes BCOR (intron) and NCOR2 (intron) by RGN11155 and RGN11189 respectively. These two RGNs were used for HD (RGN11155) and β-Thalassemia (RGN11189) therapy[38]. For the LIOTs, despite low indel frequency (below 3%), significantly detectable indels were found in the exon of nine cancer genes, such as U2AF2 and NKTR causing Acute myeloid leukemia (AML) (**Supplementary Data 3 and Figure S10**). SURRO-seq thus offers a high throughput and targeted in cell evaluation of RGN OTs, particularly in the detection of very low indel frequency OTs that could not be captured by other methods.

Validation of SURRO-seq identified OTs by targeted deep sequencing of endogenous genomic loci

To validate if OTs captured by SURRO-seq were also presented at the corresponding endogenous sites, we analyzed 23 OTs from seven RGNs in five human cell lines: human embryonic kidney cells (HEK293T), human primary fibroblasts, lung cancer cells (PC-9), ovarian cancer cells (SKOV3), and bone osteosarcoma epithelial cells (U2OS). Of these 23 OTs, 16 and 7 OTs were detected with significant and non-significant indels by SURRO-seq, respectively (Fig. 4a **and Supplementary Data 4**). Instead of using lentivirus-based delivery of CRISPRs, we applied an optimized CRISPR delivery approach based on CRISPR ribonucleoprotein (RNP). Highly efficient delivery of CRISPR into various types of cells have been reports by us and many other groupshig[2, 5, 39, 40]. We also validated that an enhance green fluorescent protein (*EGFP*) mRNA can be delivered to nearly 100% of cells in all five cell lines (**Supplementary Figure S11**). Seven on-target and 23 off-target endogenous genomic loci from the five cell lines were analyzed by targeted deep sequencing 48 hours after RNP treatments (Fig. 4b). Several Cas9 mutants have been reported with improved specificity[41–43]. In addition to the wild type SpCas9, we also analyzed the indel frequencies in the 23 off-targets and 7 RGN on-target sites in cells transfected with a high-fidelity Cas9 variant (HiFi-Cas9), of which a single point mutation (p.R691A) was introduced[41].

Analyses of deep sequencing results showed that significant indel frequencies (one-way ANOVA, $p < 0.05$) were detected in all seven RGN on-target sites in the five cell lines (Fig. 4c, **Supplementary Figure S12, Supplementary Data 4**). Consistently with early results, the high-fidelity HiFi-Cas9 retained similarly high on-target activity as the wild type Cas9 (Fig. 4c), except RGN11208 which is low in GC content (GC% = 20%, **Supplementary Figure S13**). Analyses of indel frequencies in the 23 off-target sites showed that there is a good agreement (20 out of 23, 87%) between SURRO-seq and targeted sequencing of endogenous sites. 15 out of 17 (88%) of the SURRO-seq off-target sites with significantly detectable indels were validated by targeted deep sequencing of in Cas9 RNP treated cells (Fig. 4d-e, **Supplementary Data 4**). Due to differences in RGN delivery, editing time and RGN expression level between SURRO-seq and RNP nucleofection, the indel frequencies from the endogenous off-target sites are much lower than that measured by SURRO-seq (Fig. 4e, **Supplementary Data 4**). However, the nature that SURRO-seq can amplify the indel frequency introduced at the surrogate off-target sites offers a great technical advantage for detecting true endogenous off-target sites with very low indel rates. For example, the endogenous indel frequencies of OT5679 and OT5675 introduced by RGN111189 were below 0.1% in some cell lines tested. Such a low indel frequency could not be significantly by methods such as T7E1 or Sanger sequencing, while SURRO-seq can amplify the indel frequency up to 4% (Fig. 4e). Most importantly, both the number of OTs with significantly detectable indels and the indel frequency were significantly reduced in cells edited with the high-fidelity HiFi-Cas9, which corroborates with previous finding and highlights the importance and necessity of using high-fidelity Cas9 mutations in gene therapy application to minimize the off-target effect [41]. Collectively, we demonstrated that SURRO-seq is a sensitive method for high throughput targeted evaluation of RGN off-targets in cells.

Effect of mismatch positions, mismatch types, and free binding energy on RGN specificity

The RGN off-target data generated by SURRO-seq also allow us to explore how the genomic context affects RGN off-target cleavage. Analysis of indel frequencies of the 753 Sig. OTs (both LIOT and HIOT) showed that indel frequencies were significantly decreased in OTs with more mismatches (Fig. 5a, **Supplementary Figure S10**), which corroborated with previous findings and is expected[30, 44, 45]. While it is generally believed that OTs with 3–4 mismatches are unlikely to be cleaved by RGNs, our results showed that there exists a great heterogeneity in mismatch tolerance among the 110 RGNs and between the OTs with same number of mismatches (Fig. 5a). This phenomenon was also observed for the VEGFA-T2 RGN[17] and validated by SURRO-seq (Fig. 2). We speculated that the heterogeneity of indel frequencies between OTs with same number of mismatches in our dataset is caused by the positions and types of mismatches between the RGN spacer and the target site. It has been well characterized that the CRISPR is less tolerated to mismatches in the PAM-proximal 10–12 nucleotides[46–48].

To address if this position-dependent mismatch tolerance contributes to the heterogeneity of OTs, we analyzed the frequency of mismatches occurred in each position of the 20-nt protospacer region for all RGN OTs with 3 or 4 mismatches (**Supplementary Data 5**). There is a significant (Hypermetric test p-value < 0.05) over-representation of mismatches occurred at N1 and N2 positions (the two most PAM-distal nucleotides) and an under-representation of mismatches occurred at the N12-N18 (PAM-proximal seed regions) in OTs with significantly detectable. Interestingly, our analysis also revealed that RGNs seem to be more tolerant to mismatches at the N19 and N20 position as compared to other nucleotides of the seed region **(**Fig. 5b**).**

We next analyzed the effect of mismatch types on RGN OT. Twelve types of mismatches can occur between RGN and off-target sites (**Supplementary Figure S13**). To provide a simple description, we only refer to mismatches between the gRNA spacer and the protospacer sequences (the non-targeting strand). Cumulative studies have suggested that RGN exhibits different tolerances to the different types of mismatches. Once such example is the GA mismatch, which generates a wobble base pair (rG:dT) between gRNA spacer and the complementary strand DNA. We analyzed the frequencies of these 12 mismatch types in the OTs from LibB (**Supplementary Data 5**). Our results showed that GA mismatch (a less degree of AG mismatch, for OTs with 3MM) was significantly (hypergeometric test p-value < 0.05, compared to NS OTs or total OTs) enriched in the OTs with significantly detectable indel (Fig. 5c). To further validate the effects of mismatch and mismatch type on CRISPR specificity, we generated a small SURRO-seq library (libC) carrying artificially generated OTs with all possible combinations of one mismatch for five RNGs (**Supplementary Figure S14**). SURRO-seq-based libC further showed similar findings about the effect of mismatch position and type (GA and AG wobble pairs) on RGN specificity (Fig. 5d-g, **Supplementary Figure S14, Supplementary Data 5**). Notably, the RGN 11157, which is low in GC and particularly G content, seems to be less tolerant of mismatches (**Supplementary Figure S14**).

Since a large number of regression-based, machine-learning and deep-learning models have already been developed for *in silico* prediction of RGN off-targets, we benchmarked six RGN OT scoring models: MIT [49], deepCRISPR [50], Cutting Frequency Determination (CFD) score [44], CROP-IT [51], CCtop [52], and CRISPRoff [53] with the LibB Sig. OTs data. Our results showed that CRISPRoff (Pearson R = 0.50, Spearman R = 0.48, p-value < 0.001) outperforms the other four RGN OT scorers (**Supplementary Figure S15**). Compared to the other OT prediction scorers, the CRISPRoff has included the free energy feature. We hypothesized that the energy features are the main contributing factors to the RGN OT prediction. To prove that we next analyzed the correlation between the OT indel efficiencies and position-weighted binding energy between gRNA and the (off-)target DNA (ΔG_H), the free energy of the DNA duplex (ΔG_O), or the folding energy of the gRNA only (ΔG_U) as defined by the CRISPRoff energy model [53]. Our results showed that there is significant correlation between indel efficiencies of Sig. OTs and the ΔG_H (Pearson's R = 0.53, Spearman's R = 0.52, p-value < 0.001), ΔG_O (Pearson's R = 0.25, Spearman's R = 0.26, p-value < 0.001), and ΔG_U (Pearson's R = 0.23, Spearman R's = 0.30, p-value < 0.001) (**Supplementary Figure S16**). Notably, the feature of gRNA and the (off-)target DNA binding energy (ΔG_H) yields even high correction compared to the seven RGN off-target predicting, corroborating that ΔG_H is the major energy feature determining RGN OT effect [53]. Our data collectively highlighted the importance of mismatch positions (where), mismatch types (which), free binding energy (a combined feature of mismatch positions and types) on RGN off-target effect.

Discussions

In conclusion, we validated and demonstrated that surrogate off-target site-based capturing of RGN cleavage can be used for massively targeted evaluation of SpCas9-based RGN off-targets in cells. Similar to our approach, Fu et al., very recently reported a similar library-based approach of which a pair of on-target and off-target surrogate site was introduced to allow direct comparison of on and off target efficiencies, as well as understanding effect of sequence contexts on RGN specificity[54]. Several generations of CRISPR-derived technologies have successfully reported for gene editing purposes. These include the different classes and types of CRISPR Cas systems and variants, such as SpCas9, SaCas9, NmCas9, Cas-X, Cas-Y, Cas12a, Cas13 (just to mention a few) [48, 55–59]. Most importantly, a large number of CRISPR-Cas9 derived genetic and epigenetic editing tools have been developed by fusing the dead Cas9 (dCas9) protein or nickase Cas9 (nCas9) protein to effector proteins or protein domains. By fusing dCas9 or nCas9 to deaminases, the CRISPR-Cas9 system have been repurposed for targeted base editing [59–62], such as A-to-G substitution (ABE), C-to-T substitution (CBE), C-to-G substitution (GBE). For a comprehensive overview of the CRISPR-derived base editors, we refer readers to the review paper by Porto et al..[63] Off-targets effects have been reported for these CRISPR base editors. Although not showed in this study, we have demonstrated that high throughput quantification of base editing efficiency can also be achieved using such a surrogate library in cells (BioRxiv. https://doi.org/10.1101/2020.05.20.103614). We anticipate that SURRO-seq could be adapted to evaluate off-targets of other DNA editing RGN systems, including prime editing[64]. Unlike genome-wide in-cell or cell-free OT screening methods, SURRO-

seq is limited by its pre-selected potential OTs for evaluation. However, SURRO-seq offers a sound complementary approach to the genome-wide OT screening methods for further high throughput validation of the RGN OTs.

The CRISPR-Cas9 gene editing technology has been in development for a full decade. We still do not completely understand factors affecting its specificity. These specificity-affecting factors include the gRNA-independent binding of the Cas9 protein to DNA, the number/type/position of mismatches between gRNA spacer and the target site, the epigenetic state (DNA methylation and chromatin accessibility), the expression level and duration of the Cas9 protein and gRNA in cells, and the usage of alternative PAMs. Our results suggest that there is a great heterogeneity in term of the specificity among different RGN gRNAs. Corroborating with previous findings [30, 31], CRISPR-Cas9 is less tolerant to mismatches at the seed region (N10-N20). Our data further showed that mismatches at the two upstream PAM proximal position (N19 and N20) were more tolerated than other nucleotides of the seed region. This site-dependent effect could be explained by our recent binding energy model about the effect of sliding PAMs on CRISPR-Cas9 specificity (Corsi G., et al., unpublished results). Indeed, when performing benchmarking of the different CRISPR-Cas9 off-target prediction tools with our data, our results also showed that energy-based predictors out-performed other tools in their accuracy of predicting true off-targets. The energy feature is also in agreement with our finding that Wobble base pair (G-U), which still can provide strong binding between the gRNA and target DNA strand, is tolerated. We therefore recommend the use of energy-based tools for in silico prediction of CRISPR potential off-targets, while future further improvements of their prediction outcome should be achieved with high quality off-target data and the integration of better energy features.

Substantial off-target indels were observed for some OTs evaluated in this study when conducted in cell line expressing high level of the wild type SpCas9 protein. However, the level of indels were significantly reduced when the SpCas9 was transiently expressed in cells by RNP delivery. Most importantly, with high fidelity SpCas9 variant (HiFi-SpCas9)[41], our results showed that near all off-target indels could not be significantly detected. Thus, our results strongly indicate that high fidelity SpCas9 variants should be used to its largest extend to avoid any potential adverse effect caused by off-target cleavage. This is particularly important when the CRISPR-Cas9 technology is used for gene therapy, both ex vivo and in vivo deliveries. One remaining major concern of CRISPR gene therapy is the off-target effect leading to oncogenesis due to off-target in cancer genes. Selection of high-fidelity Cas variants, carefully design of gRNA with less likelihood of introducing off-target indels in cancer genes, and experimentally validate these potential off-target sites RGN-edited cells are important for lowering the risk of detrimental off-targets in clinical application of RGN. While RGN off-target screening methods, such as GUIDE-seq, DISCOVER-seq, SITE-seq and CIRCLE-seq (also see **supplementary Figure S1**) can be used for genome-wide unbiased detection of RGN off-targets, SURRO-seq overcome the unmet need of high throughput and targeted evaluation of RGN OTs in cells. Our method provides the following four methodological advantages: (1) Scalable. The SURRO-seq library can be generated from a few hundred OTs to over 10,000 OTs. Unlike other methods, SURRO-seq can be used to evaluate hundreds of RGNs in cells

simultaneously. (2) Direct evaluation of indels. SURRO-seq directly quantifies the RGN introduced indels at the surrogate off-targets by comparing RGN edited and MOCK cells. (3) High sensitivity. For SURRO-seq, each OT site can be sequenced with a very deep coverage. And direct comparison of indels in RGN and MOCK cells further allow us to sensitively detect OTs with significant indels, and particularly OTs with low indel rate. (4) Clinical significance. SURRO-seq allows us to target evaluate if RGN introduces indels in clinically relevant genes such as cancer genes. However, we also highlight some limitations of SURRO-seq which require further improvement. Each synthetic SURRO-seq oligonucleotide is 170nt. Synthetic errors introduced in the DNA oligonucleotide library could cause dropout of some OTs after data filtering. As epigenetic features (e.g. chromatin accessibility, DNA methylation) and gene activity affect RGN activity[65–67], the surrogate OTs in the SURRO-seq library are randomly inserted in the genome and might not fully capture the indel efficiency introduced in the corresponding endogenous genomic loci. However, since the SURRO-seq system is based on a lentiviral vector which is transcriptionally active (triggered by puromycin selection) in cells and thousands of surrogate OT integration sites are analyzed, SURRO-seq thus amplifies the indel frequencies and allow us to detect off-target indels below 0.1%.

In conclusion, we report a high throughput method for targeted evaluation of CRISPR-Cas9 off-target in cells. The SURRO-seq offers a great complementary method to the existing tools for CRISPR-Cas9 off-target evaluation, off-target data generation, improvement of prediction, understanding of off-target effect, and facilitate the applications of CRISPR-based gene editing tools in clinical applications.

Methods

Cell culture

Human embryonic kidney (HEK293T), primary human skin-derived fibroblasts (Fib), U2OS, SKOV-3, and PC9 cells were cultured in DMEM media containing 10% fetal bovine serum (FBS) and 1% penicillin–streptomycin in a tissue culture incubator at 37 °C with 5% CO_2. PCR mycoplasma detection kit (cat no. PM008) was routinely used to test the mycoplasma contamination. The cells used in this study have given negative results in mycoplasma contamination test. SpCas9-expressing HEK293T (HEK293T-SpCas9) cells were generated by a PiggyBac transposon system followed by selection in the presence of 50 μg/ml hygromycin to ensure high Cas9 activity. HEK293T cells were transient transduced with pPB-TRE-spCas9-Hygromycin vector and pCMV-hybase vector with a 9:1 ratio to generate SpCas9-expressing HEK293T.

Vector construction

The LentiU6-LacZ-GFP-Puro (BB) vector was generated by our group previously (Addgene ID: 170459). This plasmid can also be acquired from the Luo lab (https://dream.au.dk/tools-and-resources).

SURRO-seq library design

Each SURRO-seq oligo consists of a BsmBI recognition site “cgtctc” with 4 bp specific nucleotides “acca” upstream, following the GGA cloning linker “aCACC”, one bp “g” for initiating transcription from U6 promoter, 20 bp gRNA sequences of “gN20”, 82bp gRNA scaffold sequence, 37bp surrogate target sequences (10bp barcode sequences, 20 bp protospacer and 3 bp PAM sequences, 4 bp downstream sequences), the downstream linker “GTTTg” and another BsmBI binding site and its downstream flanking sequences “acgg”.

The SURRO-seq pool was designed as follows: (1) LibA contains 11 on target and corresponding 170 off target gRNAs from three published off-target detection methods (T7E1, GUIDESeq, CIRCLESeq); (2) LibB contains 110 gRNAs retrieved from published studies, which we expect to have sequence characteristics representative of gRNAs in gene therapy use (Cancers, PD-1, DMD, β-hemoglobinopathies, SCD, CCR5, HTT, CEP290). **(3)** We predicted off-target sites of each gRNA with FlashFry (v 1.80) and retrieved potential off-target with up to 4bp mismatches in human genome hg19. (**4)** For each surrogate site, we added 10bp barcode (fixed “AC” for the first two nucleotides + 8 bp Unique molecular identifiers (UMIs) sequences) to the upstream sequence of each selected gRNA, constructed the surrogate target sequence as 10 bp barcode+ 23 bp gRNA (include PAM) + 4 bp downstream = 37 bp; (**5)** Off target sites with BsmBI recognition site were discarded, because of GGA cloning; (**6)** LibC contains surrogate sites with all possible 1bp mismatch for five RGNs. The oligo pools were synthesized in Genscript® (Nanjing, China), and all sgRNA sequences and their oligos are provided in Supplementary Data.

Construction of SURRO-seq plasmid library

PCR amplification was used to amplify the 170-nt oligonucleotide pool. Firstly, the SURRO-seq oligos diluted to 1 ng/μl and then performed PCR amplifications using the primers: SURRO (BsmBI GGA)-F and SURRO (BsmBI GGA)-R (**Supplementary Data 6**). The PCR reaction was carried out using PrimeSTAR HS DNA Polymerase (Takara, Japan) following the manufacturer’s instruction.

The PCR products of SURRO-seq oligos were then used for Golden Gate Assembly (GGA) to generate the plasmids library. 36 parallel GGA reactions were performed, and the ligation products were pooled into one tube. Transformation was then carried out using chemically competent DH5α cells. For each reaction, 10 μl GGA ligation product was transformed in to 50 μl competent cells and all the transformed cells were plated on one LB plate (15 cm dish in diameter) with Xgal, IPTG and Amp selection. High ligation efficiency was determined by the presence of very few blue colonies. To ensure that there was sufficient coverage of each surrogate vector in the oligonucleotide library. For one library containing 12,000 synthetic oligos, 42 parallel transformations were performed, and all the bacterial colonies were scraped

off and pooled together for plasmids midi-prep (PureLink™ HiPure Plasmid DNA Midiprep Kit). For small library, equal ratio reduction can be adjusted accordingly. For NGS-based quality quantification of library coverage, midi-prep plasmids were used as DNA templates for PCR amplifications, followed by gel purification and NGS sequencing.

SURRO-seq plasmid library lentivirus packaging.

Supernatants containing lentiviral particles were produced by transient transfection of HEK293T cells using PEI 40000 (Polyethylenimine Linear, MW 40000). For 10 cm dish transfection, the DNA/PEI mixture contains 13 μg pLenti-TRAPseq vectors, 3 μg prove-REV, 3.75 μg pMD.2G, 13 μg pMDGP-Lg/p-RRE, 100 μg PEI 40000 solution (1 μg/μl in sterilized ddH_2O) and supplemented by Opti-MEM without phenol red (Invitrogen) to a final volume of 1 mL. The transfection mixture was pipetted up and down gently several times, and further incubated and kept at room temperature (RT) for 20 min. The transfection complex was added to 80%-confluent HEK293T cells in a 10-cm dish containing 10 ml of culture medium. After 48 h viral supernatant was harvested and filtered with a 0.45 μm filter. Polybrene solution (Sigma-Aldrich) was added to the crude virus solution to a final concentration of 8 μg/mL. The crude virus solution was aliquoted into 15 mL tubes (5 mL/tube) and store in -80 °C freezer until used.

Lentivirus titer quantification by flow cytometry (FCM)

The LentiU6-LacZ-GFP-Puro (BB) vector expresses an EGFP gene. The functional titer of our lentivirus prep was assayed by FCM. Briefly, 1) **Day 1**: Seed HEK293T cells to 24-well plate. 18 wells were used to perform the titter detection, a gradient volume of the crude lentivirus was added into the cells and each volume was tested with two replicates; 2) **Day 2**: Transduce cells at 60~80% confluence. Before transduction, determine the total number of cells using one well of cells. The remaining wells were changed to fresh culture medium containing 8μg/mL polybrene. A gradient volume of crude virus was added to each well and mix gently; 3) **Day 3**: Change to fresh medium without polybrene; 4) **Day 4:** Harvest all the cells and wash with PBS twice. The cells were fixed in 4% formalin solution at RT for 20 min and spun down the cell pellet at 2,000 rpm for 5 min. Cell pellet was washed with PBS twice and re-suspended in PBS solution, followed by FCM analysis. FCM was performed using a BD LSRFortessaTM cell analyzer with at least 30,000 events collected for each sample in duplicates. The FCM output data was analyzed by the software Flowjo vX.0.7. Percentage of GFP-positive cells was calculated as: Y% = $N_{\text{GFP-positive cells}}$ / $N_{\text{total cells}}$ x 100 %. For accurate titter determination, there should be a linear relationship between the GFP positive percentages and crude volume. The titter (Transducing Units (TU/mL) calculation according to this formula: TU/mL = (N_{initial} x Y% x 1000)/V. V represents the crude volume (μl) used for initial transduction.

SURRO-seq library lentivirus transduction

HEK293T-SpCas9 cells were cultured in D10 medium with 50 μg/ml hygromycin throughout the whole experiment. For SURRO-seq library transduction, at Day -1: 2.5 x 10^6 cells per 10 cm dish were seeded. For a 12K SURRO-seq library, transductions were performed in 10 replicates to reach 4000X coverage. For each group, one plate was used for cell number determination before transduction and another plate was used for drug-resistance (puromycin) test control. The remaining 10 plates were used for the SURRO-seq lentivirus library transduction (transduction coverage per gRNA exceeds 4000X of a 12K library); 2) Day 0: We first determined the approximate cell number per dish. This was used to determine the volume of crude lentivirus used for transduction using a multiplicity of infection (MOI) of 0.3. The low MOI (0.3) ensured that most infected cells receive only 1 copy of the lentivirus construct with high probability. The calculation formula is V = N x 0.3 / TU. V = volume of crude lentivirus used for infection (ml); N = cell number in the dish before infection; TU = the titter of crude lentivirus (IFU/mL). The infected cells were cultured in a 37 ℃ incubator; 3) Day 1: 24 hours after transduction, the cell was passaged at a ratio of 3 folds. 4) Day 2: The transduced cells were cultured in D10 medium containing 50 μg/ml hygromycin, 1 μg/mL puromycin, and 1 μg/mL doxycycline to induce Cas9 overexpression. 5) The transduced cells were spitted every 2~3 days when cell confluence reaches up to 90% at a ratio of 1:3. Cells from day 10 were harvested for further genomic DNA extraction. Parallel experiments were performed using wildtype HEK293T cells as MOCK controls.

PCR amplicons of TRAPs from cells

Genomic DNA was extracted using the phenol-chloroform method. Then the genomic DNA was purified and subjected to SURRO-seq PCR. The PCR primers were SURRO-NGS-F and SURRO-NGS-R1 (**Supplementary Data 6**). In this study, 5 ug genomic DNA was used as template in one PCR reaction which contained approximately 7.6 x 10^5 copies of surrogate construct which covered about 63 times coverage of a 12K SURRO-seq library. For a 12K library, 32 parallel PCR reactions were performed to achieve approximately 2,016 times coverage of each construct. Then the PCR products were purified by 1.5% gel and mixed with equal amounts and deep sequenced.

Synthetic gRNAs

All synthetic gRNAs used for validation of OTs were chemically modified to increase stability in cells and synthesized by Synthego Co. (California).

RNP nucleofection

The CRISPR RNP was delivered into cells by nucleofection. For one nucleofection, 6 μg SpCas9 protein (Cat# 1081059, IDT) and the 3.2 μg synthetic gRNA (Synthego) was mixed in a PCR tube by pipetting and incubated at room temperature for at least 10 min and maximum 1hr. Then 200,000 suspended cells were gently resuspended cells in 20 uL nucleofection buffer (OptiMEM) by pipetting up and down. The cells and RNP complex were then transferred to a 4D-Nucleofector 16-well nucleocuvette strip (Catalog #: AXP-1004, Lonza). The samples should cover the bottom of the wells, and any air bubble must be avoided. Nucleofection was performed with program CM-138. Immediately after electroporation, prewarmed culture media was added to the cells (180μL per well of the Nucleocuvette strip). The cells were then transferred into one well of a 12-well cell culture plate with prewarmed medium. Cells were harvested for amplicon PCR and deep sequencing 48 hours after transfection.

Deep amplicon sequencing

The on-target and off-target site were amplified by PCRs. All primers used for PCR were showed in **Supplementary Data 6**. The amplicons were subjected to deep sequencing on the MGISEQ-2000 (MGI of BGI in China) platform. All the samples were subjected to pair-ended 150 bp deep-sequencing on MGISEQ-2000 platform.

Raw data processing

FastaQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and fastp (https://github.com/OpenGene/fastp) with default options were used for data quality control and filtering with the default parameters. The pair-end data was assembled using flash (http://www.cbcb.umd.edu/software/flash). BWA-MEM with default options was used to map the assembled data to the designed oligos sequence to preliminarily distinguish the data of each TRAP site.

Data filtering

The pysam module of Python-3.8 was used to split the aligned data according to the site number of the chip, and the reads of different sites were obtained. Then, we used three steps of strictly controlling parameters to filter the data of each site. Firstly, according to the structure of the chip, g + gRNA (20bp) + scaffold (82bp) + barcode (10bp) + GTTT should remain unchanged at the beginning and end of each site. Then, in order to remove the chip synthesis errors, the pseudo editing sequences found in WT group were removed from spcas9 group. Finally, in order to remove the interference of sequencing errors on the data, the extracted sequence of each site was re-aligned to the reference sequence, and the 1bp indel on N1-N14 and N22-N27 of TRAP (27bp) sequence were removed. The above three filtering steps were completed with julia-1.5.3 language.

Fisher's exact test and statistical analysis

In order to obtain stable and effective off-target efficiency, false positive results must be excluded. We used the number of reads of indel and no indel in spcas9 group and WT group to form a 2 × 2 matrix. Fisher's exact test was used to confirm whether the editing of each site was effective. In order to reduce False Discovery Rate (FDR), all p-values were corrected by BH (Benjaminiand Hochberg) method. Next, we used strict parameters (Total read numbers(spCas9) 32, Indel read numbers (spCas9) 5, Indel Frequency (IF%) (WT) 25) to filter off-target efficiency with bias. Then we used parameters (Fold Change (FC) > 2, p-value (adjusted by BH) < 0.05) to divide the off-target data set into two parts for downstream analysis. The calculation formula of indel efficiency is as follows:

$$Indel\ Frequency(\%) = \frac{Indel\ read\ numbers}{Total\ read\ numbers} \times 100$$

And fold change is as follows:

$$FC = \frac{Indel\ efficiency[spCas9]}{Indel\ efficiency[WT]}$$

Fisher's exact test and other statistical analysis were performed in R-4.0.3. Visualization was completed by R and excel.

Data availability

All NGS data generated by this study have been shared via the CNGB public data depository with the following accession numbers: CNP0001979, CNP0002648. A complete list of 704 NGS samples were summarized in **Supplementary Data 7**.

Declarations

Acknowledgements

This project was partially supported by the Sanming Project of Medicine in Shenzhen (SZSM201612074, to L.B. and Y.L), Qingdao-Europe Advanced Institute for Life Sciences Grant, Guangdong Provincial Key Laboratory of Genome Read and Write (No. 2017B030301011 to X.Xu.), Guangdong Provincial Academician Workstation of BGI Synthetic Genomics (No. 2017B090904014 to X.Xu.), Danish Research Council (9041-00317B to J.G. and Y.L.), European Union's Horizon 2020 research and innovation program under grant agreement No 899417 (Y.L.), the DFF Sapere Aude Starting grant (8048-00072A to L.L.) and the National Human Genome Research Institute of the National Institutes of Health (RM1HG008525 to G.C.). We thank the China National GeneBank for the support of executing the project under the framework of Genome Read and Write.

Conflict of interests

Authors declare no conflict of interests

References

1. Doudna, J.A. The promise and challenge of therapeutic genome editing. Nature **578**, 229–236 (2020).
2. Xiang, X. et al. Efficient correction of Duchenne muscular dystrophy mutations by SpCas9 and dual gRNAs. Mol Ther Nucleic Acids **24**, 403–415 (2021).
3. Frangoul, H. et al. CRISPR-Cas9 Gene Editing for Sickle Cell Disease and beta-Thalassemia. N Engl J Med **384**, 252–260 (2021).
4. Esrick, E.B. et al. Post-Transcriptional Genetic Silencing of BCL11A to Treat Sickle Cell Disease. N Engl J Med **384**, 205–215 (2021).
5. Hendel, A. et al. Chemically modified guide RNAs enhance CRISPR-Cas genome editing in human primary cells. Nat Biotechnol **33**, 985–989 (2015).
6. Ding, X. et al. Improving CRISPR-Cas9 Genome Editing Efficiency by Fusion with Chromatin-Modulating Peptides. CRISPR J **2**, 51–63 (2019).
7. Lin, L. et al. Fusion of SpCas9 to E. coli Rec A protein enhances CRISPR-Cas9 mediated gene knockout in mammalian cells. J Biotechnol **247**, 42–49 (2017).
8. Ma, L. et al. MiCas9 increases large size gene knock-in rates and reduces undesirable on-target and off-target indel edits. Nat Commun **11**, 6082 (2020).
9. Suzuki, K. et al. In vivo genome editing via CRISPR/Cas9 mediated homology-independent targeted integration. Nature **540**, 144–149 (2016).
10. Liang, X., Potter, J., Kumar, S., Ravinder, N. & Chesnut, J.D. Enhanced CRISPR/Cas9-mediated precise genome editing by improved design and delivery of gRNA, Cas9 nuclease, and donor DNA. J Biotechnol **241**, 136–146 (2017).
11. Lino, C.A., Harper, J.C., Carney, J.P. & Timlin, J.A. Delivering CRISPR: a review of the challenges and approaches. Drug Deliv **25**, 1234–1257 (2018).
12. Tsai, S.Q. et al. CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR-Cas9 nuclease off-targets. Nat Methods **14**, 607–614 (2017).
13. Kim, D. et al. Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nat Methods* **12**, 237–243, 231 p following 243 (2015).
14. Cameron, P. et al. Mapping the genomic landscape of CRISPR-Cas9 cleavage. Nat Methods **14**, 600–606 (2017).

15. Yan, W.X. et al. BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks. Nat Commun **8**, 15058 (2017).
16. Kim, D. & Kim, J.S. DIG-seq: a genome-wide CRISPR off-target profiling method using chromatin DNA. Genome Res **28**, 1894–1900 (2018).
17. 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).
18. Ortinski, P.I., O'Donovan, B., Dong, X. & Kantor, B. Integrase-Deficient Lentiviral Vector as an All-in-One Platform for Highly Efficient CRISPR/Cas9-Mediated Gene Editing. Mol Ther Methods Clin Dev **5**, 153–164 (2017).
19. Hu, J. et al. Detecting DNA double-stranded breaks in mammalian genomes by linear amplification-mediated high-throughput genome-wide translocation sequencing. Nat Protoc **11**, 853–871 (2016).
20. Yin, J. et al. Optimizing genome editing strategy by primer-extension-mediated sequencing. Cell Discov **5**, 18 (2019).
21. Wienert, B. et al. Unbiased detection of CRISPR off-targets in vivo using DISCOVER-Seq. Science **364**, 286–289 (2019).
22. Lee, S.H. et al. CUT-PCR: CRISPR-mediated, ultrasensitive detection of target DNA using PCR. Oncogene **36**, 6823–6829 (2017).
23. Brinkman, E.K. & van Steensel, B. Rapid Quantitative Evaluation of CRISPR Genome Editing by TIDE and TIDER. Methods Mol Biol **1961**, 29–44 (2019).
24. Shen, M.W. et al. Predictable and precise template-free CRISPR editing of pathogenic variants. Nature **563**, 646–651 (2018).
25. Allen, F. et al. Predicting the mutations generated by repair of Cas9-induced double-strand breaks. Nat Biotechnol (2018).
26. VI Congress of the International Xenotransplantation Association. Chicago, Illinois, USA. 29 September-3 October 2001. Abstracts. Xenotransplantation **8 Suppl 1**, 1–134 (2001).
27. Xiang, X. et al. Enhancing CRISPR-Cas9 gRNA efficiency prediction by data integration and deep learning. Nat Commun **12**, 3238 (2021).
28. Fu, Y. et al. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. Nat Biotechnol **31**, 822–826 (2013).
29. Pattabhi, S. et al. In Vivo Outcome of Homology-Directed Repair at the HBB Gene in HSC Using Alternative Donor Template Delivery Methods. Mol Ther Nucleic Acids **17**, 277–288 (2019).
30. Anderson, E.M. et al. Systematic analysis of CRISPR-Cas9 mismatch tolerance reveals low levels of off-target activity. J Biotechnol **211**, 56–65 (2015).
31. Zheng, T. et al. Profiling single-guide RNA specificity reveals a mismatch sensitive core sequence. Sci Rep **7**, 40638 (2017).
32. Fu, B.X., 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).

33. Dabrowska, M., Juzwa, W., Krzyzosiak, W.J. & Olejniczak, M. Precise Excision of the CAG Tract from the Huntingtin Gene by Cas9 Nickases. Front Neurosci **12**, 75 (2018).
34. Grasso, C.S. et al. The mutational landscape of lethal castration-resistant prostate cancer. Nature **487**, 239–243 (2012).
35. Walker, C.J. et al. Patterns of CTCF and ZFHX3 Mutation and Associated Outcomes in Endometrial Cancer. J Natl Cancer Inst **107** (2015).
36. Kawaguchi, M. et al. A diagnostic marker for superficial urothelial bladder carcinoma: lack of nuclear ATBF1 (ZFHX3) by immunohistochemistry suggests malignant progression. BMC Cancer **16**, 805 (2016).
37. Song, Z. et al. Genomic profiles and tumor immune microenvironment of primary lung carcinoma and brain oligo-metastasis. Cell Death Dis **12**, 106 (2021).
38. Antony, J.S. et al. Gene correction of HBB mutations in CD34(+) hematopoietic stem cells using Cas9 mRNA and ssODN donors. Mol Cell Pediatr **5**, 9 (2018).
39. Seki, A. & Rutz, S. Optimized RNP transfection for highly efficient CRISPR/Cas9-mediated gene knockout in primary T cells. J Exp Med **215**, 985–997 (2018).
40. Hoshijima, K. et al. Highly Efficient CRISPR-Cas9-Based Methods for Generating Deletion Mutations and F0 Embryos that Lack Gene Function in Zebrafish. Dev Cell **51**, 645–657 e644 (2019).
41. Vakulskas, C.A. et al. A high-fidelity Cas9 mutant delivered as a ribonucleoprotein complex enables efficient gene editing in human hematopoietic stem and progenitor cells. Nat Med **24**, 1216–1224 (2018).
42. Kleinstiver, B.P. et al. High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. Nature **529**, 490–495 (2016).
43. Slaymaker, I.M. et al. Rationally engineered Cas9 nucleases with improved specificity. Science **351**, 84–88 (2016).
44. Doench, J.G. et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol **34**, 184–191 (2016).
45. Cho, S.W. et al. Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. Genome Res **24**, 132–141 (2014).
46. Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. Science **339**, 819–823 (2013).
47. Jiang, W., Bikard, D., Cox, D., Zhang, F. & Marraffini, L.A. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nat Biotechnol **31**, 233–239 (2013).
48. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science **337**, 816–821 (2012).
49. Hsu, P.D. et al. DNA targeting specificity of RNA-guided Cas9 nucleases. Nat Biotechnol **31**, 827–832 (2013).

50. Chuai, G. et al. DeepCRISPR: optimized CRISPR guide RNA design by deep learning. Genome Biol **19**, 80 (2018).
51. Singh, R., Kuscu, C., Quinlan, A., Qi, Y. & Adli, M. Cas9-chromatin binding information enables more accurate CRISPR off-target prediction. Nucleic Acids Res **43**, e118 (2015).
52. 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).
53. Alkan, F., Wenzel, A., Anthon, C., Havgaard, J.H. & Gorodkin, J. CRISPR-Cas9 off-targeting assessment with nucleic acid duplex energy parameters. Genome Biol **19**, 177 (2018).
54. Fu, R. et al. Systematic decomposition of sequence determinants governing CRISPR/Cas9 specificity. Nat Commun **13**, 474 (2022).
55. Ran, F.A. et al. In vivo genome editing using Staphylococcus aureus Cas9. Nature **520**, 186–191 (2015).
56. Hou, Z. et al. Efficient genome engineering in human pluripotent stem cells using Cas9 from Neisseria meningitidis. Proc Natl Acad Sci U S A **110**, 15644–15649 (2013).
57. Burstein, D. et al. New CRISPR-Cas systems from uncultivated microbes. Nature **542**, 237–241 (2017).
58. Zetsche, B. et al. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. Cell **163**, 759–771 (2015).
59. Cox, D.B.T. et al. RNA editing with CRISPR-Cas13. Science **358**, 1019–1027 (2017).
60. Komor, A.C., Kim, Y.B., Packer, M.S., Zuris, J.A. & Liu, D.R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature **533**, 420–424 (2016).
61. Mok, B.Y. et al. A bacterial cytidine deaminase toxin enables CRISPR-free mitochondrial base editing. Nature **583**, 631–637 (2020).
62. Sakata, R.C. et al. Base editors for simultaneous introduction of C-to-T and A-to-G mutations. Nat Biotechnol **38**, 865–869 (2020).
63. Porto, E.M., Komor, A.C., Slaymaker, I.M. & Yeo, G.W. Base editing: advances and therapeutic opportunities. Nat Rev Drug Discov **19**, 839–859 (2020).
64. Anzalone, A.V. et al. Search-and-replace genome editing without double-strand breaks or donor DNA. Nature **576**, 149–157 (2019).
65. Jensen, K.T. et al. Chromatin accessibility and guide sequence secondary structure affect CRISPR-Cas9 gene editing efficiency. FEBS Lett **591**, 1892–1901 (2017).
66. Kallimasioti-Pazi, E.M. et al. Heterochromatin delays CRISPR-Cas9 mutagenesis but does not influence the outcome of mutagenic DNA repair. PLoS Biol **16**, e2005595 (2018).
67. Chung, C.H. et al. Computational Analysis Concerning the Impact of DNA Accessibility on CRISPR-Cas9 Cleavage Efficiency. Mol Ther **28**, 19–28 (2020).

Figures

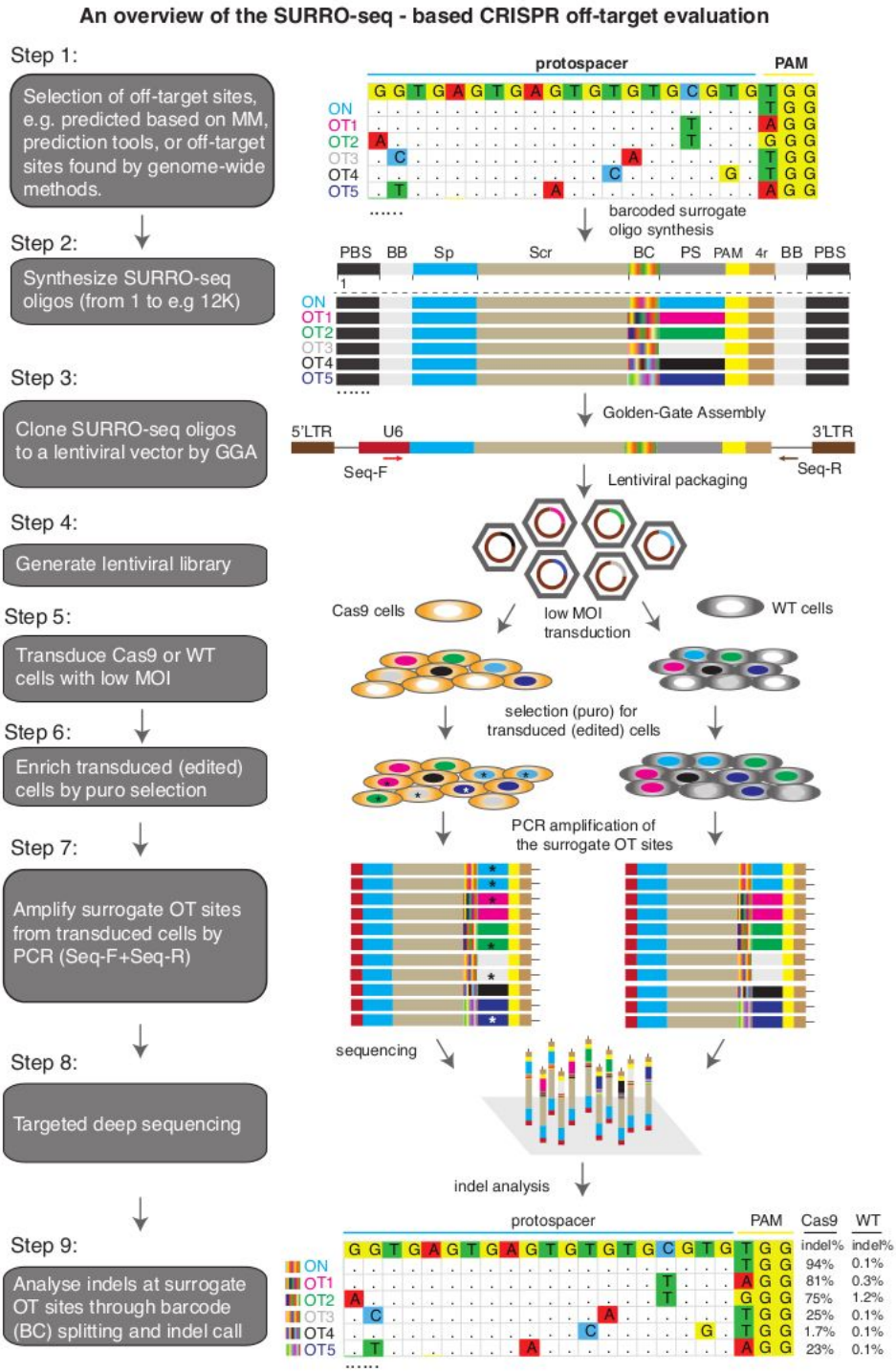


Figure 1

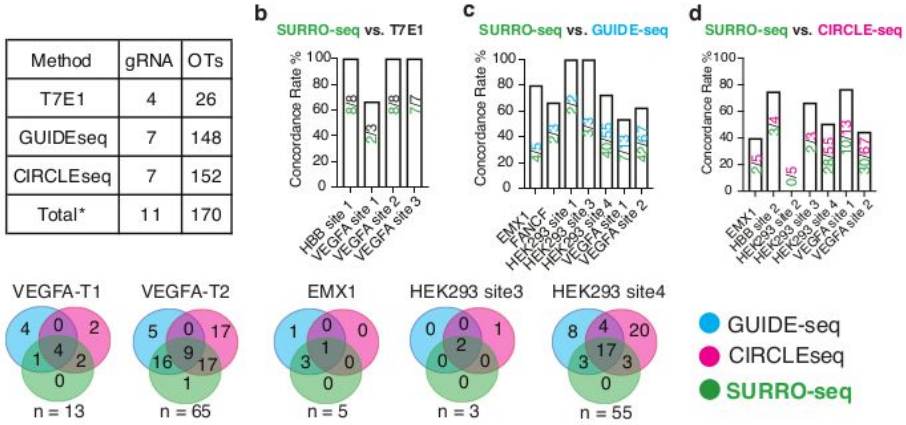
Design of SURRO-seq

An overview of the nine major steps of SURRO-seq is schematically presented. MM, mismatches; ON, on-target; OT, off-targets; PBS, primer binding sites; BB, BsmBI binding site; Sp, spacer; Scr, SpCas9 gRNA scaffold; BC, barcodes; PS, protospacer; PAM, protospacer adjacent motif; 4r, 4bp downstream sequences;

GGA, Golden Gate Assembly; MOI, multiplexity of infection; WT, wildtype; Results presented in Step 9 are indel frequencies captured by SURRO-seq for RGN VEGFA T3.

a

Method	gRNA	OTs
T7E1	4	26
GUIDEseq	7	148
CIRCLEseq	7	152
Total*	11	170



f

Comparison of VEGFA-T1 OTs detection between T7E1, GUIDEseq, CIRCLEseq and SURRO-seq

protospacer	PAM	T7E1	GUIDE-seq		CIRCLE-seq		SURRO-seq					
GGGTGGGGGGAGTTTGCTCCT	GG	RGI%	RE%	reads	RE%	reads	IF%, SpCas9	I/T reads	IF%, MOCK	I/T reads	p.adj.	
..........A..	..	10.5	100	2685	100	736	63.29	1236/1953	0.09	2/2264	0E+00	*
..........C..A..	..	NA	9.27	249	ND	ND	31.49	137/435	0.6	3/501	9E-47	*
TA....A.....C.....	..	NA	72.07	1935	130.98	964	20.88	3363/16110	0.12	24/20189	0E+00	*
.C.....T.....C..	..	NA	9.16	246	55.71	410	13.46	349/2592	0.08	2/2427	2E-101	*
C..G...A.....	..	3.67	13.97	375	136.68	1006	2.39	73/3058	0.12	3/2581	6E-16	*
...GA.....A.....	..	8.54	30.5	819	68.75	506	1.79	34/1897	0.73	9/1227	2E-02	*
.A.G....A.C.....A..	..	NA	ND	ND	87.23	642	1.14	26/2278	0.04	1/2518	2E-07	*
...G.CA.....A.....	..	NA	ND	ND	81.25	598	0.87	33/3814	0.1	4/4049	7E-07	*
CT.GT.....C.....A..	..	NA	12.07	324	ND	ND	0.65	7/1073	0	0/656	8E-02	NS
..A...T.....	..	8.32	22.23	597	ND	ND	0.27	13/4742	0.62	29/4679	2E-02	NS
.T.G...TA.....A..	..	NA	8.98	241	ND	ND	0.13	29/22221	0.15	28/18274	7E-01	NS
.A.GT.....A.....	..	NA	4.28	115	ND	ND	0.1	6/6051	0.12	6/4818	9E-01	NS
..C...A...GA.....	..	NA	ND	ND	58.97	434	0.07	4/6075	0.13	7/5308	5E-01	NS
AT.....T.....A.....A..	..	NA	ND	ND	70.38	518	0.04	3/7997	0.07	5/7587	6E-01	NS

g

Comparison of VEGFA-T2 OTs detection between T7E1, GUIDEseq, CIRCLEseq and SURRO-seq

protospacer	PAM	T7E1	GUIDE-seq		CIRCLE-seq		SURRO-seq					
GACCCCCTCCACCCCGCCTCCGG	RGI%	RE%	reads	RE%	reads	IF%, SpCas9	I/T reads	IF%, MOCK	I/T reads	p.adj.	
..........A..	..	38.6	100	540	100	294	97	8350/8608	0.75	155/20568	0E+00	*
ATT.....C.....A..	..	19.4	286.85	1549	448.98	1320	96.38	1492/1548	2.21	74/3352	0E+00	*
ACA.....C.....A..	..	12.3	218.15	1178	259.18	762	95.53	2458/2573	0.2	17/8481	0E+00	*
CTA.....	..	15	100	540	211.56	622	93.21	851/913	0.4	2/494	8E-306	*
T.....CA.....T..	..	6.97	148.7	803	ND	ND	90.69	3818/4210	0.46	62/13387	0E+00	*
.GG.....T..	..	15.6	89.44	483	682.99	2008	87.57	3495/3991	0.07	8/11623	0E+00	*
.....C.....C.....	..	33.6	72.41	391	ND	ND	83.79	1685/2011	0.3	8/2693	0E+00	*
.CTT.....A.....	..	NA	170.19	919	425.85	1252	77.91	3407/4373	1.4	196/14011	0E+00	*
CCG.....A.T..	..	NA	ND	ND	182.99	538	63.37	263/415	1.19	24/2014	3E-209	*
.....TC.....A.....	..	NA	142.04	767	182.31	536	56.19	608/1082	0.12	3/2547	0E+00	*
ATT.....A..C.T..	..	NA	22.41	121	ND	ND	55.3	5460/9874	0.32	99/30750	0E+00	*
CTG.....C.....A.T..	..	NA	61.11	330	ND	ND	55.16	540/979	0.56	12/2130	1E-307	*
C.T.....C.....A..C.G..	..	NA	62.22	336	ND	ND	52.29	503/962	0.36	10/2784	0E+00	*
C.A.....C.....T..A..	..	NA	20.56	111	ND	ND	49.48	1085/2193	0.06	3/5328	0E+00	*
CG...T.C.....	..	28.7	82.59	446	ND	ND	48.77	4578/9387	0.41	114/28092	0E+00	*
.T.....T.....A.....	..	NA	ND	ND	214.29	630	38.94	493/1266	0.04	1/2264	8E-250	*
AG.....A..T.....G..	..	NA	ND	ND	304.76	896	37.45	3668/9795	0.17	38/21758	0E+00	*
CCT.....CA.....A..	..	NA	115.74	625	ND	ND	33.24	1950/5867	1.47	167/11357	0E+00	*
AGG.....CA.....A..	..	NA	ND	ND	ND	ND	31.91	90/282	0.15	1/664	8E-51	*
C..T.....C.....C..A..	..	NA	23.7	128	ND	ND	27.44	517/1884	1.47	67/4572	5E-220	*
TC.A.....C.....C.G..	..	NA	25	135	ND	ND	27.09	1510/5574	0.25	28/11165	0E+00	*
CTT.....CA.....C.A..	..	NA	47.41	256	ND	ND	24.37	97/398	0.46	7/1524	3E-61	*
AGG.....C...G.....A..	..	NA	46.67	252	ND	ND	21.15	386/1825	0.2	5/2523	3E-147	*
CT.....AC.....A..	..	29.8	ND	ND	207.48	610	20.01	871/4353	0.24	20/8464	0E+00	*
.C.....A.....A..C..	..	NA	ND	ND	154.42	454	17.9	194/1084	0.72	19/2645	1E-85	*
C..TA.....C.....C.G..	..	NA	43.52	235	ND	ND	17.73	1172/6610	0.51	87/17133	0E+00	*
C..TT.....C.....A..A..	..	NA	22.78	123	ND	ND	12.92	131/1014	0.39	10/2556	4E-61	*
TG.....TC.....T..	..	NA	146.48	791	308.16	906	11.51	3539/30753	0.18	127/70368	0E+00	*
CC..A...C.....A..	..	NA	130.37	704	153.74	452	9.58	377/3934	0.63	60/9515	8E-141	*
C.....C.T.....A..	..	NA	ND	ND	200.68	590	9.17	415/4527	0.71	84/11797	2E-151	*
TG.....C.....A.....T..	..	NA	205	1107	ND	ND	5.8	4/69	0	0/197	7E-03	*

(Total VEGFA-T2 OTs data were shown in Extended Tables S2)

Figure 2

Validation of RGN OTs detection between T7E1, GUIDE-seq, and/or CIRCLE-seq by SURRO-seq

A. Overview of RGN gRNAs and OTs selected for validated with SURRO-seq. B-D, Comparison of the OT detection concordance rate between SURRO-seq and T7E1 (B), GUIDE-seq (C) and CIRCLE-seq (D). Numbers are total OTs for each RGN (upper) evaluated with the compared method and OTs agreed with SURRO-seq (lower). (e) Venn diagram comparison of OTs with significant significantly detectable off-targets (SURRO-seq) and OTs with deep sequencing reads detected by GUIDE-seq or CIRCLE-seq. Numbers are OT sites. (F-G) Comparison of VEGFA-T1 (F) and VEGFA-T2 (G) OT detections between T7E1, GUIDE-seq, CIRCLE-seq and SURRO-seq. Full results are showed in Supplementary Data 2. RGI, relative gel intensity; RE%, percentage of relative efficiency, calculated by % reads in OT per reads in ON; IF, indel frequency; I/T reads, indel/total reads.

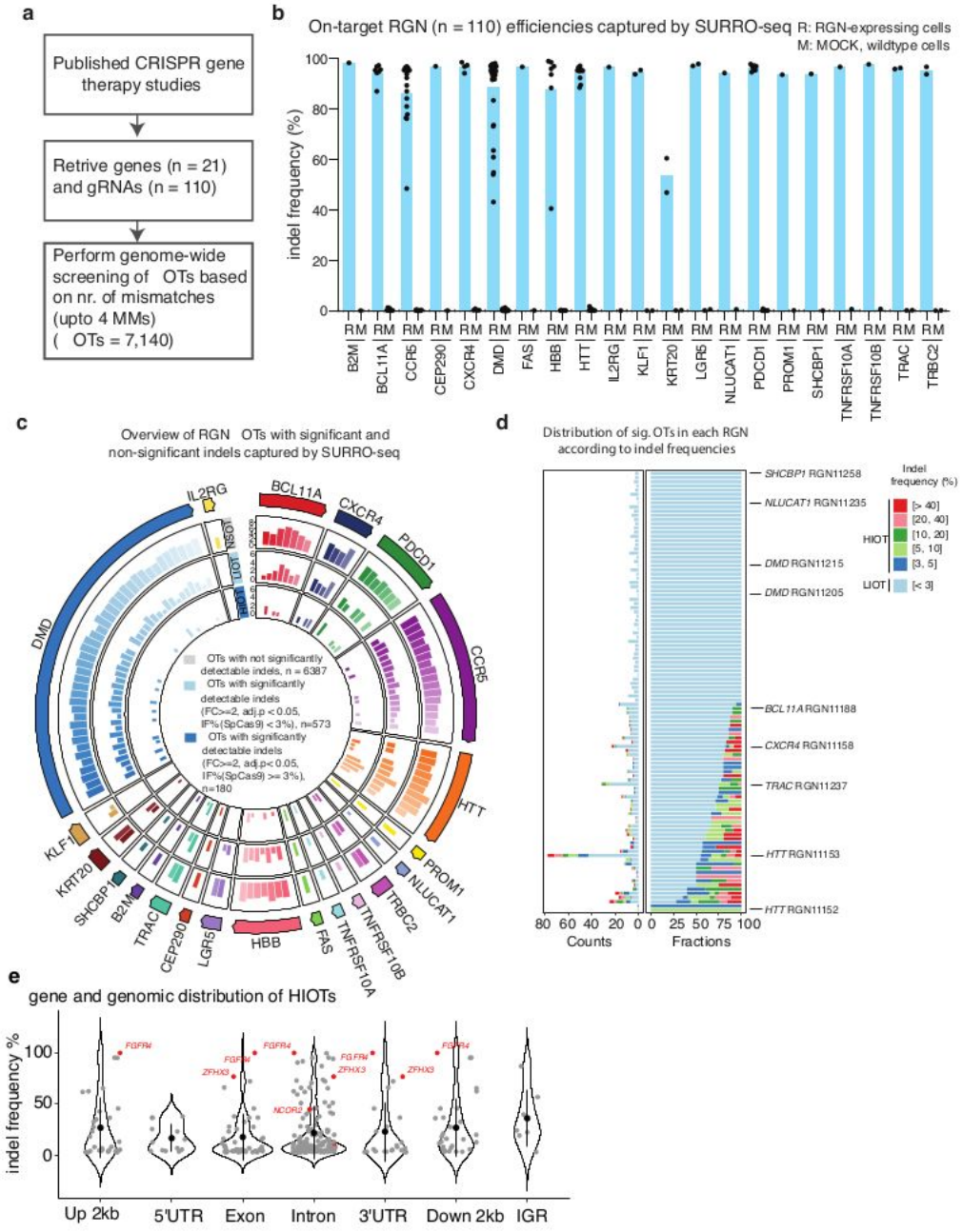
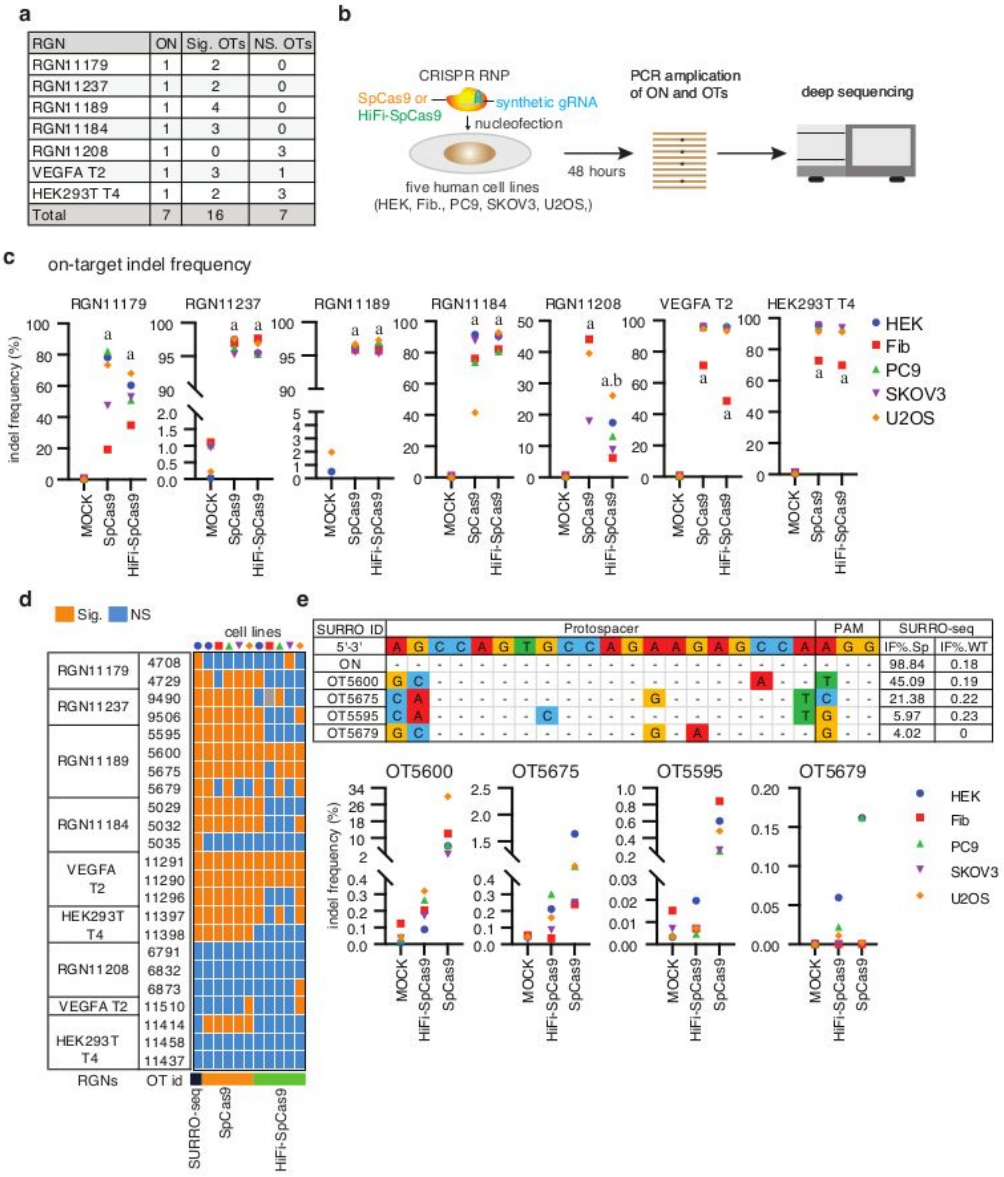


Figure 3

High throughput evaluation of gene therapy RGN OTs with SURRO-seq

A. Overview of gene therapy RGN selection and number of OTs captured. **B**. Quantification of indel frequencies for the 110 RGNs by SURRO-seq. R, RGN edited, M, MOCK control. C. Overview of the number of RGNs OTs with no significantly detectable indel (NSOT, outer circle), with significantly detectable indels

but low indel frequency (< 3%, LIOT, middle circle), and significantly detectable indels with high indel frequency (>= 3%, HIOT, inner circle). **D**. Bar plot of total number (left) and fraction (right) of LIOTs and HIOTs for the RGNs. **E**. Violin plot of the gene and genomics location of the HIOTs and indel frequency (mean and 1SD). OTs in cancer genes are highlighted in read.



a

RGN	ON	Sig. OTs	NS. OTs
RGN11179	1	2	0
RGN11237	1	2	0
RGN11189	1	4	0
RGN11184	1	3	0
RGN11208	1	0	3
VEGFA T2	1	3	1
HEK293T T4	1	2	3
Total	7	16	7

Figure 4

Validation of endogenous OTs in five human cell lines by deep sequencing

A. Overview of RGNs, SURRO-seq identified Sig. OTs and NS. OTs selected for validation. **B**. Schematic illustration of the experiments. RNP, ribonucleoprotein; HEK, HEK293T cell; Fib, human skin-derived fibroblasts. **C**. Dot plot of on-target indel frequencies (indel reads/total reads %) in the CRISPR RNP edited cells. D. Heatmap summary of the RGN OTs evaluated by SURRO-seq and deep sequencing in five human cells lines. Indel frequency values were showed in supplementary data 4. **E**. Example of indel frequencies of four OTs for RGN11189 measured by SURRO-seq and by deep sequencing of RGN edited human cell lines.

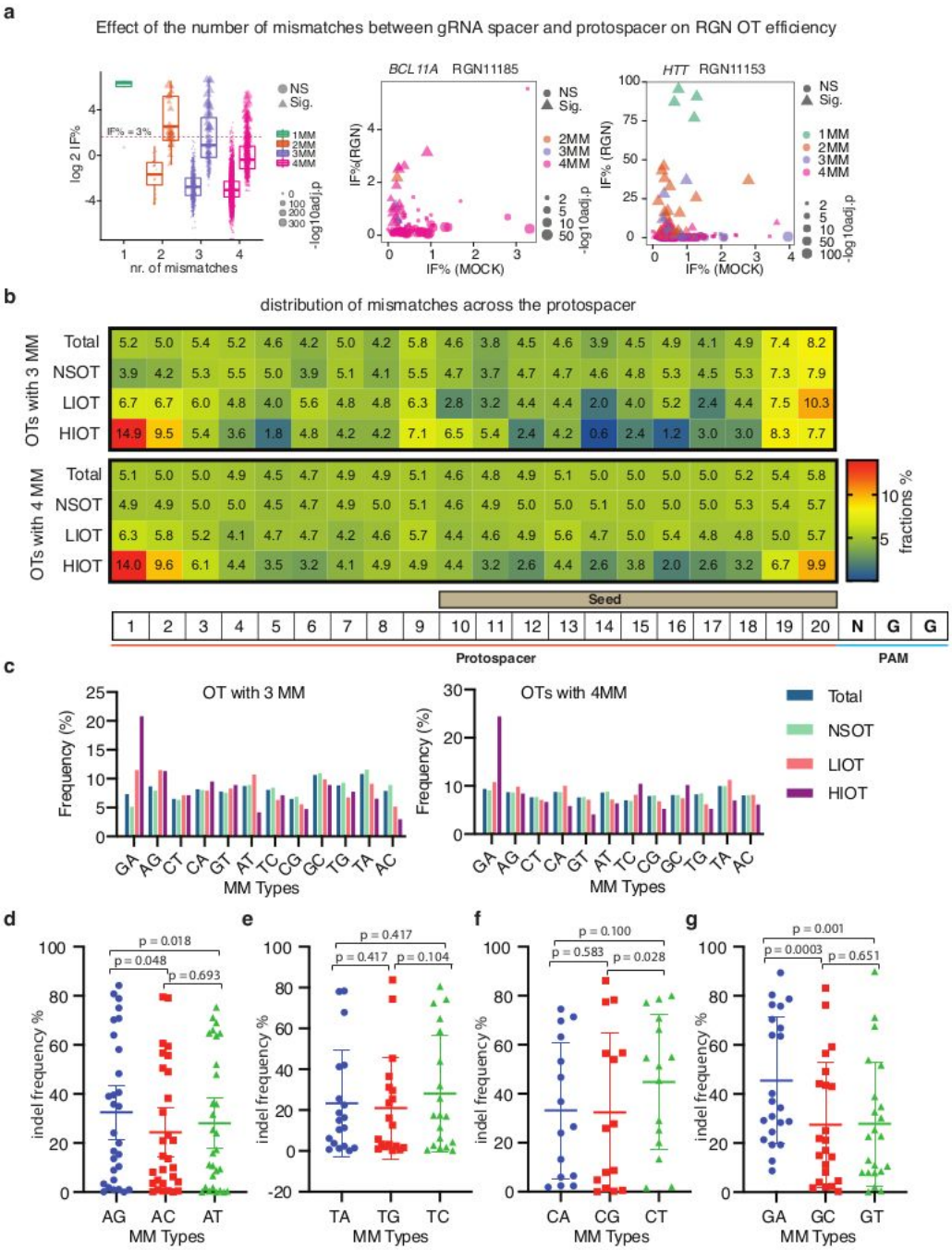


Figure 5

Effects of mismatch number, position and type on RGN off-target activity

A. Box plot of log 2 indel frequency for RGN OTs evaluated by SURRO-seq in LibB. Sites were grouped based the number of mismatches, plotted according to significance and log10 adj. p-values. **B**. Heatmap presentation of the fraction of mismatches occurred in each position of the gRNA for OTs in LibB,

grouped based on total OTs, NSOTs, LIOTs and HIOTs. **C**. Bar plot of appearance frequencies of each type of mismatches occurred in the different groups of RGN OTs in LibB. **D-G**. Dot plots of indel frequencies for OTs with one mismatch measured in LibC. Pair-wise ANOVA analysis was performed for A type mismatches (D), T type mismatches (E), C type mismatches (F), and G type mismatches (G).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryData17.zip
- SupplementaryMaterialsandFigures.pdf
- 3580490related8r8dh7d.pdf