Interrogation of CRISPR-Cas targeting specificity for mammalian genome engineering

by

David Scott

B.S. University of California, San Diego (2012)

Submitted to the Department of Brain and Cognitive Sciences in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2017

Signature redacted

Author .....
David Scott
Department of Brain and Cognitive Sciences
December 21, 2016

Signature redacted

Certified by .....
Feng Zhang
W. M. Keck Career Development Professor of Biomedical Engineering
Thesis Supervisor

Signature redacted

Accepted by .....
Matthew A. Wilson
Sherman Fairchild Professor of Neuroscience and Picower Scholar
Director of Graduate Education for Brain and Cognitive Sciences

ABSTRACT

Class II CRISPR-Cas RNA programmable DNA endonucleases enable high efficiency genome editing across the biological diversity for research, industrial, and biomedical applications. Human genome editing with CRISPR-Cas just recently made its debut in human clinical trials and has immense therapeutic potential to fix disease-causing mutations at the level of DNA. Ensuring the integrity and safety of research, industrial, and biomedical applications of CRISPR-Cas necessitates efficient, versatile, and comprehensive methods to evaluate of the specificity of genome editing. Here, we optimize the efficiency and characterize the targeting specificity of SpCas9 to ensure robust cleavage activity while minimizing off-target activity in human cells. We characterize SpCas9 mismatch tolerance between the guide RNA and target, and provide data-driven design software to guide the selection of high fidelity Cas9 targets. We find that SpCas9 binding activity is not predictive of DNA cleavage, limiting the efficacy of Cas9 ChIP for unbiased evaluation of Cas9 off-target activity. Alternatively, we demonstrate that insert capture – insertion of short DNA fragments at double strand breaks (DSBs) by non-homologous end-joining (NHEJ) – provides unbiased genome-wide identification of off-target cleavage by Cas9 as well as relative rates of indel, chromosomal rearrangement, and translocation accompanying NHEJ repair. However, insert capture is largely limited to use in model cell lines and is fundamentally limited in sensitivity due to labeling of low frequency errors in DSB repair. To directly label DSBs from cell culture or tissue samples, we adapted BLESS (direct in situ breaks labeling, enrichment on streptavidin and next-generation sequencing) and BLISS (Breaks Labeling In Situ and Sequencing) for unbiased genome-wide analysis of CRISPR-Cas specificity. Finally, we consider how human genetic variation will affect the targeting specificity of CRISPR-Cas endonucleases for therapeutic applications. Using the ExAC and 1000 Genomes datasets we find that human variation has important implications for Cas enzyme choice as well as target efficacy and safety. From this analysis, we provide a framework for the design of CRISPR-based therapeutics to maximize efficacy and safety across patient populations.

TABLE OF CONTENTS

Introduction: Inroads to programmable genome engineering: considerations for specificity

Programmable genome editing, the ability to specifically manipulate genomic DNA in a living cell, has enabled breakthrough progress in the interrogation and manipulation of functional genetic elements for research, industrial, and biomedical applications. The targeting specificity of programmable nucleases is critical for ensuring the validity of functional inferences accompanying genome manipulation and the safety of industrial and biomedical applications. To properly frame the topic of CRISPR/Cas specificity addressed by this work, it is useful to consider a brief history of genome engineering with programmable nucleases, cellular repair of targeted DNA cleavage events, and past challenges and investigation of targeting specificity.

Programmable genome engineering was first enabled with the discovery of modular DNA sequence recognition by Zinc finger domains[1]. Biochemical characterization of individual zinc finger domains suggested concordance between individual protein zinc finger domains and recognition of specific 2 – 3 nucleotide DNA motifs[2,3]. Hence, programmable DNA recognition arose from the assembly of polydactyl zinc finger proteins – fusions of 3 – 6 different zinc finger domains cumulatively conferring recognition of DNA targets 6 – 18 nucleotides in length. Zinc finger nucleases (ZFNs) were subsequently created by fusing a FokI DNA nuclease domain to the end of a polydactyl zinc finger protein[4]. FokI cleaves DNA as a dimer[5], and hence, DNA cleavage with zinc finger nucleases was achieved by targeting opposing top and bottom

strand DNA sequences on either side of a desired genomic target, with the FokI nuclease domains forming a dimer at the center to cleave the target DNA.

Despite this breakthrough demonstration of programmable DNA cleavage with ZFNs, the targeting specificity of this technology is challenged by the fact that zinc finger domains often are not perfectly specific for individual 2 – 3 nucleotide motifs[3,6]. The targeting specificity of individual zinc finger domains can also be altered by the DNA sequence flanking a predicted binding site. Despite efforts to hone the specificity of individual zinc finger domains using techniques such as directed evolution, use of ZFNs for manipulation of living cells is often associated with cytotoxicity, indicative of promiscuous DNA modification[7].

Genome engineering with programmable nucleases involves an intricate interplay between nuclease cleavage and DNA repair mechanisms in living cells. Following the introduction of a targeted double strand break in the genome of a living cell, it is the cell's endogenous repair machinery that gives rise to modification of the genomic sequence due to homology directed repair or errors in non-homologous end joining (NHEJ)[8]. Homology directed repair describes cell cycle dependent DNA repair mechanisms active only in G2 and S phase of dividing cells. Homology directed repair precisely alters the genomic sequence at a double strand break based on DNA templates encoding 1) a desired modification to the genomic sequence, and 2) long flanking DNA sequences homologous to the target locus for modification. Prior to the use of programmable nucleases for genome engineering, homologous recombination has been used to incorporate exogenous DNA sequences into the genomes of stem cells for the creation of transgenic animals[9,10]. However, the low efficiency of DNA

modification (typically <0.001%) precluded the use of homology directed repair in isolation for therapeutic genome editing. However, it was discovered that the efficiency of homology directed genome modification increases by 3 – 4 orders of magnitude using ZFNs programmed to repeatedly cleave the target locus for gene repair[11,12,13]. The coupling of ZFNs and homology directed repair marked a critical leap towards therapeutic genome editing and the direct treatment of genetic diseases and epidemics such as HIV/AIDS. However, a major hurdle to the realization of widespread gene therapy is that homology-directed genomic modification also accompanies imprecise modification of the genome by non-homologous end joining (NHEJ).

Quickly rejoining broken strands of DNA in a cell, NHEJ is the dominant form of DNA repair in mammalian cells and is active in all phases of the cell cycle[14]. While NHEJ is a highly accurate machinery for repairing double strand breaks, repeated cleavage of target DNA by programmable nucleases often results in the deletion or insertion of nucleotides at target sites. These errors of NHEJ are termed indels, and if introduced into the coding sequence of a gene, indels often shift the reading frame, resulting in gene silencing via nonsense-mediated decay of the gene product. Unlike homology directed repair, which precisely occurs at the target locus due to requirement for long sequences of target homology in DNA templates, any spurious double strand breaks introduced outside of the target locus by a programmable nuclease are repaired by NHEJ and can result in undesirable gene knockout due to the creation of off-target indels.

The high level of cytotoxicity observed accompanying gene targeting with initial implementations of zinc finger nucleases is the result of widespread off-target cleavage,

overwhelming the capacity of NHEJ to repair widespread genomic lesions. The specificity of cleavage was improved by the creation of FokI variants that prevent homodimerization of FokI domains, however, problems with off-target genome modification persisted largely due to ambiguous ZFN target affinity[15,16]. A solution to limitations in the programmability and specificity of ZFNs ultimately arose not from directed evolution of zinc finger binding domains in a laboratory, but from TAL effectors (TALEs) – modular DNA-binding proteins critical to the pathogenesis of Xanthamonas sp. in plant hosts[17]. TALEs consist of a repetitive core domain containing tandem 32 – 34 amino acid motifs, of which, residues 12 – 13 were observed to be hyper variable. The identities of these hyper variable amino acids show a strong correspondence with single nucleotide residues in TALE binding sites, providing a basis for the assembly of TALE binding domains with programmable specificity[18]. Fusing FokI nuclease domains to TALE DNA-binding domains, pairs of TALE nucleases (TALENs) can be programmed to bind opposing genomic loci and form a FokI, cleaving DNA in a manner very similar to ZFNs[19,20].

Throughout the early development of ZFNs and TALENs, reporter assays and cytotoxicity were the primary techniques used to assess targeting efficiency and specificity respectively. However, cytotoxicity results only from widespread off-target effects that overwhelm cellular DNA repair mechanisms, and much more sensitive screening for off-target effects is critical to ensuring the specificity of genome engineering for research, industrial, and therapeutic applications. Utilizing mismatch dependent nucleases CEL I or CEL II, Surveyor assays allow the detection of indels resulting from DSBs at on-target and off-target genomic loci in a population of cells[16,21].

Surveyor assays have been used to investigate the efficiency and specificity of indel formation accompanying treatment of cells by ZFNs as follows: 1) target loci or predicted off-target loci with high target homology are amplified by PCR; 2) double stranded PCR products are denatured and heterogeneous single strand products are reannealed creating mismatches between modified and unmodified DNA strands; 3) annealed products are treated with CEL I or II nucleases to cleave mismatches; 4) resulting cleaved and full length amplicons are quantified using gel electrophoresis to assess the frequency of modification at the genomic locus. Alternatively, Next Generation Sequencing (NGS) has also been used to directly observe on- and off-target genomic modification by ZFNs.

Beyond the targeted investigation of genome modification frequencies at ZFN target and computationally predicted off-target loci, integrase deficient lentiviral vectors (IDLVs) have been used for unbiased genome-wide investigation of ZFN off-target activity[22]. Researchers found that IDLV DNA fragments lacking lentiviral genomic integration capabilities are incorporated at sites of genomic double strand breaks. After integrating at the site of a double strand break, a known sequence inside the IDLV is used to prime a PCR extension into the unknown genomic sequence flanking each site of integration. Next Generation Sequencing of these genomic products is then used to provide an unbiased map of IDLV integration events at ZFN target and off-target sites throughout the genome. Although limited in sensitivity, IDLV insertion assays provided the first unbiased investigation of programmable nuclease off-target activity.

The demonstration of RNA programmable mammalian genome engineering with CRISPR/Cas9 marked a paradigm shift from ZFNs and TALENs[23,24]. CRISPR/Cas9 is

not an engineered assembly of DNA binding and effector domains such as ZFNs and TALENs, but a single protein endowed by evolution with programmability and highly efficient nuclease activity. Programmed by RNA guides, CRISPR DNA endonucleases mediate cleavage of DNA targets that are complementary to the guide RNA protospacer sequence and flanked by a short protospacer adjacent motif (PAM) specific to each endonuclease[25,26] The ease of programmability of CRISPR/Cas nucleases has fueled rapid development of this technology and adoption for research, industrial, and therapeutic applications. Due to the novel DNA targeting mechanism of CRISPR-Cas nucleases and the rapid adoption of this technology, comprehensive characterization of CRISPR/Cas specificity is critically important. During my graduate studies in the lab of Feng Zhang, I have had the privilege to work with remarkable teams to optimize Cas9 efficacy, characterize the specificity of CRISPR-Cas systems, and develop efficient and versatile methods with the goal of comprehensively evaluating the specificity of genome engineering technologies present and future.

CHAPTER 1: Optimize spCas9 sgRNA for enhanced nuclease activity and evaluate spCas9 target recognition and specificity.

Adapted from:

Hsu, P. D.*, Scott, D. A.*, et al. DNA targeting specificity of RNA-guided Cas9 nucleases. Nat. Biotechnol. 31, 827–832 (2013).

The *Streptococcus pyogenes* Cas9 (SpCas9) nuclease can be efficiently targeted to genomic loci by means of single-guide RNAs (sgRNAs) to enable genome editing. Here, we characterize SpCas9 targeting specificity in human cells to inform the selection of target sites and avoid off-target effects. Our study evaluates >700 guide RNA variants and SpCas9-induced indel mutation levels at >100 predicted genomic off-target loci in 293T and 293FT cells. We find that SpCas9 tolerates mismatches between guide RNA and target DNA at different positions in a sequence-dependent manner, sensitive to the number, position and distribution of mismatches. We also show that SpCas9-mediated cleavage is unaffected by DNA methylation and that the dosage of SpCas9 and sgRNA can be titrated to minimize off-target modification. To facilitate mammalian genome engineering applications, we provide a web-based software tool to guide the selection and validation of target sequences as well as off-target analyses.

1.1 Introduction

The bacterial class II clustered, regularly interspaced, short palindromic repeats (CRISPR) system from S. pyogenes can be reconstituted in mammalian cells using three minimal components[23,24]; the CRISPR-associated nuclease Cas9 (SpCas9), a specificity-determining CRISPR RNA (crRNA), and an auxiliary trans-activating crRNA (tracrRNA)[27]. Following crRNA and tracrRNA hybridization, SpCas9 is targeted to genomic loci matching a 20-nt guide sequence within the crRNA, immediately upstream of a required 5′-NGG protospacer adjacent motif (PAM)[27]. crRNA and tracrRNA duplexes can also be fused to generate a chimeric sgRNA[28] that mimics the natural crRNA-tracrRNA hybrid. Both crRNA-tracrRNA duplexes and sgRNAs can be used to target SpCas9 for multiplexed genome editing in eukaryotic cells[23,24,29]. Although an sgRNA design consisting of a truncated crRNA and tracrRNA has been previously shown to mediate efficient cleavage in vitro[28], it failed to achieve detectable cleavage at several loci that were efficiently modified by crRNA-tracrRNA duplexes bearing identical guide sequences[23,24].

Since the early characterization of restriction endonucleases, it has been shown that the activity of many bacterial nucleases is altered by epigenetic modification of DNA, such as DNA methylation[30,31,32]. Compared to the bacterial genome, the epigenetic landscape in mammalian cells is far more complex, providing challenges to the translation of Cas9 for robust genome-wide editing. Beyond enhanced epigenetic complexity, the human genome sequence is vastly more complex than short bacterial

genome sequences. In order to characterize the fidelity mammalian genome editing with Cas9, it is important to understand how mismatches between the guide RNA and target affect nuclease activity.

Here, we engineer the Cas9 sgRNA for high efficiency cleavage of genomic DNA and demonstrate robust Cas9 nuclease activity in varying epigenetic and cell expression contexts. To characterize of the DNA targeting specificity of SpCas9, we provide a comprehensive investigation of Cas9 single and multiple mismatch tolerance at genomic targets. Additionally, we leverage these principles for the *in silico* prediction of Cas9 off-target activity in the human genome and validate off-target activity at predicted off-target sites using targeted next generation sequencing.

1.2.1 Results: Optimization of spCas9 sgRNA for enhanced nuclease activity

Because the major difference between this sgRNA design and the native crRNA-tracrRNA duplex is the length of the tracrRNA sequence, we tested whether extension of the tracrRNA tail would improve SpCas9 activity. We generated a set of sgRNAs targeting multiple sites within the human EMX1 and PVALB loci with different tracrRNA 3′ truncations (Fig. 1a). Using the SURVEYOR nuclease assay[33], we assessed the ability of each Cas9-sgRNA complex to generate indels in human embryonic kidney (HEK) 293FT cells through the induction of DNA double-stranded breaks (DSBs) and subsequent nonhomologous end joining (NHEJ) DNA damage repair (Methods). sgRNAs with +67 or +85 nucleotide (nt) tracrRNA tails mediated DNA cleavage at all target sites tested, with up to fivefold higher levels of indels than the corresponding crRNA-tracrRNA duplexes (Fig. 1b and Supplementary Fig. 1a). Furthermore, both sgRNA designs efficiently modified PVALB loci that were previously not targetable using crRNA-tracrRNA duplexes[23] (Fig. 1b and Supplementary Fig. 1b). For all five tested targets, we observed a consistent increase in modification efficiency with increasing tracrRNA length. We performed northern blot analyses for the guide RNA truncations and found increased levels of expression for the longer tracrRNA sequences, suggesting that improved target cleavage was at least partially due to higher sgRNA expression or stability (Fig. 1c). Taken together, these data indicate that the tracrRNA tail is important for optimal SpCas9 expression and activity in vivo.

We further investigated the sgRNA architecture by extending the duplex length from 12 to the 22 nt found in the native crRNA-tracrRNA duplex (Supplementary Fig. 2a). We also mutated the sequence encoding the sgRNAs to abolish any poly-T tracts that could serve as premature transcriptional terminators for U6-driven transcription[34]. We tested these new sgRNA scaffolds on three targets within the human EMX1 gene (Supplementary Fig. 2b) and observed only modest changes in modification efficiency. Thus, we established sgRNA(+67) as a minimum effective SpCas9 guide RNA architecture and for all subsequent studies we used the most active sgRNA(+85) architecture.

We have previously shown that a catalytic mutant of SpCas9 (D10A nickase) can mediate gene editing by homology-directed repair without detectable indel formation[23]. Given its higher cleavage efficiency, we tested whether sgRNA(+85), in complex with the Cas9 nickase, can likewise facilitate homology-directed repair without incurring on-target NHEJ. Using single-stranded oligonucleotides as repair templates, we observed that both the wild-type and the D10A SpCas9 mediate homology-directed repair in HEK 293FT cells, whereas only the former does so in human embryonic stem cells (hESCs; Fig. 1d and Supplementary Fig. 3a–c). We further confirmed using SURVEYOR assay that no target indel mutations are induced by the SpCas9 D10A nickase (Supplementary Fig. 3d).

1.2.2 Results: Characterization of epigenetic constraints on spCas9 activity

To explore whether the genome targeting ability of sgRNA(+85) is influenced by epigenetic factors[35,36] that constrain the alternative transcription activator-like effector nuclease (TALENs)[19,37,38,39,40] and potentially also zinc finger nuclease (ZFNs)[12,16,41,20,42] technologies, we further tested the ability of SpCas9 to cleave methylated DNA. Using either unmethylated or M. SssI-methylated pUC19 as DNA targets (Supplementary Fig. 4a,b) in a cell-free cleavage assay, we showed that SpCas9 efficiently cleaves pUC19 regardless of CpG methylation status in either the 20-bp target sequence or the PAM (Supplementary Fig. 4c). To test whether this is also true in vivo, we designed sgRNAs to target a highly methylated region of the human SERPINB5 locus (Fig. 1e,f). All three sgRNAs tested were able to mediate indel mutations in endogenously methylated targets (Fig. 1g).

1.2.3 Results: Characterization of Cas9 nuclease specificity

Evaluation of spCas9 target recognition and target mismatch tolerance

Having established the optimal guide RNA architecture for SpCas9 and having demonstrated its insensitivity to genomic CpG methylation, we sought to conduct a comprehensive characterization of the DNA targeting specificity of SpCas9. Previous studies on SpCas9 cleavage specificity[23], were limited to a small set of single-nucleotide mismatches between the guide sequence and DNA target, suggesting that perfect base-pairing within 10–12 bp directly 5′ of the PAM (PAM-proximal) determines Cas9 specificity, whereas multiple PAM-distal mismatches can be tolerated. In addition, a recent study using catalytically inactive SpCas9 as a transcriptional repressor found no significant off-target effects throughout the Escherichia coli transcriptome[43]. However, a systematic analysis of Cas9 specificity within the context of a larger mammalian genome has not yet been reported.

To address this, we first evaluated the effect of imperfect complementarity between the guide RNA and its genomic target on SpCas9 activity, and then assessed the cleavage activity resulting from a single sgRNA on multiple genomic off-target loci with sequence similarity. To facilitate large-scale testing of mismatched guide sequences, we developed a simple sgRNA testing assay by generating expression cassettes encoding U6-driven sgRNAs using PCR and transfecting the resulting amplicons (Supplementary Fig. 5). We then performed deep sequencing of the region flanking each target site (Supplementary Fig. 6) for two independent biological replicates. From these data, we

applied a binomial model to detect true indel events resulting from SpCas9 cleavage and NHEJ misrepair and calculated 95% confidence intervals for all reported NHEJ frequencies.

We systematically investigated the effect of base-pairing mismatches between guide RNA sequences and target DNA on target modification efficiency. We chose four target sites within the human EMX1 gene[23,44,29,45] and, for each, generated a set of 57 different guide RNAs containing all possible single-nucleotide substitutions in positions 1–19 directly 5′ of the requisite NGG PAM (Fig. 2a). The 5′ guanine at position 20 is preserved, given that the U6 promoter requires guanine as the first base of its transcript. These 'off-target' guide RNAs were then assessed for cleavage activity at the on-target genomic locus.

Consistent with previous findings[23,44,28], SpCas9 tolerates single-base mismatches in the PAM-distal region to a greater extent than in the PAM-proximal region. In contrast to a model that implies that a prototypical 10–12 bp PAM-proximal seed sequence largely determines target specificity[23,44,28], we found that most bases within the 20-bp target site provide varying degrees of specificity. Single-base specificity generally ranges from 8 to 14 bp immediately upstream of the PAM, indicating a sequence-dependent, mismatch-sensitive boundary that varies in length (Fig. 2b, Supplementary Fig. 7).

To further investigate the contributions of base identity and position within the guide RNA to SpCas9 specificity, we generated additional sets of mismatched guide RNAs for 11 more target sites within the EMX1 locus (Supplementary Fig. 8), totaling over 400 sgRNAs. These guide RNAs were designed to cover all 12 possible RNA:DNA

mismatches for each position in the guide sequence with at least 2× coverage for positions 1–10. Our aggregate single-mismatch data reveal multiple exceptions to the seed sequence model of SpCas9 specificity[23,44,45] (Fig. 2c). Within the PAM-proximal region, the degree of tolerance varied with the identity of a particular mismatch, with rC:dC base-pairing exhibiting the highest level of disruption to SpCas9 cleavage activity (Fig. 2c).

In addition to the target specificity, we also investigated the NGG PAM requirement of SpCas9. To vary the second and third positions of PAM, we selected 32 target sites within the EMX1 locus encompassing all 16 possible alternate PAMs with 2× coverage. Using the SURVEYOR assay, we showed that SpCas9 also cleaves targets with NAG PAMs, albeit with one-fifth of the efficiency for target sites with 5′-NGG PAMs (Fig. 2d). The tolerance for an NAG PAM is in agreement with previous bacterial studies[44] and expands the S. pyogenes Cas9 target space to every 4 bp on average within the human genome, not accounting for constraining factors such as guide RNA secondary structure or certain epigenetic modifications (Fig. 2e). Although we have shown here that methylated DNA sequences can be cleaved, by SpCas9 further characterization of the implications of epigenetic factors on CRISPR editing efficiency are needed.

We next explored the effect of multiple base mismatches on SpCas9 target activity. For four targets within the EMX1 gene, we designed sets of guide RNAs that contained varying combinations of mismatches to investigate the effect of mismatch number, position and spacing on SpCas9 target cleavage activity (Fig. 3a,b). In general, we observed that the total number of mismatched base-pairs is a key determinant for SpCas9 cleavage efficiency. Two mismatches, particularly those occurring in a PAM-

proximal region, considerably reduced SpCas9 activity whether these mismatches are concatenated or interspaced (Fig. 3a,b); this effect is further magnified for three concatenated mismatches (Fig. 3a). Furthermore, three or more interspaced (Fig. 3c) and five concatenated (Fig. 3a) mismatches eliminated detectable SpCas9 cleavage in the vast majority of loci.

The position of mismatches within the guide sequence also affected the activity of SpCas9. PAM-proximal mismatches are less tolerated than PAM-distal counterparts (Fig. 3a), recapitulating our observations from the single base-pair mismatch data (Fig. 2c). This effect is particularly salient in guide sequences bearing a small number of total mismatches, whether those are consecutive (Fig. 3a) or interspaced (Fig. 3b). Additionally, guide sequences with mismatches spaced four or more bases apart also mediated SpCas9 cleavage in some cases (Fig. 3c). Thus, together with the identity of mismatched base-pairing, we observed that many off-target cleavage effects can be explained by a combination of mismatch number and position.

Cas9 mismatch tolerance is consistent with genomic off-target modification

Given these mismatched guide RNA results, we expected that for any particular sgRNA, SpCas9 may cleave genomic loci that contain small numbers of mismatched bases. For the four EMX1 targets described above, we computationally selected 117 candidate off-target sites in the human genome that are followed by a 5′-NRG PAM and meet any of the following additional criteria: (i) up to five mismatches, (ii) short insertions or deletions or (iii) mismatches only in the PAM-distal region. Additionally, we assessed off-target

loci of high sequence similarity without the PAM requirement. The majority of off-target sites tested for each sgRNA (30/31, 23/23, 48/51 and 12/12 sites for EMX1 targets 1, 2, 3 and 6, respectively) exhibited modification efficiencies at least 2 magnitudes lower than that of corresponding on-targets (Fig. 4a,b, Supplementary Fig. 9). Of the four off-target sites that exhibit substantial modification efficiencies, three contained only mismatches in the PAM-distal region, consistent with our multiple mismatch sgRNA observations (Fig. 3). Notably, these three loci were followed by 5′-NAG PAMs, demonstrating that off-target analyses of SpCas9 must include 5′-NAG as well as 5′-NGG candidate loci.

Enzymatic specificity and activity strength are often highly dependent on reaction conditions, which at high enzyme concentration might amplify off-target activity[46,47]. One potential strategy for minimizing nonspecific cleavage is to limit the enzyme concentration, namely the level of SpCas9-sgRNA complex. Cleavage specificity, measured as the ratio of on- to off-target cleavage, increased dramatically as we decreased the equimolar amounts of SpCas9 and sgRNA transfected into 293FT cells (Fig. 4c,d) from 7.1 × 10−10 to 1.8 × 10−11 nmol/cell (400 ng to 10 ng of Cas9-sgRNA plasmid). qRT-PCR assay confirmed that the level of hSpCas9 mRNA and sgRNA decreased proportionally to the amount of transfected DNA (Supplementary Fig. 10). Whereas specificity increased gradually by nearly fourfold as we decreased the transfected DNA amount from 7.1 × 10−10 to 9.0 × 10−11 nmol/cell (400 ng to 50 ng plasmid), we observed a notable additional sevenfold increase in specificity upon further decreasing transfected DNA from 9.0 × 10−11 to 1.8 × 10−11 nmol/cell (50 ng to 10 ng plasmid; Fig. 4c). These findings suggest that we can minimize the level of off-target

activity by titrating the amount of SpCas9 and sgRNA DNA delivered. However, increasing specificity by reducing the amount of transfected DNA also leads to a reduction in on-target cleavage. These measurements enable quantitative integration of specificity and efficiency criteria into dosage choice to optimize SpCas9 activity for different applications. Additional work to explore modifications in SpCas9 and sgRNA design may improve SpCas9-intrinsic specificity without sacrificing cleavage efficiency.

1.3 Discussion

The ability to program SpCas9 to target specific sites in the genome by simply designing a short guide RNA complementary to the desired target site holds enormous potential for applications throughout biology and medicine. Our results demonstrate that the specificity of SpCas9-mediated DNA cleavage is sequence- and locus-dependent and governed by the quantity, position and identity of mismatching bases. Whereas the PAM-proximal 8–12 bp of the guide sequence generally defines specificity, the PAM-distal sequences also contribute to the overall specificity of SpCas9-mediated DNA cleavage. Although there may be off-target cleavage for a given guide sequence, they can be predicted and likely minimized by following general design guidelines.

In silico prediction and validation of spCas9 off-target activity

To maximize SpCas9 specificity for editing a particular gene, one should identify potential 'off-target' genomic sequences by considering the following four constraints. First and foremost, they should not be followed by a PAM with either 5′-NGG or 5′-NAG sequences. Second, their global sequence similarity to the target sequence should be minimized, and guide sequences with genomic off-target loci that have fewer than three mismatches should be avoided. Third, at least two mismatches should lie within the PAM-proximal region of the off-target site. Fourth, a maximal number of mismatches should be consecutive or spaced less than four bases apart. Finally, the amount of SpCas9 and sgRNA can be titrated to optimize on- to off-target cleavage ratio.

Using these criteria, we formulated a scoring algorithm to integrate and quantify the contributions of mismatch location, density and identity on SpCas9 on-target and off-target cleavage. We applied the aggregate cleavage efficiencies of single-mismatch guide RNAs to test this scoring scheme separately on genome-wide targets and found that these factors, taken together, accounted for >50% of the variance in cutting-frequency rank among the genome-wide targets studied (Supplementary Fig. 11).

Implementing the guidelines delineated above, we designed a computational tool to facilitate the selection and validation of sgRNAs as well as to predict off-target loci for specificity analyses; this tool can be accessed at http://www.genome-engineering.org/. These results and tools further extend the SpCas9 system as a versatile alternative to ZFNs and TALENs for genome editing applications. Further work examining the thermodynamics and in vivo stability of sgRNA-DNA duplexes will likely yield additional predictive power for off-target activity, whereas exploration of SpCas9 mutants and orthologs may yield novel variants with improved specificity.

1.4 Methods

Cell culture and transfection

Human embryonic kidney (HEK) cell line 293FT (Life Technologies) was maintained in Dulbecco's modified Eagle's Medium (DMEM) supplemented with 10% FBS (HyClone), 2 mM GlutaMAX (Life Technologies), 100 U/ml penicillin, and 100 µg/ml streptomycin at 37 °C with 5% CO2 incubation.

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

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

SURVEYOR nuclease assay for genome modification

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

The genomic region flanking the CRISPR target site for each gene was PCR amplified, and products were purified using QiaQuick Spin Column (Qiagen) following the manufacturer's protocol. 400 ng total of the purified PCR products were mixed with 2 µl 10× Taq DNA Polymerase PCR buffer (Enzymatics) and ultrapure water to a final volume of 20 µl, and subjected to a re-annealing process to enable heteroduplex formation: 95 °C for 10 min, 95 °C to 85 °C ramping at −2 °C/s, 85 °C to 25 °C at −0.25 °C/s, and 25 °C hold for 1 min. After re-annealing, products were treated with SURVEYOR nuclease and SURVEYOR enhancer S (Transgenomics) following the manufacturer's recommended protocol, and analyzed on 4–20% Novex TBE polyacrylamide gels (Life Technologies). Gels were stained with SYBR Gold DNA stain (Life Technologies) for 30 min and imaged with a Gel Doc gel imaging system (Bio-rad). Quantification was based on relative band intensities. Indel percentage was determined by the formula, $100 \times (1 - (1 - (b + c)/(a + b + c))1/2)$, where a is the integrated intensity of the undigested PCR product, and b and c are the integrated intensities of each cleavage product.

Northern blot analysis of tracrRNA expression in human cells

Northern blots were done as previously described[23]. Briefly, RNAs were extracted using the mirPremier microRNA Isolation Kit (Sigma) and heated to 95 °C for 5 min before loading on 8% denaturing polyacrylamide gels (SequaGel, National Diagnostics). Afterwards, RNA was transferred to a Hybond N+ membrane (GE Healthcare) and crosslinked with Stratagene UV Crosslinker (Stratagene). Probes were labeled with (gamma-32P) ATP (PerkinElmer) with T4 polynucleotide kinase (New England Biolabs). After washing, membrane was exposed to phosphor screen for 1 h and scanned with phosphorimager (Typhoon).

Bisulfite sequencing to assess DNA methylation status

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

In vitro transcription and cleavage assay

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

Deep sequencing to assess targeting specificity

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

Barcoded and purified DNA samples were quantified by Quant-iT PicoGreen dsDNA Assay Kit or Qubit 2.0 Fluorometer (Life Technologies) and pooled in an equimolar ratio.

Sequencing libraries were then sequenced with the Illumina MiSeq Personal Sequencer (Life Technologies).

Sequencing data analysis and indel detection

MiSeq reads were filtered by requiring an average Phred quality (Q score) of at least 23, as well as perfect sequence matches to barcodes and amplicon forward primers. Reads from on- and off-target loci were analyzed by first performing Smith-Waterman alignments against amplicon sequences that included 50 nucleotides upstream and downstream of the target site (a total of 120 bp). Alignments, meanwhile, were analyzed for indels from 5 nucleotides upstream to 5 nucleotides downstream of the target site (a total of 30 bp). Analyzed target regions were discarded if part of their alignment fell outside the MiSeq read itself, or if matched base-pairs comprised less than 85% of their total length.

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

In order to place error bounds on the true-indel read frequencies in the sequencing libraries themselves, Wilson score intervals[48] were calculated for each sample, given the MLE-estimate for true-indel target-regions, Rp, and the number of reads R.

qRT-PCR analysis of relative Cas9 and sgRNA expression

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

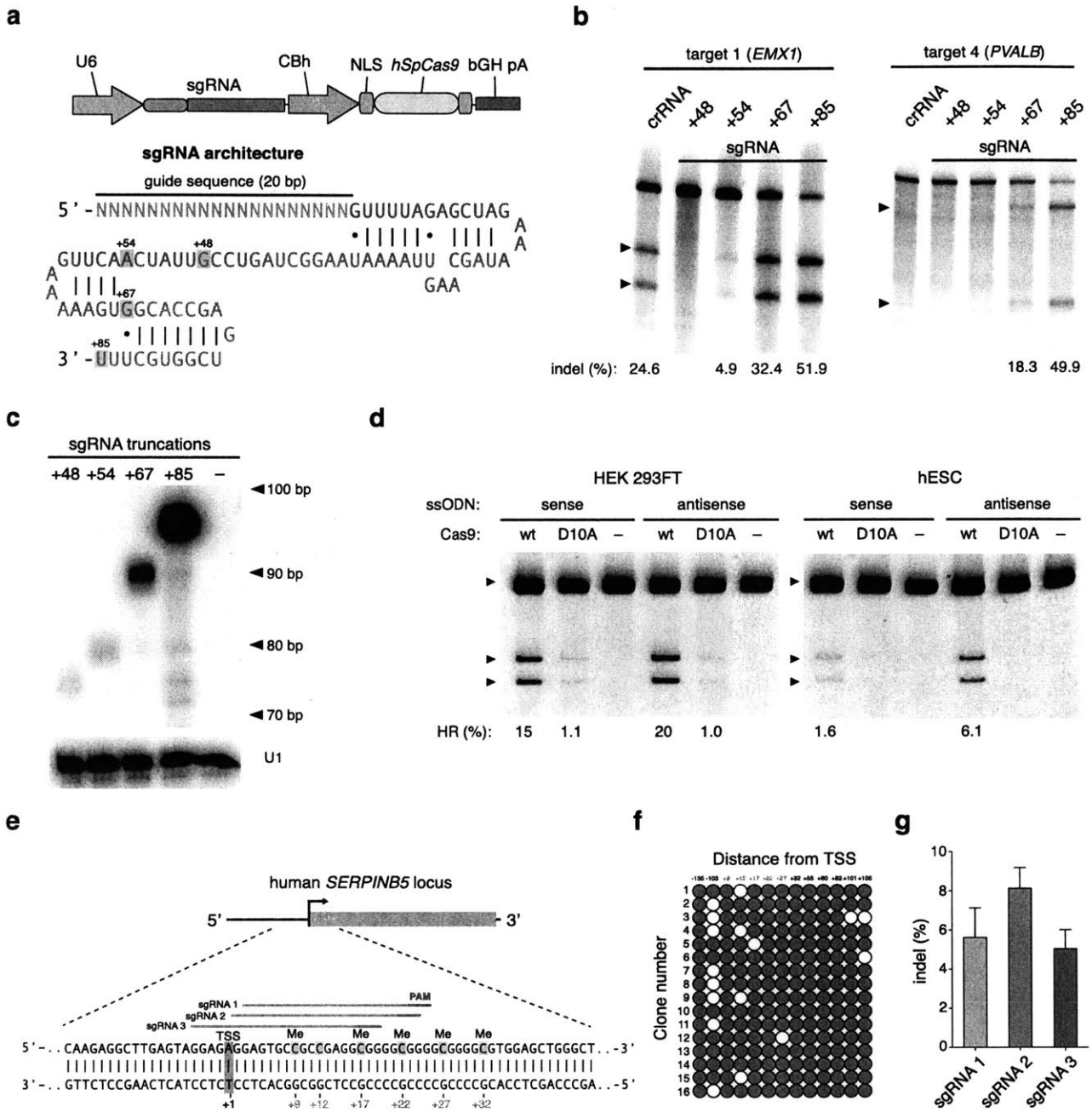


Figure 1: Optimization of guide RNA architecture for SpCas9-mediated mammalian genome editing. (a) Schematic of bicistronic expression vector (PX330) for U6 promoter-driven sgRNA and CBh promoter-driven human codon-optimized S. pyogenes Cas9 (hSpCas9) used for all subsequent experiments. The sgRNA consists of a 20-nt guide sequence (blue) and scaffold (red), truncated at various positions as indicated. (b) SURVEYOR assay for SpCas9-mediated indels at the human EMX1 and PVALB loci. Arrowheads indicate the expected SURVEYOR fragments (n = 3). (c) Northern blot analysis for the four sgRNA truncation architectures, with U1 as loading control. (d) Both wild-type (WT) or nickase mutant (D10A) of SpCas9 promoted insertion

of a HindIII site into the human EMX1 gene. Single-stranded oligonucleotides, oriented in either the sense or antisense direction relative to genome sequence, were used as homologous recombination templates (Supplementary Fig. 3). (e) Schematic of the human SERPINB5 locus. sgRNAs and PAMs are indicated by colored bars above sequence; methylcytosine (Me) are highlighted (pink) and numbered relative to the transcriptional start site (TSS, +1). (f) Methylation status of SERPINB5 assayed by bisulfite sequencing of 16 clones. Filled circles, methylated CpG; open circles, unmethylated CpG. (g) Modification efficiency by three sgRNAs targeting the methylated region of SERPINB5, assayed by deep sequencing (n = 2). Error bars indicate Wilson intervals (Methods).

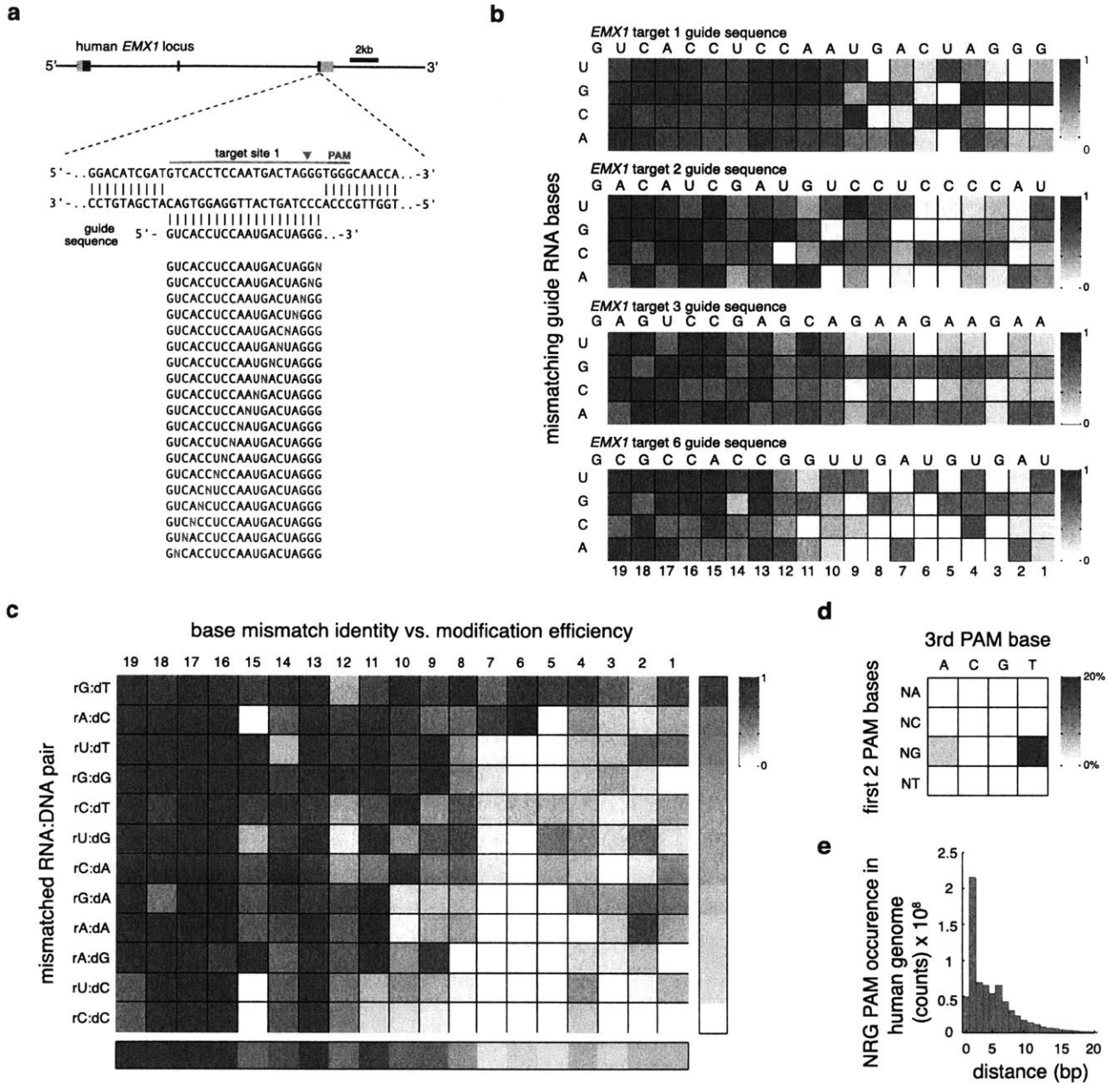


Figure 2: Single-nucleotide specificity of SpCas9. (a) Schematic of the experimental design. sgRNAs carrying all possible single base-pair mismatches (blue Ns) throughout the guide sequence were tested for each EMX1 target site (target site 1 shown as example). (b) Heatmap representation of relative SpCas9 cleavage efficiency by 57 single-mutated and 1 nonmutated sgRNA each for four EMX1 target sites. For each EMX1 target, the identities of single base-pair substitutions are indicated on the left; original guide sequence is shown above and highlighted in the heatmap (gray squares). Modification efficiencies (increasing from white to dark blue) are normalized to the original guide sequence. Sequence logo representation of the same data can be found in Supplementary Figure 7. (c) Heatmap for relative SpCas9 cleavage efficiency for

each possible RNA:DNA base pair, compiled from aggregate data from single-mismatch guide RNAs for 15 EMX1 targets (Supplementary Fig. 8). Mean cleavage levels were calculated for the 10 PAM-proximal bases (right bar) and across all substitutions at each position (bottom bar); positions in gray were not covered by the 469 single-mutated and 15 unmutated sgRNAs tested. (d) SpCas9-mediated indel frequencies at targets with all possible PAM sequences, determined using the SURVEYOR nuclease assay. Two target sites from the EMX1 locus were tested for each PAM. (e) Histogram of distances between 5′-NRG PAM occurrences within the human genome. Putative targets were identified using both strands of human chromosomal sequences (GRCh37/hg19).

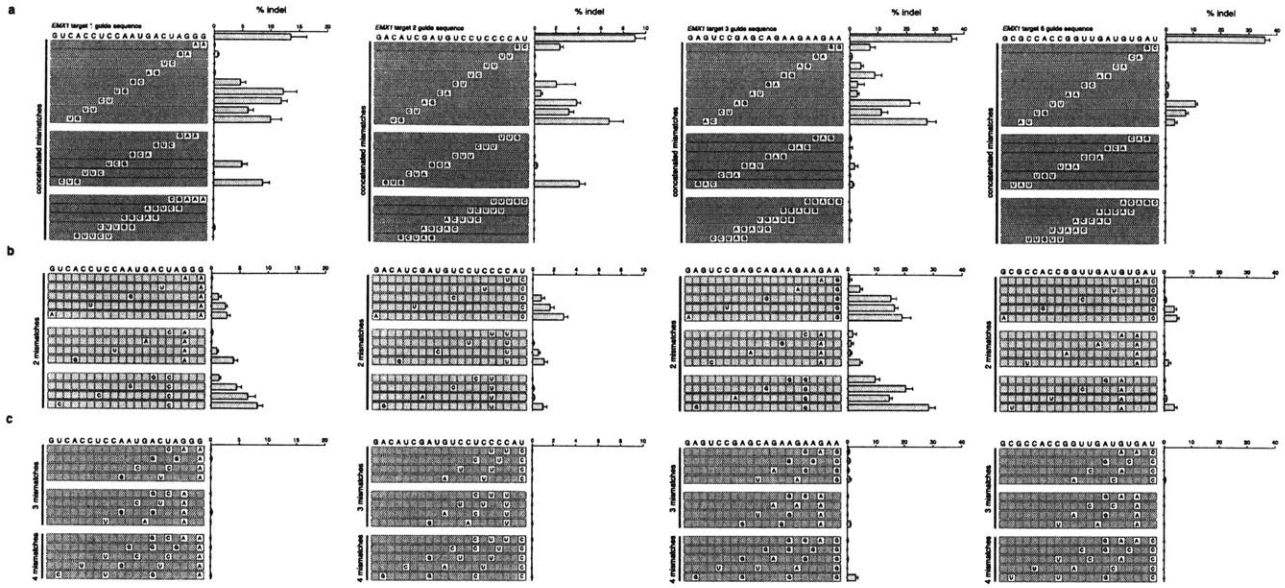


Figure 3: Multiple mismatch specificity of SpCas9. (a–c) SpCas9 cleavage efficiency with guide RNAs containing consecutive mismatches of 2, 3 or 5 bases (a), or multiple mismatches separated by different numbers of unmutated bases for EMX1 targets 1, 2, 3 and 6 (b,c). Rows represent each mutated guide RNA; nucleotide substitutions are shown in white cells; gray cells denote unmutated bases. All indel frequencies are absolute and analyzed by deep sequencing from two biological replicas. Error bars indicate Wilson intervals

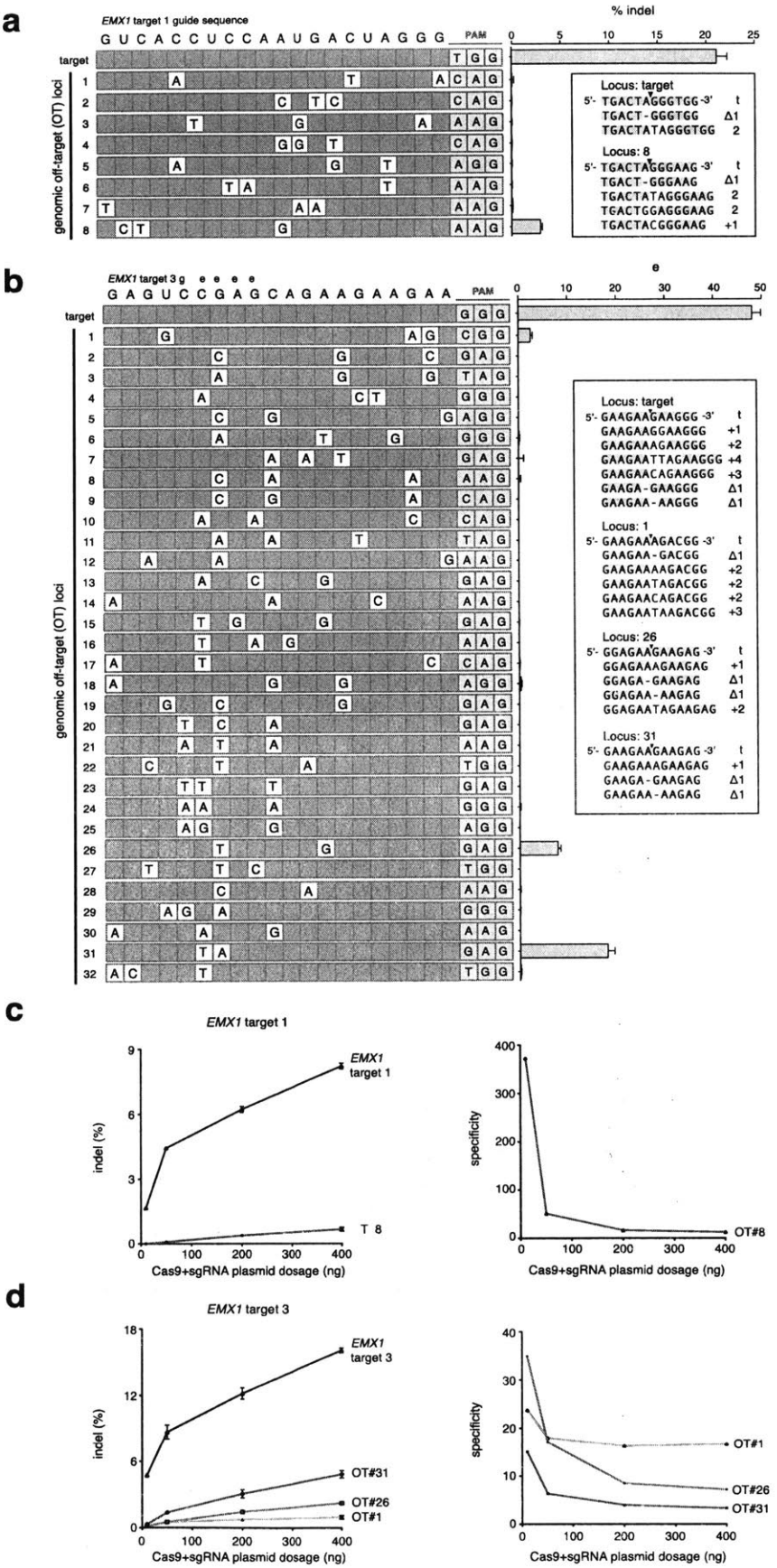
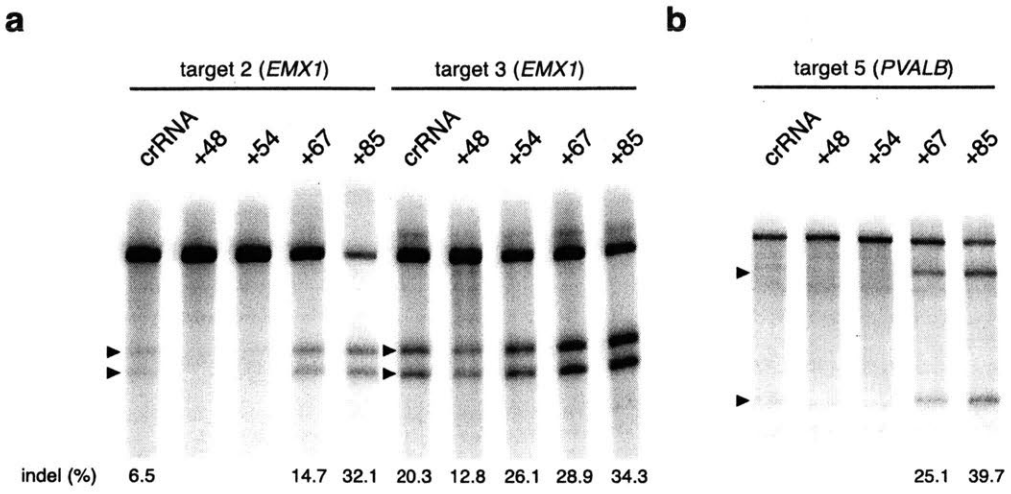
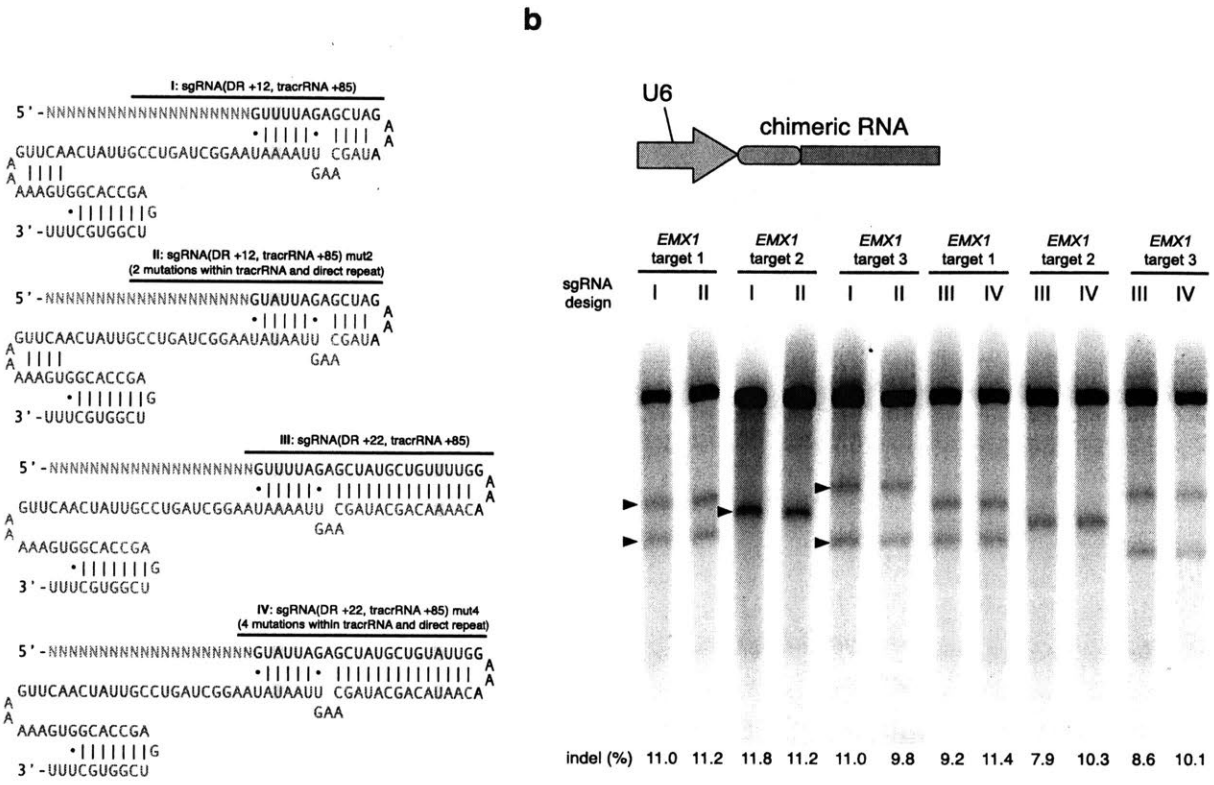

a
EMX1 target 1 guide sequence
G U C A C C U C C A A U G A C U A G G G
PAM
% indel
target
genomic off-target (OT) loci
Locus: target
5'- TGACTAGGGTGG -3' t
TGACT-GGGTGG Δ1
TGACTATAGGGTGG 2
Locus: 8
5'- TGACTAGGGAAG -3' t
TGACT-GGGAAG Δ1
TGACTATAGGGAAG 2
TGACTGGAGGGAAG 2
TGACTACGGGAAG +1
b
EMX1 target 3 g
G A G U C C G A G C A G A A G A A G A A
PAM
target
genomic off-target (OT) loci
Locus: target
5'- GAAGAAGAAGGG -3' t
GAAGAAGGAAGGG +1
GAAGAAAGAAGGG +2
GAAGAATTAGAAGGG +4
GAAGAACAGAAGGG +3
GAAGA-GAAGGG Δ1
GAAGAA-AAGGG Δ1
Locus: 1
5'- GAAGAAAGACGG -3' t
GAAGAA-GACGG Δ1
GAAGAAAAGACGG +2
GAAGAATAGACGG +2
GAAGAACAGACGG +2
GAAGAATAAGACGG +3
Locus: 26
5'- GGAGAAGAAGAG -3' t
GGAGAAAGAAGAG +1
GGAGA-GAAGAG Δ1
GGAGAA-AAGAG Δ1
GGAGAATAGAAGAG +2
Locus: 31
5'- GAAGAAGAAGAG -3' t
GAAGAAAGAAGAG +1
GAAGA-GAAGAG Δ1
GAAGAA-AAGAG Δ1
c
EMX1 target 1
EMX1 target 1
T 8
indel (%)
Cas9+sgRNA plasmid dosage (ng)
specificity
OT#8
Cas9+sgRNA plasmid dosage (ng)
d
EMX1 target 3
EMX1 target 3
OT#31
OT#26
OT#1
indel (%)
Cas9+sgRNA plasmid dosage (ng)
specificity
OT#1
OT#26
OT#31
Cas9+sgRNA plasmid dosage (ng)

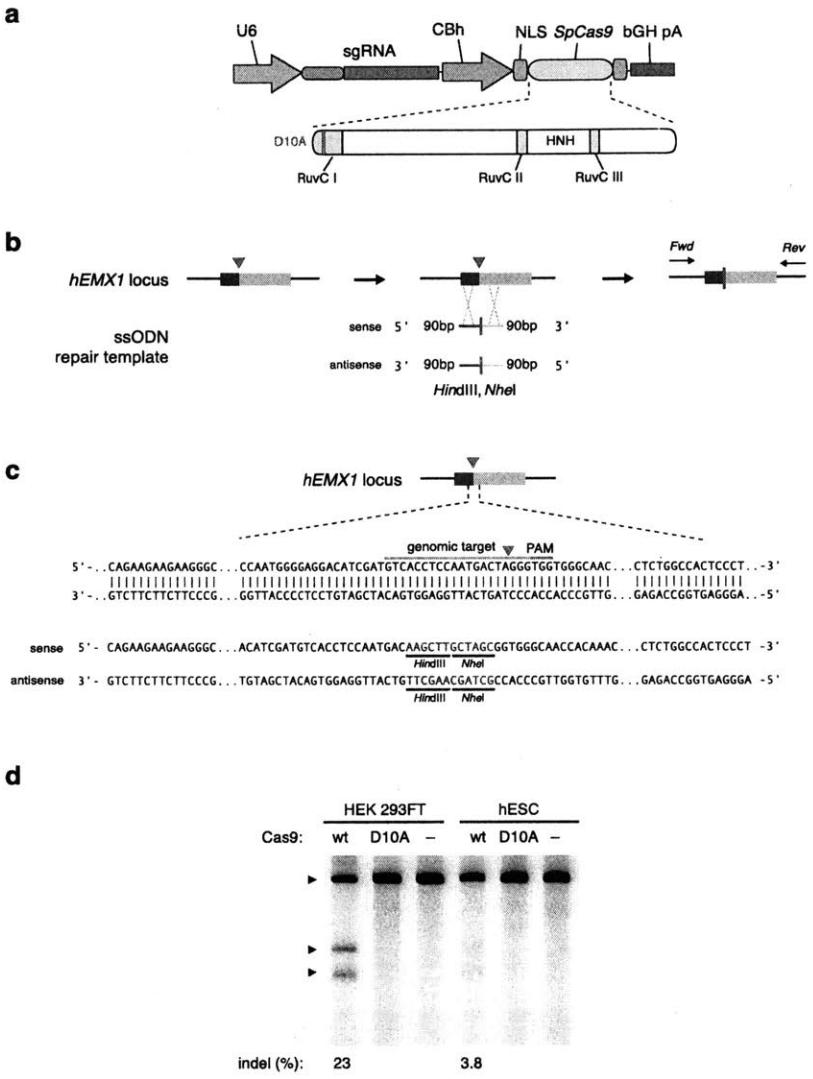
Figure 4: SpCas9-mediated indel frequencies at predicted genomic off-target loci. (a,b) Cleavage levels at putative genomic off-target loci containing two or three individual mismatches (white cells) for EMX1 target 1 and target 3 are analyzed by deep sequencing. List of off-target sites are ordered by median position of mutations. Putative off-target sites with additional mutations did not have detectable indels. The Cas9 dosage was 3 × 10−10 nmol/cell, with equimolar sgRNA delivery. Error bars indicate Wilson intervals (Online Methods). (c,d) Indel frequencies for EMX1 targets 1 and 3 and selected off-target loci (OT) as a function of SpCas9 and sgRNA dosage, (n = 2, Wilson intervals). 400 ng to 10 ng of Cas9-sgRNA plasmid corresponds to 7.1 × 10−10 to 1.8 × 10−11 nmol/cell. Cleavage specificity is measured as a ratio of on- to off-target cleavage.



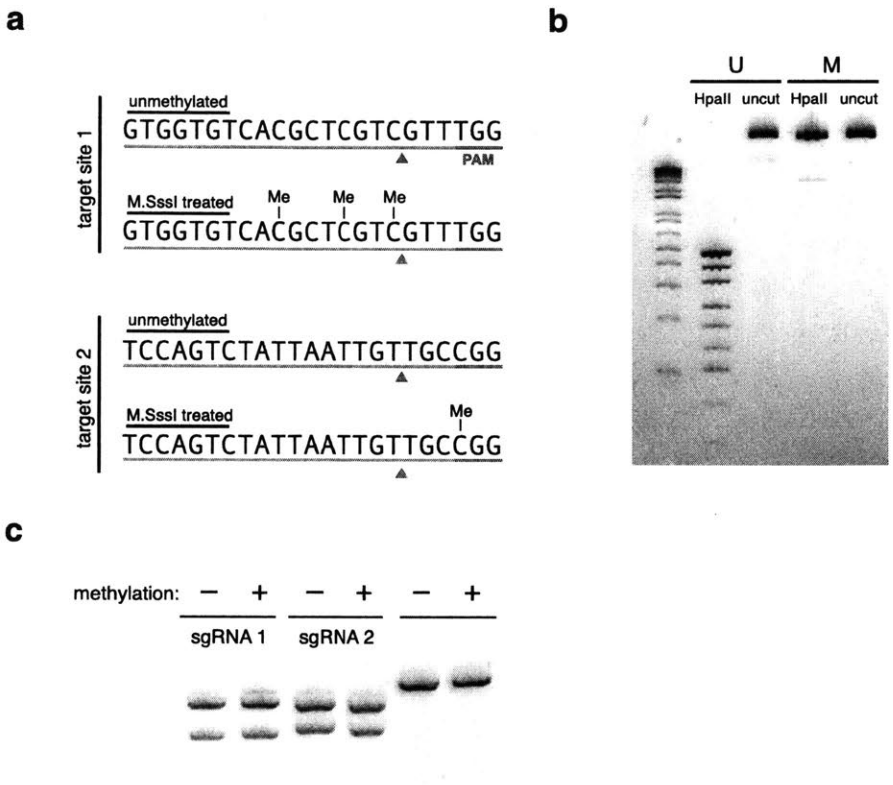
Supplementary Figure 1: Modification efficiencies of CRISPR-Cas system for additional human genomic targets. DNA expression vectors carrying SpCas9 and crRNA-tracrRNA pair or single guide RNA (sgRNA) are co-transfected into 293FT cells. Cleavage efficiency (% indel) is assessed using the SURVEYOR nuclease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



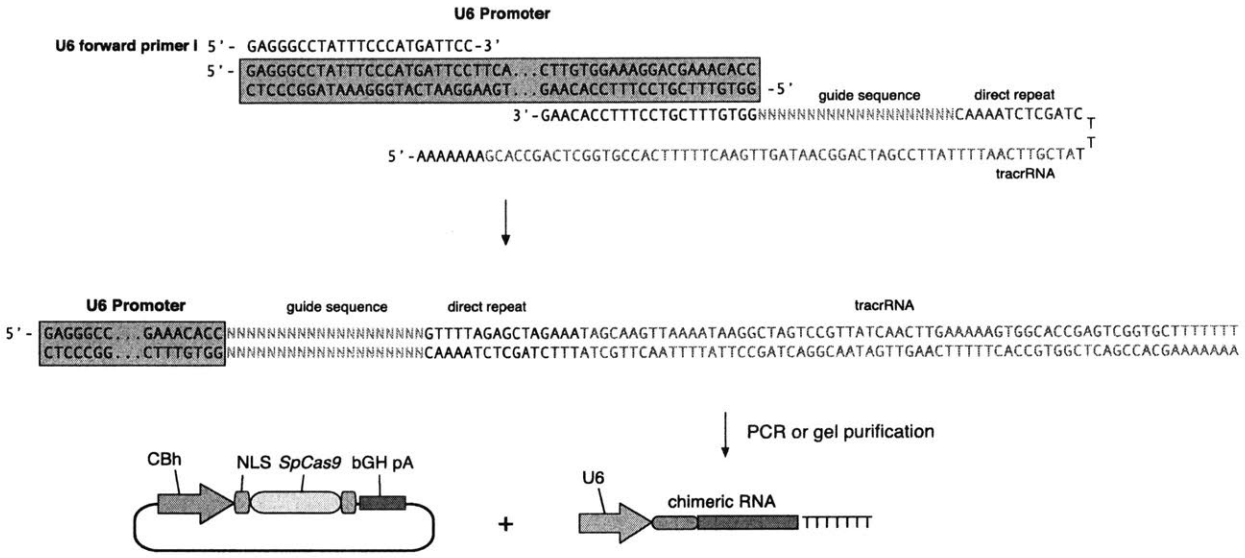
Supplementary Figure 2: Further optimization of CRISPR-Cas sgRNA architecture. a,Schematic of four additional sgRNA architectures, I-IV. Each consists of a 20-nt guide sequence(blue) joined to the direct repeat (DR, grey), which base-pairs with the tracrRNA (red). The DRtracrRNA hybrid is truncated at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



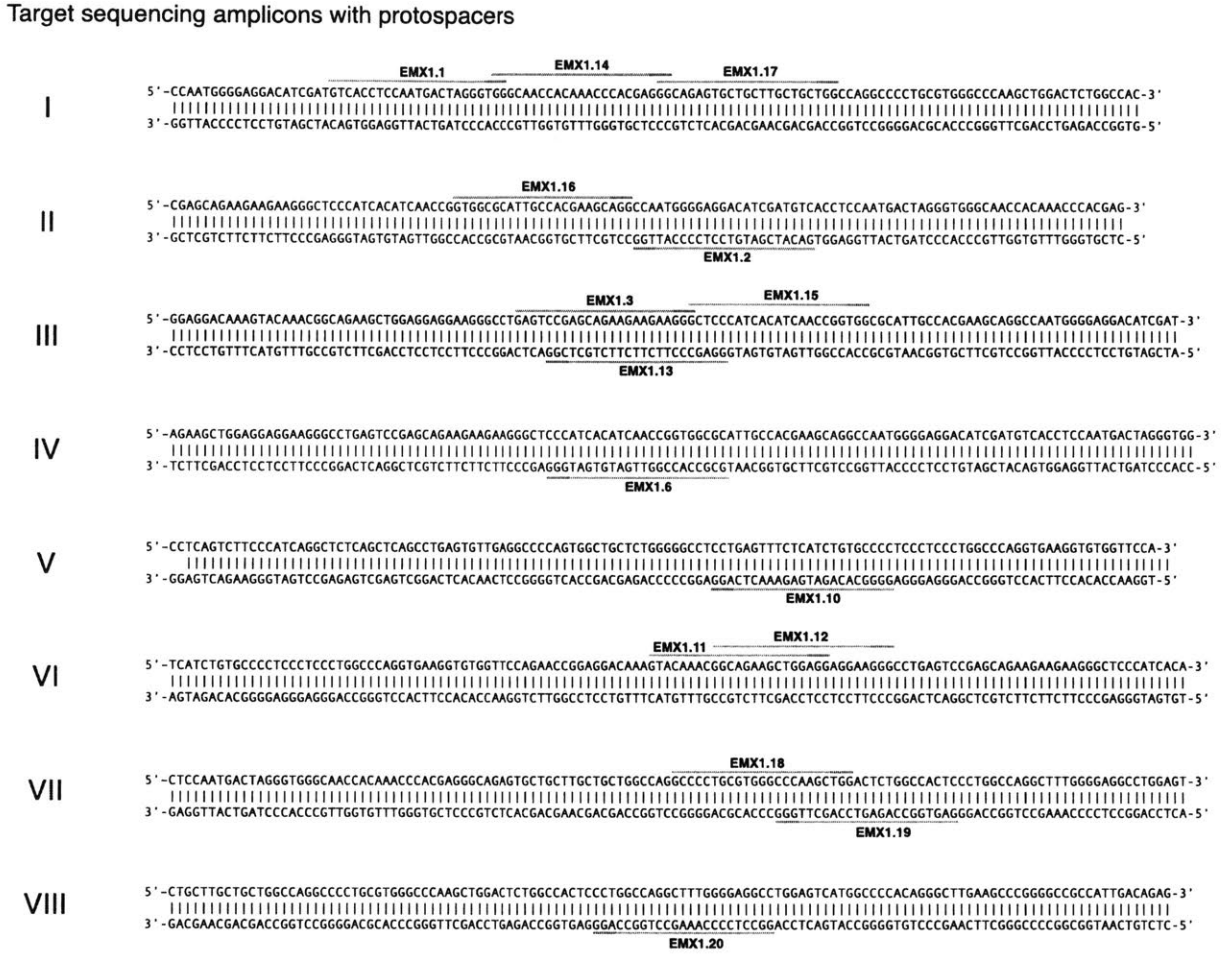
Supplementary Figure 3: Genome editing via homologous recombination. a, Schematic of SpCas9 nickase, with D10A mutation in the RuvC I catalytic domain. b, Schematic representing homologous recombination (HR) at the human EMX1 locus using either sense or antisense single stranded oligonucleotides as repair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



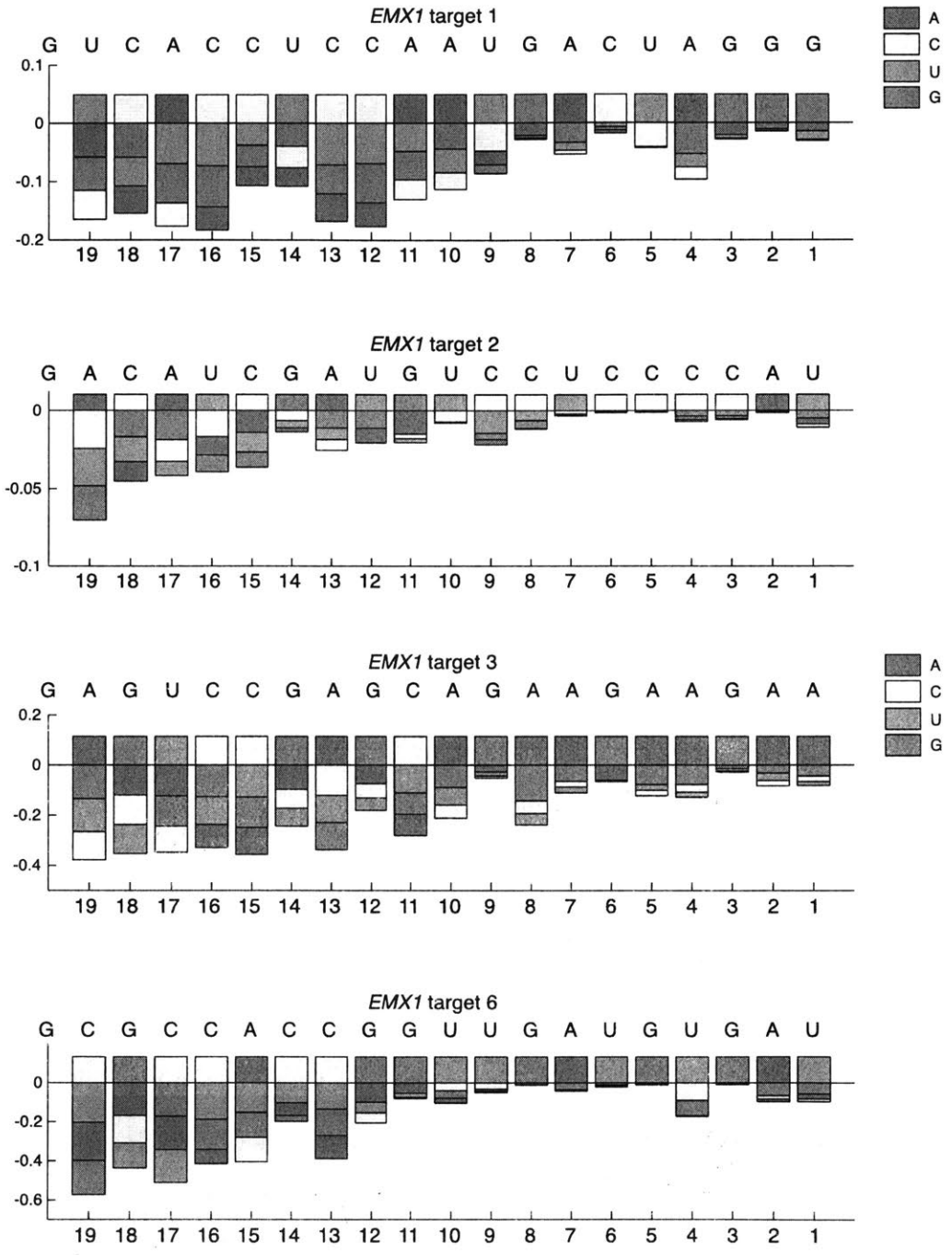
Supplementary Figure 4: SpCas9 cleaves methylated targets in vitro. a, Sequence of CpG dinucleotide-containing targets in pUC19 plasmid methylated in vitro by M.SssI. Methyl-CpGs in either the target sequence or PAM are indicated; arrows indicate expected cleavage site. b, Unmethylated (U) or methylated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



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

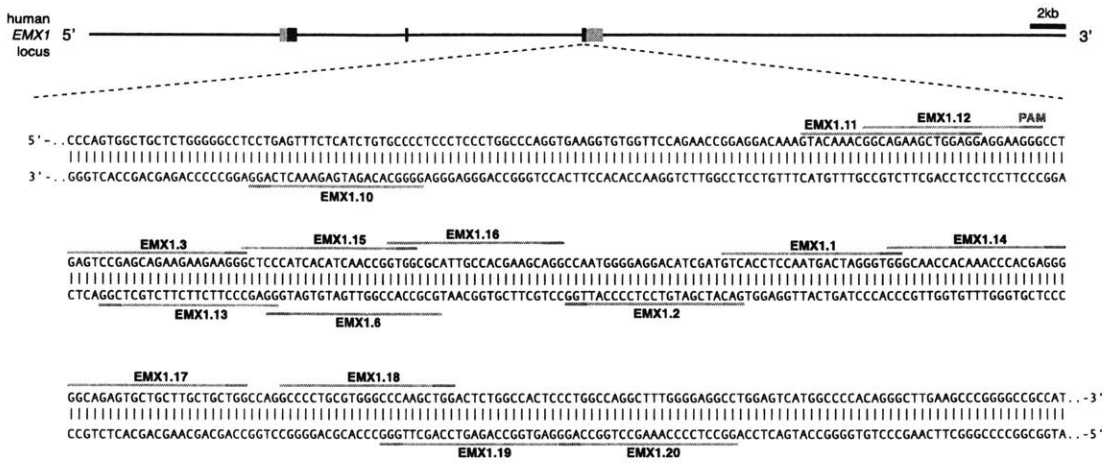


Supplementary Figure 6: The human EMX1 locus with target sites. Schematic of the human EMX1 locus showing the location of 15 target DNA sites, indicated by blue lines with corresponding PAM in magenta.



Supplementary Figure 7: Base frequency plots of relative SpCas9 cleavage efficiency for four EMX1 target sites. Relative contribution of each base per guide sequence position to SpCas9 cleavage efficiency for EMX1 targets 1, 2, 3, and 6. Modification efficiencies are normalized to cleavage levels mediated by the original guide sequence.

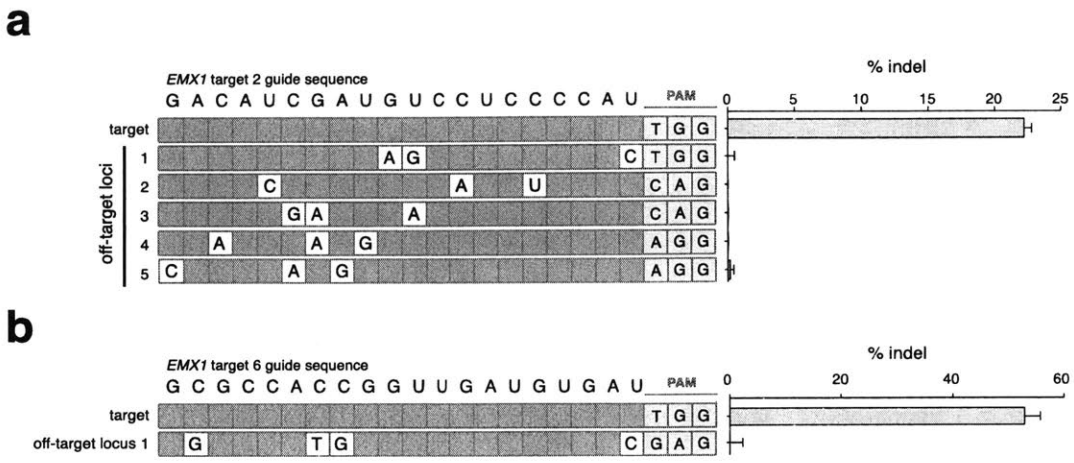
a



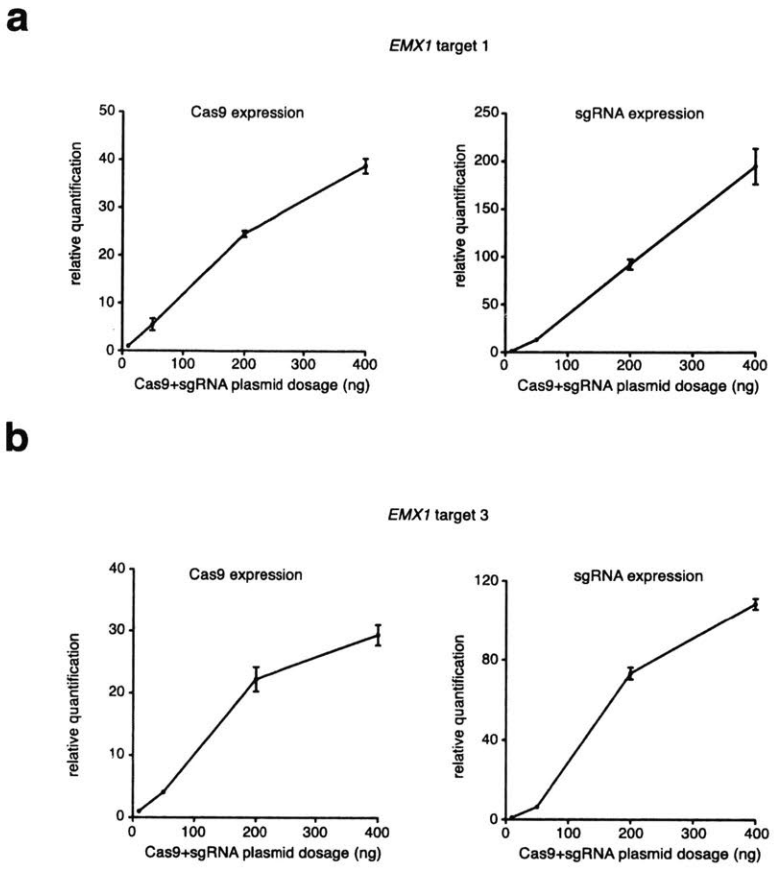
b

target species	gene	protospacer ID	target site (5' to 3')	PAM	strand
Homo sapiens	*EMX1*	1	GTCACCTCCAATGACTAGGG	TGG	+
	EMX1	2	GACATCGATGTCCTCCCCAT	TGG	–
	EMX1	3	GAGTCCGAGCAGAAGAAGAA	GGG	+
	EMX1	6	GCGCCACCGGTTGATGTGAT	GGG	–
	EMX1	10	GGGGCACAGATGAGAAACTC	AGG	–
	EMX1	11	GTACAAACGGCAGAAGCTGG	AGG	+
	EMX1	12	GGCAGAAGCTGGAGGAGGAA	GGG	+
	EMX1	13	GGAGCCCTTCTTCTTCTGCT	CGG	–
	EMX1	14	GGGCAACCACAAACCCACGA	GGG	+
	EMX1	15	GCTCCCATCACATCAACCGG	TGG	+
	EMX1	16	GTGGCGCATTGCCACGAAGC	AGG	+
	EMX1	17	GGCAGAGTGCTGCTTGCTGC	TGG	+
	EMX1	18	GCCCCTGCGTGGGCCCAAGC	TGG	+
	EMX1	19	GAGTGGCCAGAGTCCAGCTT	GGG	–
	EMX1	20	GGCCTCCCCAAAGCCTGGCC	AGG	–

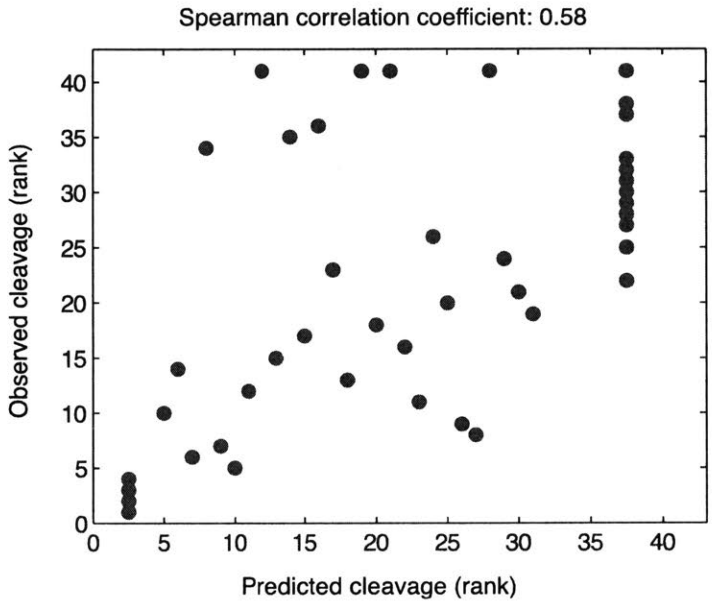
Supplementary Figure 8: Target amplicon sequencing for assessing CRISPR-Cas modification efficiency. a, Schematic of the human EMX1 locus and a target site of interest. Target amplicons are PCR-amplified by a fusion PCR method to add sequencing adapters for the deep sequencing. Each sample is uniquely barcoded for multiplexed sequencing and pooled in an equimolar ratio into a sequencing library. b, target sequencing amplicons with associated target sites. Experimental samples are tracked by their barcode and amplicon identities.



Supplementary Figure 9: Additional genomic off-target site analysis. Cleavage levels at candidate genomic off-target loci (white cells indicating mismatches) for a, EMX1 target 2 and b, EMX1 target 6 were analyzed by deep sequencing. All indel frequencies are absolute and analyzed by deep sequencing from 2 biological replicates. Error bars indicate Wilson confidence intervals (Supplementary Methods).



Supplementary Figure 10: qRT-PCR analysis of Cas9 and sgRNA expression. Relative mRNA expression levels of hSpCas9 and sgRNA(+85) at different DNA transfection dosages for a, EMX1 target 1 and b, EMX1 target 3. The DNA transfected contains both hSpCas9 as well as the corresponding sgRNA on a single plasmid (guide sequence cloned into pX330). Data are plotted as the mean of independent biological replicates (n = 3) and relative quantification over the lowest SpCas9 dosage.



Supplementary Figure 11: Predicted and observed cutting frequency-ranks among genome-wide targets.

CHAPTER 2: Exploring the utility of spCas9 binding and NHEJ event labeling for the unbiased genome wide detection of editing.

Adapted from:

Wu, X., Scott, D. A., et al. Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. Nat. Biotechnol. 32, 670–676 (2014).

Scott, D. A., Smargon, A., et al. Unbiased genome-wide detection of Cas9 nuclease specificity using Insert Capture. Unpublished.

Having characterized the mechanism of Cas9 on and off target recognition for nuclease activity and used targeted Next Generation Sequencing to investigate Cas9 off-target activity in the human genome[49], we began investigating Cas9 specificity in an unbiased fashion. To investigate the efficacy of Cas9 ChIP binding assays for predicting Cas9 off-target nuclease activity, catalytically inactive Cas9 (dCas9) ChIP was used to identify Cas9 binding loci for individual targets. While off-target binding sites for dCas9 numbered in the thousands for many RNA guides, treatment of cells with the same RNA guides and catalytically active Cas9 resulted in undetectable indels using our previously established targeted NGS assay with an indel detection sensitivity of approximately 0.1 – 0.3%. In addition to investigating the efficacy of dCas9 ChIP for the unbiased characterization of off-target activity, we developed insert capture assays for the sensitive labeling of NHEJ events in cells following Cas9 genomic manipulation. Insert capture events resulting from DSBs are stably propagated in the genome, allowing accumulation of signal from DSBs over the duration of genome editing in dividing cells.

We show that insert capture is a sensitive method for the unbiased genome-wide detection of Cas9 off-target modification.

2.1 Introduction:

In silico algorithms for predicting CRISPR-Cas9 off-targets suffer from both high numbers of false positives and false negatives, and it has been shown that numerous off-target sites predicted using these algorithms show no detectable editing when experimentally measured[49]. Additionally, these algorithms fail to detect multiple off targets that show experimentally validated editing[50]. In general, available computational algorithms for off-target detection provide minimal power for off target prediction, besides providing a basic estimate of how many similar sequences exist in the genome. While in the future training better algorithms based on expanding data sets for off target activity may resolve these problems, such efforts will require comprehensive experimental validation of off-target activity using unbiased methods with high sensitivity and throughput.

Unbiased methods for the detection of nuclease activity can either measure DNA cleavage events by directly labeling free DNA ends resulting from a double strand break, precursory events predictive of a double strand break formation, or outcomes of targeted double strand break repair. Programmable nucleases fundamentally consist of a DNA binding component, such that target binding precedes DNA cleavage by the nuclease component of the protein. Combined with enthusiasm for using catalytically dead CRISPR/Cas9 (dCas9) as a DNA binding protein for epigenetic and transcriptional regulation, we sought to both characterize the specificity of Cas9 DNA binding and

investigate whether Cas9 off-target binding is predictive of Cas9 off-target DNA cleavage.

Following Cas9 DNA cleavage, NHEJ is the most prevalent form of DNA double strand break repair in mammalian cells[14]. Individual double strand breaks do not occur in isolation in a cell, and when multiple breaks occur simultaneously, different outcomes of DNA repair can occur, including: 1) perfect re-ligation of DNA ends at a single break site, 2) imperfect re-ligation of the broken DNA ends resulting in a short deletion or insertion (indel), 3) ligation of DNA ends from different DSBs occurring on the same chromosome (chromosomal rearrangement), or 4) ligation of DNA ends from DSBs occurring on separate chromosomes (translocation). During repair of double strand breaks, it has been previously shown that exogenous integrase deficient lentiviral linear DNA (IDLV) is inserted at low frequency between cut DNA ends, allowing unbiased genome-wide mapping of ZFN off-target cleavage[22]. However, current implementations of IDLV insertion enable only disparate identification of genomic DNA flanking each end of the IDLV. Hence, it is impossible to distinguish between an indel, a chromosomal rearrangement, and a translocation event using IDLV.

Here we show that DNA binding is not predictive of genome modification by Cas9 and present a two-state model describing target interrogation and cleavage by Cas9. In an effort to provide an efficient and unbiased genome wide evaluation of each different type of DNA repair, we developed insert capture. The insert capture approach relies on the insertion of short DNA sequences at double strand breaks and paired identification of the genomic sequence flanking both sides of the DNA insert. We show that insert capture not only provides unbiased genome wide detection of double strand breaks

sites but recapitulation of the frequency of indel formation, chromosomal rearrangement, and translocation accompanying genome modification with Cas9.

2.2.1 Results: Implications of spCas9 binding for targeted nuclease activity

To test if dCas9 binding correlates with Cas9 nuclease–induced mutation, we examined the indel frequencies of the four on-target sites and 295 selected off-target sites by targeted PCR and sequencing[49]. These sites were selected to cover a broad range of binding levels and numbers of mismatches to the sgRNA: we ranked all peaks by binding (background-subtracted read counts) and, for each binding level, selected a peak with the fewest mismatches and another peak with most mismatches to the guide.

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

Using our previously validated model[49], the background indel frequencies due to sequencing errors were determined for each individual target using two biological replicates transfected with only Cas9 but no sgRNA (control). Importantly the control samples showed no evidence of targeted mutations by Cas9 (note that background indels in the absence of Cas9 might also occur). We manually reviewed sequencing

alignments of all loci with indel frequencies >0.03%. We found that 12–37% of sequencing reads from the on-target sites contained indels. One off-target site, which was from Nanog-sg2, was mutated at a frequency of 0.7% (Fig. 1). There was no detectable correlation between binding and indel frequency (sites in Fig. 1 are ranked by decreasing binding from left to right for each sgRNA). The selected sites include 7 of the top 10 (including all the top 6) and 36 of the top 50 Nanog-sg3 binding sites with the strongest ChIP signals, and 4 of the 8 Nanog-sg3 off-target binding sites that had fewer than four mismatches to the sgRNA; none of these off-target sites showed cleavage significantly above the background level. Cumulatively, these results indicate a two state model for Cas9 DNA cleavage, where following DNA unwinding at the target PAM and subsequent target binding, additional thermodynamic barriers exist to DNA cleavage by Cas9 (Fig. 2).

2.2.2 Results: Unbiased detection of spCas9 off-target activity by Insert Capture

DNA inserts contain a unique molecular identifier (UMI) composed of 5 - 30 random nucleotides used to encode individual insertion events (18 random nucleotides for the current implementation), here forward referred to as the event barcode (EBC). Importantly, the EBC allows for the identification of unique insertion events at individual genomic loci, and determination of the outcome of NHEJ (indel, chromosomal rearrangement, or translocation). EBCs are flanked on both sides by two priming sites for extraction of genomic sequences on either side of the insert by non-restrictive linear amplification (nrLAM)[51] and later exponential amplification (Fig. 3a). To allow for specific amplification from each individual priming site on the DNA insert, all priming sites contain different sequences.

Double stranded DNA inserts (dsDNA insert) as well as single stranded DNA inserts containing a short 3` double stranded region (ssDNA insert) show similar efficiencies of capture at DSB sites in the genome. Following incorporation into the genome by NHEJ, inner priming sites (P1a and P1b) on either side of the EBC facilitate the use of non-restrictive linear amplification to capture the EBC and genomic sequence flanking both sides of the insertion site in a single stranded DNA (ssDNA) amplicon. Use of biotinylated primers for nrLAM enables purification of ssDNA amplicons on biotin/streptavidin beads. Thereafter, an ssDNA adapter is ligated to the unknown end of the ssDNA amplicon, and a 5' RACE PCR strategy is used to prepare amplicons for next generation sequencing (P2a/P3 or P2b/P3 priming sites).

During experiments, 2-3x10^5 293T cells were transfected with Cas9, sgRNA, and DNA insert, and passaged for 24 – 30 days prior to isolation of genomic DNA (Fig. 3c). Insert capture events showed minimal or no truncation of the insert and no evidence of homology between the insert and surrounding genomic loci, indicating that inserts are primarily integrated into the genome during non-homologous end joining (NHEJ) repair at DSBs. Expansion of cell populations containing inserts amplifies representation of individual insertion events. Cell populations are then split into two fractions, each containing representation for the amplified insert capture events. Uniquely specifying each insertion event, pairing of EBC sequences in for sequences captured from each fraction enables the reconstruction of the complete NHEJ junction containing an insert (Fig. 3e). This enables the classification of each insertion event as an indel, chromosomal rearrangement, or translocation event. Additionally, extraction of DNA at multiple time points allows for tracking of individual EBC-labeled clones at multiple time points and can be used to assess the phenotypic effects of mutations at individual genomic loci. Cumulatively, insert capture enables the comprehensive assessment of DSB sites, NHEJ outcomes, and phenotypic effects associated with off-target events.

We used insert capture to study off-target genome modification by Cas9 targeting the VEGFA3 locus. More than 2000 unique insertion events were observed at the VEGFA3 on-target locus, and multiple off-target loci with high homology to the VEGFA3 on-target were identified with decreasing enrichment (Fig. 4). Notably, these loci included all sites computationally predicted and previously validated for Cas9 off-target activity as well as additional loci not observed in previous studies[49,52]. Validation of cutting frequencies at

these loci showed that the frequency of insert capture events correlates with the frequency of genomic modification identified at on and off-target loci (Fig. 5).

To investigate genomic rearrangement resulting from expression of Cas9 and VEGFA3 sgRNA, NHEJ junctions were reconstructed by matching EBCs in sequencing data for split fractions of cells used to capture the 3' or 5' genomic sequence of insert capture events (Fig. 6, 7). This data provided an unbiased map of chromosomal rearrangement and translocation events occurring between any two genomic loci for each sample. Genomic rearrangement events associated with Cas9 on or off-target genomic modification were filtered by selecting all chromosomal rearrangement or translocation events where one element of the junction mapped to either a VEGFA3 on or off-target locus validated above (Fig. 6, 7, vertical red lines: on-target, solid; off-target, dotted). This analysis demonstrates that Cas9 on and off-target modification results in genomic rearrangement between Cas9 on and off-targets as well as between Cas9 on or off-targets and spontaneous double strand breaks occurring elsewhere in the genome.

2.3 Discussion

Our results show that Cas9 DNA binding activity is not predictive of Cas9 DNA endonuclease activity. The observation that most of the sites bound by Cas9 do not seem to be cleaved substantially is reminiscent of the eukaryotic Argonaute-microRNA system, in which most target mRNAs bearing partial microRNA matches are bound without cleavage and only a few targets with extensive pairing are cleaved[53]. We propose a two-state model (Fig. 2) similar to the Argonaute-microRNA system, in which pairing of a short seed region triggers binding after PAM recognition and subsequent DNA unwinding. In this model, targets with only seed complementarity remain bound by Cas9 without cleavage; only those with extensive pairing undergo efficient cleavage. This suggests a conformation change between binding and cleavage as observed for Argonaute-microRNA complexes[53,54]. While this paper was under review, a pair of Cas9 structural studies were published[55,56], including a crystal structure of dCas9 in complex with sgRNA and target DNA, which not only supports our observation of a PAM-proximal 5-nucleotide seed but also suggests a large conformation change during the inactive-active state transition[56].

While the dCas9 binding at off-target loci does not appear enable the prediction of sites for Cas9 off-target genome modification, we found that insert capture provides sensitive detection of NHEJ events at sites of on and off-target genome manipulation by Cas9. In addition to the detection of individual double strand break sites, insert capture enables comparison of the relative rates of indel, chromosomal rearrangement, and

translocation accompanying on-target and off-target manipulation at different sites in the genome.

Significant drawbacks of techniques such as insert capture and IDLV insertion are two-fold. 1) these approaches are inherently limited in absolute sensitivity due to the requirement for low probability errors in NHEJ in order to capture exogenous DNA fragments. Ideally unbiased assays for the genome wide detection of on and off-target DNA cleavage should be capable of detecting a single cleavage event. 2) These approaches are largely limited to applications in model cell lines where exogenous DNA inserts or IDLV particles can be efficiently delivered to cells in high concentrations. The use of such artificial cell lines for the analysis of the targeting specificity of programmable nucleases may be biased due to specific contextual factors associated with model cell lines, such as chromatin structure, rapid cell division, and differences in DNA repair. Taking these considerations into account, we refocused our efforts on the development of direct DNA double strand break labeling technologies as will be discussed in the next chapter. A similar approach to insert capture, GUIDEseq, has been subsequently published by the group of Dr. Keith Joung[57]. However, GUIDEseq enables only the mapping of DNA double strand break sites and not assessment of relative rates of indel, chromosomal rearrangement, and translocation.

2.4 Methods

Chromatin immunoprecipitation and analysis

Cells were passaged at 24 h post-transfection into a 150-mm dish, and fixed for ChIP processing at 48 h post-transfection. For each condition, 10 million cells are used for ChIP input, following experimental protocols and analyses as previously described[58] with the following modifications: instead of pairwise peak-calling, ChIP peaks were only required to be enriched over both 'empty' controls (dSpCas9 only, dSaCas9 only) as well as the other Cas9/other sgRNA sample (for example, SpCas9/EMX-sg2 peaks must be enriched over SaCas9/EMX-sg1 peaks in addition to the empty controls). This was done to avoid filtering out of real peaks present in two related samples as much as possible.

To identify off-targets ranked by motif or sequence similarity to guide, motif scores for ChIP peaks were calculated as follows: for a given ChIP peak, the 100-nucleotide interval around the peak summit, the target sequence, and a given sgRNA guide region of length L, the query, an alignment score is calculated for every subsequence of L in the target. The subsequence with the highest score is reported as the best match to the query. For each subsequence alignment, the score calculation begins at the 5′ end of the query. For each position in the alignment, 1 is added or subtracted for match or mismatch between the query and target, respectively. If the score becomes negative, it is set to 0 and the calculation continued for the remainder of the alignment. The score at the 3′ end of the query is reported as the final score for the alignment. MACS scores =

−10log(P value relative to the empty control) are determined as previously described[59]. For unbiased determination of PAM from ChIP peaks, the peaks were analysed for the best match by motif score to the guide region only within 50 nucleotides of the peak summit; the alignment was extended for 10 nucleotides at the 3′ end and visualized using Weblogo[60].

To calculate the motif score threshold at which false discovery rate < 0.1 for each sample, 100-nucleotide sequences centred around peak summits were shuffled while preserving dinucleotide frequency. The best match by motif score to the guide+PAM (NGG for SpCas9, NNGRRT for SaCas9) in these shuffled sequences was then found. The score threshold for false discovery rate < 0.1 was defined as the score such that less than 10% of shuffled peaks had a motif score above that score threshold.

DNA cleavage assay culture and transfection

Human embryonic kidney (HEK) cell line 293FT (Life Technologies) was maintained in Dulbecco's modified Eagle's Medium (DMEM) supplemented with 10% FBS (HyClone), 2 mM GlutaMAX (Life Technologies), 100 U/ml penicillin, and 100 μg/ml streptomycin at 37 °C with 5% CO2 incubation.

293FT cells were seeded onto 6-well plates, 24-well plates or 96-well plates (Corning) 24 h before transfection. Cells were transfected using Lipofectamine 2000 (Life Technologies) at 80–90% confluency following the manufacturer's recommended protocol. For each well of a 6-well plate, a total of 1 μg of Cas9+sgRNA plasmid was used. For each well of a 24-well plate, a total of 500 ng Cas9+sgRNA plasmid was used

unless otherwise indicated. For each well of a 96-well plate, 65 ng of Cas9 plasmid was used at a 1:1 molar ratio to the U6-sgRNA PCR product.

Deep sequencing to assess targeting specificity

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

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

Sequencing data analysis and indel detection

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

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

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

In order to place error bounds on the true-indel read frequencies in the sequencing libraries themselves, Wilson score intervals[48] were calculated for each sample, given the MLE-estimate for true-indel target-regions, Rp, and the number of reads R.

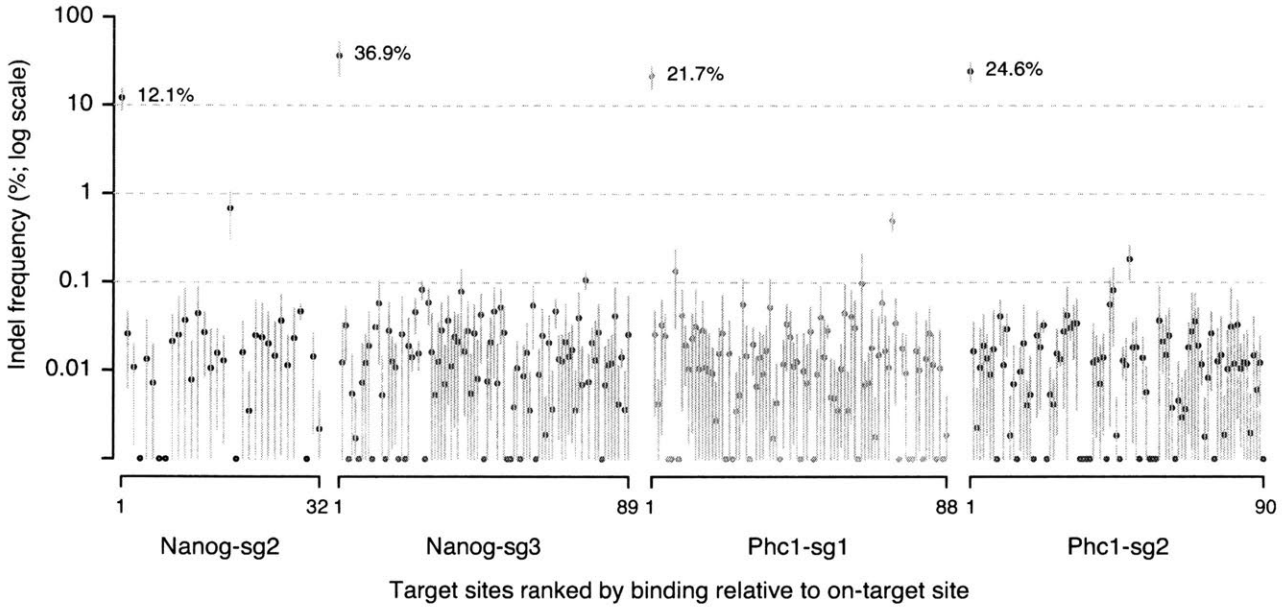


Figure 1: Indel frequencies at on-target sites and 295 off-target sites. For each sgRNA, selected target sites were ranked by decreasing ChIP binding relative to on-target site. Dots and gray bars indicate the mean and s.d. of indel frequency from three biological replicates, respectively. The y-axis was truncated at 0.001% for visualization at log scale. The indel frequencies for the four on-target sites are labeled with percentages.

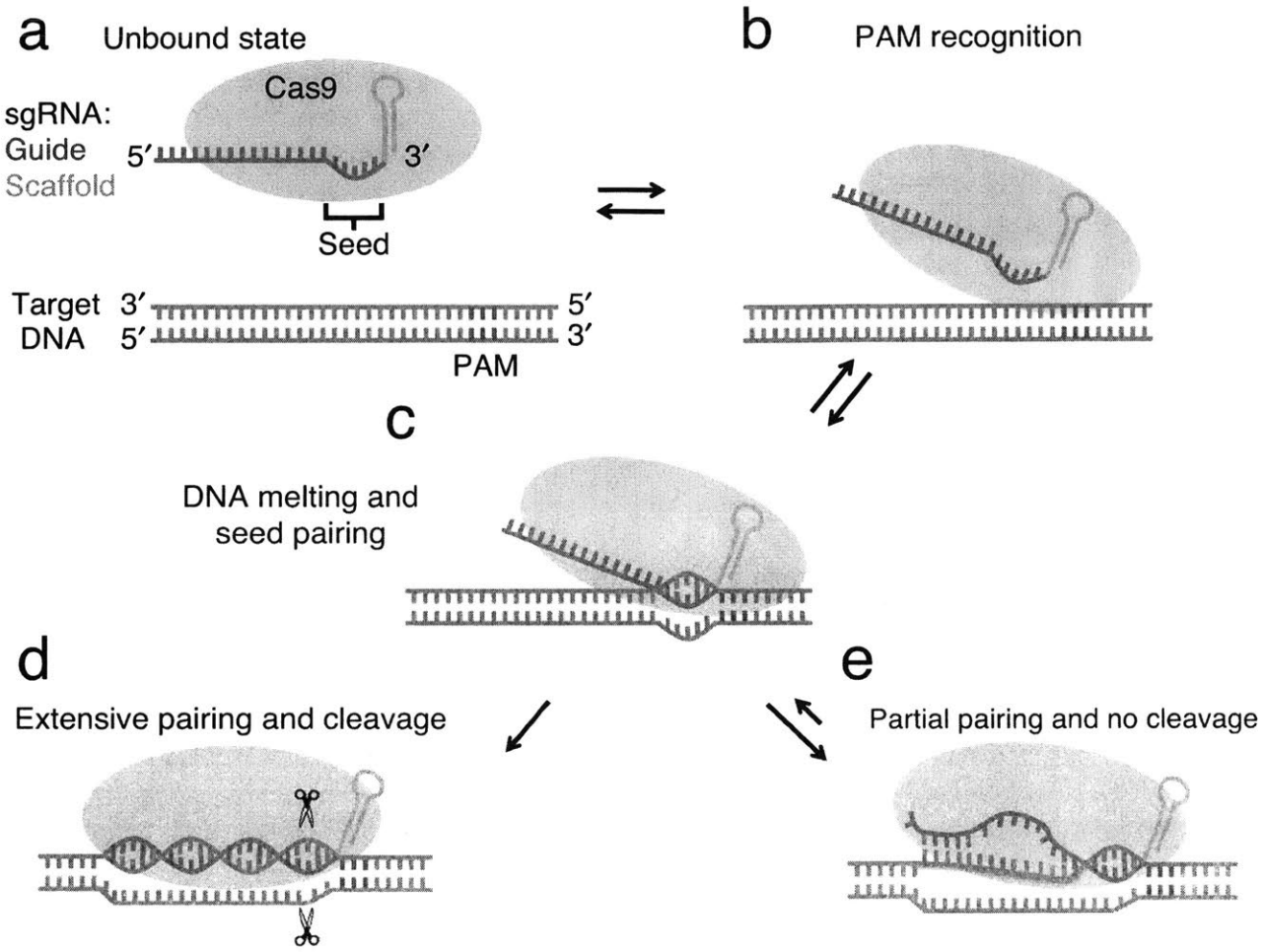


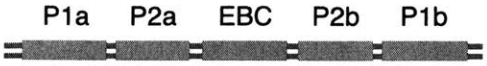
Figure 2: A model for Cas9 target binding and cleavage. (a) In the unbound state, Cas9 is loaded with sgRNA but not bound to DNA. The PAM region in the DNA is colored in red. (b) Recognition of the PAM by Cas9. (c) Cas9 melts the DNA target near the PAM to allow seed pairing. (d) If base pairing can be propagated to PAM-distal regions, the two Cas9 nuclease domains may be able to 'clamp' the target DNA and cleave it. (e) If only partial pairing occurs, there is no cleavage and Cas9 remains bound to the target.

a) insert elements

dsDNA insert ex.1.1
priming sites

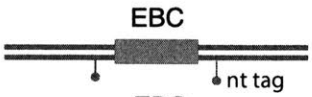
* Optional inclusion of modifi
nucleic acids to enhance insert
stability or insertion efficiency

dsDNA insert ex.1.2
event barcode



dsDNA insert ex.1.3
partial double strand

dsDNA insert ex.1.4
tagged nucleotides

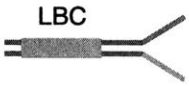


dsDNA insert ex.1.5
type IIS restiction

b) ligation handle configurati

ssDNA ligation handle

Y ligation adapter



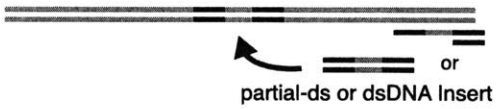
c) genomic event capt

i) genomic disr ion

ii) double str reak

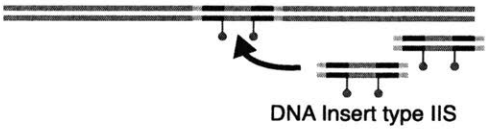
iii) event c

event capture ex.1



event capture ex.2

event capture ex.3



a) optionally: genomic insert event propagat

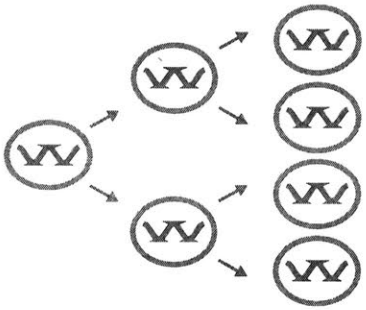


Figure 3: DNA Insertion Method (insert and library preparation elements). (a) ex.1.1: DNA insert composed of one or more priming sites; ex.1.2: DNA insert composed of an event barcode (EBC) and priming sites for bi-directional non-restrictive linear amplification (P1a, P1b) and exponential amplification (P2a, P2b); ex.1.3: single stranded DNA insert containing a partial double stranded region; ex.1.4 DNA insert containing tagged nucleotides or defined nucleotide sequence for affinity purification; ex.1.5 DNA insert containing typeIIS nuclease sites at insert ends. b) ssDNA ligation handle containing a ligation barcode (LBC) and priming site (P3) for exponential amplification; Y ligation adapter composed of LBS and handle sequences. (c) Events (indel, translocation, large insertion or deletion) are captured by the ligation of a DNA insert at a pair of free DNA ends. (d) Optionally,

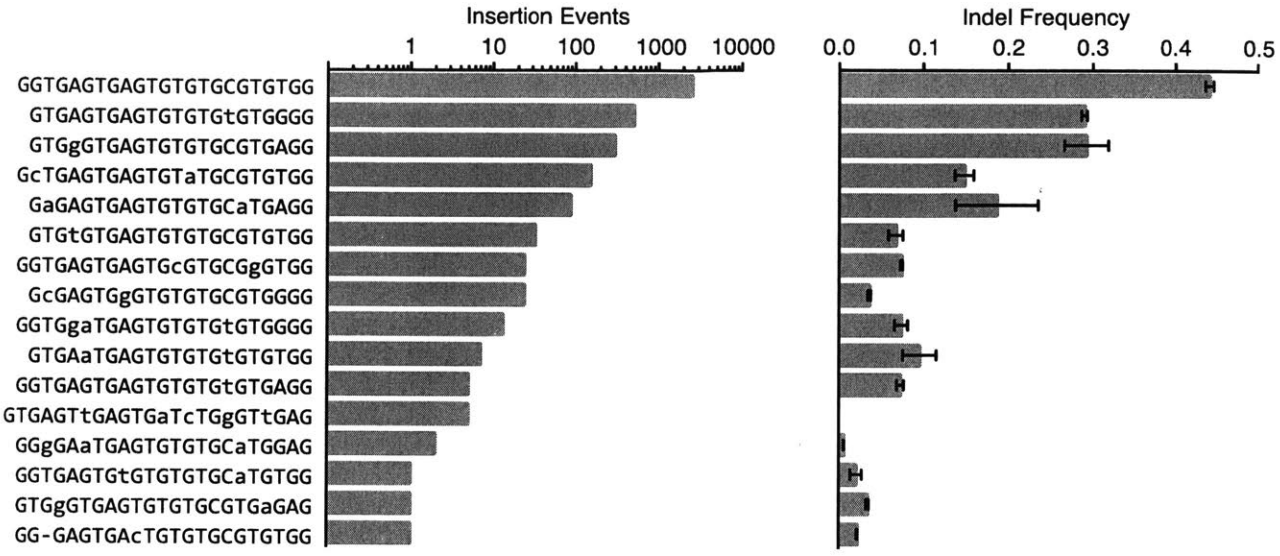


Figure 4: On and off-target insert enrichment for VEGFA3 editing. Left) enrichment plot of normalized insert event counts for on-target and validated off-target sites Right) Maximum likelihood estimate (MLE) indel frequencies for on-target and validated off-target sites.

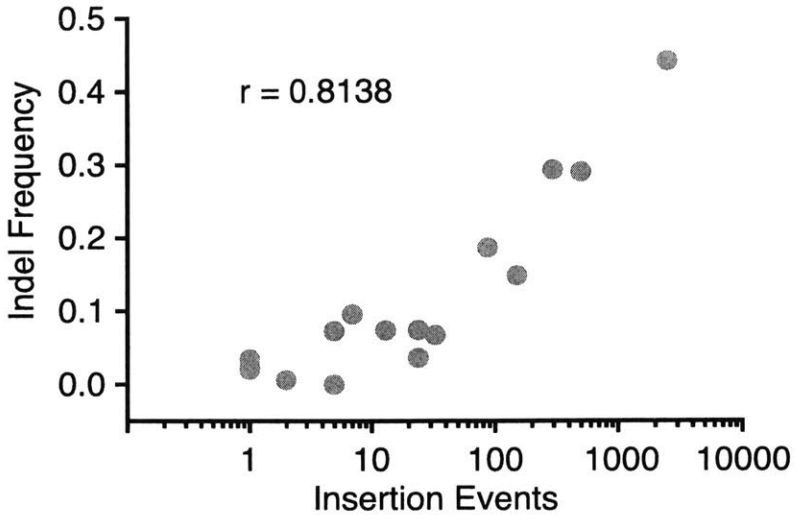


Figure 5: Correlation between insert enrichment and indel frequencies for VEGFA3 on and off-target sites. Scatter plot shows MLE indel frequencies vs. insertion events for VEGFA3 on target and validated off-target sites. Pearson product-moment correlation coefficient (r) between indel and insert frequencies is 0.8138.

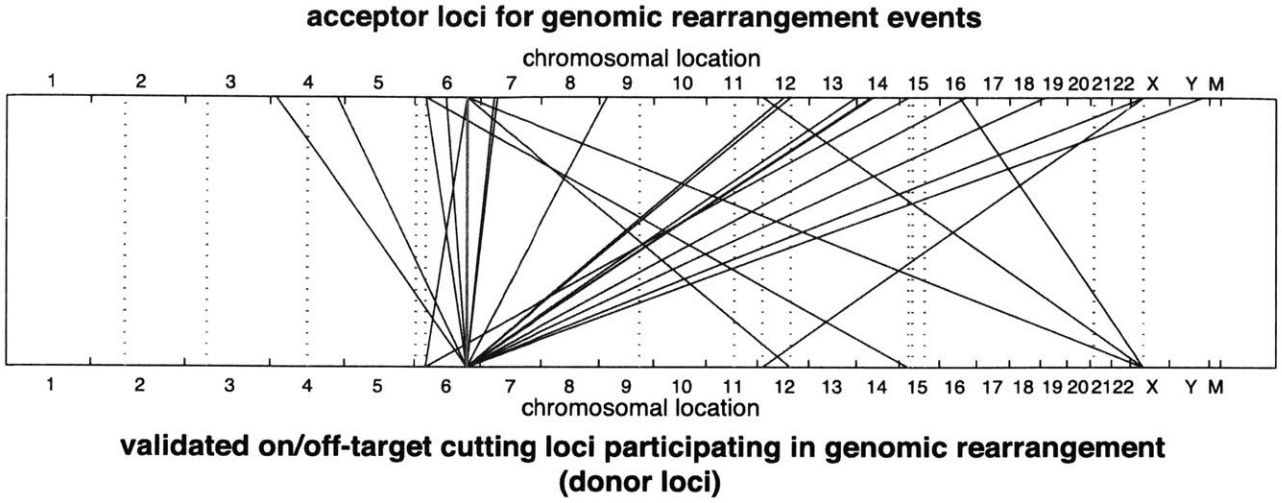


Figure 6: VEGFA3 Genomic Rearrangement Mappings. Detected mappings of genomic rearrangement events (long deletions, translocations) from validated VEGFA3 on-target and off-target loci (donor loci; lower nodes) to junction loci (acceptor loci; upper nodes).

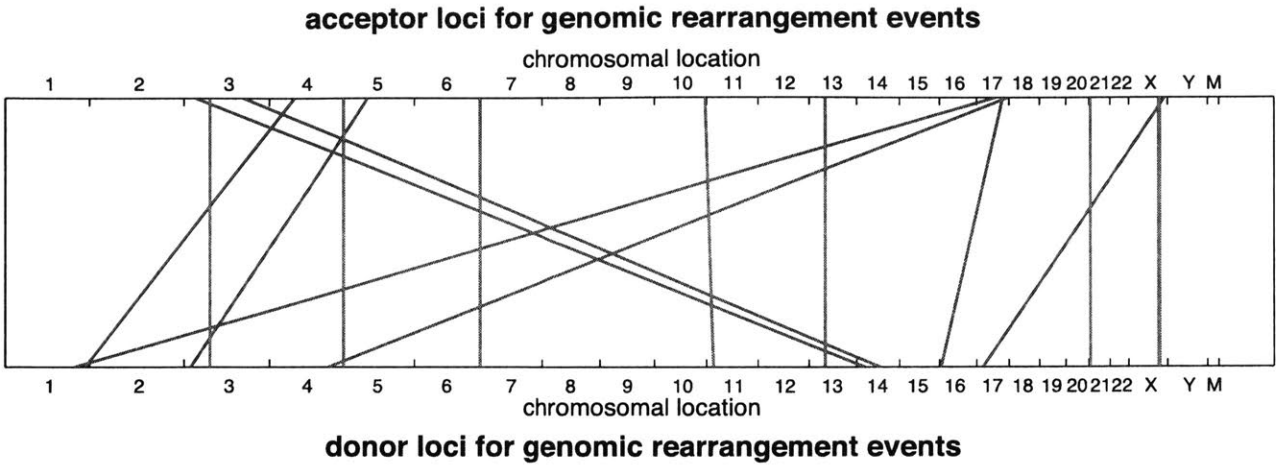


Figure 7: Baseline Genomic Rearrangement Mappings. Detected mappings of genomic rearrangement events (long deletions, translocations) from donor loci (lower nodes) to acceptor loci (upper nodes) in cells containing only insert.

CHAPTER 3: Unbiased off-target detection using direct DNA break labeling

Adapted from:

Ran, F. A.*, Cong, L.*, Yan, W. X.*, Scott, D. A., et al. In vivo genome editing using Staphylococcus aureus Cas9. *Nature* **520,** 186–191 (2015).

Yan, W. X.*, Mirzazadeh, R.*, Garnerone, S., Scott, D. A., et al. BLISS: versatile and quantitative DNA break sequencing in low-input cell and tissue samples. *Nat. Biotechnol., Under Review.*

To measure the genome-wide cleavage activity of SpCas9 and SaCas9 directly, we applied BLESS (direct in situ breaks labeling, enrichment on streptavidin and next-generation sequencing)[61] to directly label Cas9-induced DNA double-stranded breaks (DSBs) and quantify them using next-generation sequencing (NGS). Breaks Labeling In Situ and Sequencing (BLISS) was developed to further enhance the sensitivity and throughput of direct DSB labeling. Directly labeling double strand breaks in adherent cell monolayers, BLISS advances the capabilities of unbiased DSB detection towards characterization of genome editing *in vivo.*

3.1 Introduction

As advances in CRISPR-Cas technology promise to enable a broad range of *inVivo* and therapeutic applications, accurate, genome-wide identification of off-target nuclease activity has become increasingly important. Although a number of studies have employed sequence similarity-based off-target search[49,52,62,63,50,64] or dCas9-ChIP[58,65] to predict off-target sites for Cas9, such approaches cannot assess the nuclease activity of Cas9 in a comprehensive and unbiased manner. Additionally, GUIDEseq, IDLV integration, and HTGTS methods require low frequency errors in endogenous non-homologous end-joining (NHEJ) to label DSBs, thus limiting the absolute sensitivity of these assays and potentially missing DSBs that are not repaired via NHEJ. GUIDEseq and IDLV may also be limited to use in model cell lines due to limitations in the delivery of high concentration exogenous DNA fragments *inVivo.* Furthermore, HTGTS may be limited in sensitivity and biased due to the requirement for low frequency translocation events between the Cas9 on- and off-target sites. BLESS (direct in situ breaks labelling, enrichment on streptavidin and next-generation sequencing), and BLISS (Breaks Labeling In Situ and Sequencing) are methods to directly label DSBs that are not restricted by delivery of exogenous DNA fragments and do not require NHEJ events. We show that BLESS and BLISS provide sensitive detection of DSBs *in Situ* and enable visualization of the anatomy of endogenous targeted and spurious DSBs.

3.2.1 Results: Unbiased detection of Cas9 off-target activity using BLESS

To measure the genome-wide cleavage activity of SaCas9 and SpCas9 directly, we applied BLESS (direct in situ breaks labelling, enrichment on streptavidin and next-generation sequencing)[61] to capture a snapshot of Cas9-induced DNA double-stranded breaks (DSBs) in cells. We transfected 293FT cells with SaCas9 or SpCas9 and the same EMX1 targeting guides used in the previous ChIP experiment, or pUC19 as a negative control. After cells are fixed, free genomic DNA ends from DSBs are captured using biotinylated adaptors and analyzed by deep sequencing (Fig. 1a). To identify candidate Cas9-induced DSB sites genome-wide, we established a three-step analysis pipeline following alignment of the sequenced BLESS reads to the genome (Fig. S2a). First, we applied nearest-neighbor clustering on the aligned reads to identify groups of DSBs (DSB clusters) across the genome. Second, we sought to separate potential Cas9-induced DSB clusters from background DSB clusters resulting from low frequency biological processes and technical artefacts, as well as high-frequency telomeric and centromeric DSB hotspots[61]. From the on-target and a small subset of verified off-target sites (predicted by sequence similarity using a previously established method[49] and sequenced to detect indels), we found that reads in Cas9-induced DSB clusters mapped to characteristic, well-defined genomic positions compared to the more diffuse alignment pattern at background DSB clusters. To distinguish between the two types of DSB clusters, we calculated in each cluster the distance between all possible pairs of forward and reverse-oriented reads (corresponding to 3′ and 5′ ends of DSBs), and filtered out the background DSB clusters based on the distinctive pairwise-distance

distribution of these clusters (Fig. S1b, c). Third, the DSB score for a given locus was calculated by comparing the count of DSBs in the experimental and negative control samples using a maximum-likelihood estimate[49]. This analysis identified the on-target loci for both SaCas9 and SpCas9 guides as the top scoring sites, and revealed additional sites with high DSB scores (Fig. 1b–d).

Next, we sought to assess whether DSB scores correlated with indel formation. We used targeted deep sequencing to detect indel formation on the ~30 top-ranking off-target sites identified by BLESS for each Cas9 and sgRNA combination. We found that only those sites that contained a PAM and homology to the guide sequence exhibited indels (Fig. S2). We observed a strong linear correlation between DSB scores and indel levels for each Cas9 and sgRNA pairing (r2 = 0.948 and 0.989 for the two EMX1 targets with SaCas9 and r2 = 0.941 and 0.753 for those with SpCas9) (Fig. 1c, Fig. S3b–d). Furthermore, BLESS identified additional off-target sites not previously predicted by *in silico* methods or ChIP (Fig. S1, S3). These new off-target sites include not only those containing Watson–Crick base-pairing mismatches to the guide, but also the recently reported insertion and deletion mismatches in the guide:target heteroduplex (Fig. 1d)[50,64]. Together, these results highlight the need for more precise understanding of rules governing Cas9 nuclease activity, a requisite step towards improving the predictive power of computational guide design programs.

The potential for off-target mutagenesis is an important consideration for Cas9-mediated genome editing. Currently, studies on Cas9 specificity have not been able to directly measure genomewide nuclease activity in an unbiased manner[66,67,25,27,68,28]. Since BLESS (direct in situ breaks labeling, enrichment on streptavidin and next-

generation sequencing)[69] can capture DNA double-stranded breaks (DSBs) in mammalian cells, we sought to apply it towards evaluating Cas9 specificity.

BLESS captures chromosomal DSB sites via ligation of a biotinylated hairpin adaptor to open ends of genomic DNA (DSB-proximal labeling). Chromosomal DNA with DSBs can subsequently be captured using streptavidin beads and sheared to yield smaller size fragments compatible with next generation sequencing. A distal adaptor carrying the PCR priming site was then ligated on to permit enrichment prior to preparation of sequencing library (Fig. 1a).

Because DSBs can occur during natural biological processes such as replication, especially around pericentromeric and telomeric regions, as well as sample processing steps due to physical shearing, it is important to accurately identify the DSBs induced by Cas9. To do this, we empirically optimized the parameters for each step of our BLESS analysis, as explained in the following subsections: 1. Determining the clustering window for building regions of DSB enrichment ("DSB clusters"), 2. Defining the distribution of pairwise-distances within each DSB cluster, and 3. Background subtraction using negative controls (Fig. S2a).

The AAV-SaCas9 system is able to mediate efficient and rapid editing of Pcsk9 in the mouse liver, resulting in reductions of serum Pcsk9 and total cholesterol levels. To assess the specificity of SaCas9, we used an unbiased DSB detection method, BLESS, to identify a list of candidate off-target cleavage sites in a mouse cell line. We examined these sites in liver tissue transduced by AAV-SaCas9 and did not observe any indel formation within the detection limits of in vitro BLESS and targeted deep sequencing.

Importantly, the off-target sites identified in vitro might differ from those in vivo, which need to be further evaluated by the applications of BLESS or other unbiased techniques such as those published during the revision of this work[57,70].

3.2.2 Results: BLESS computational analysis

I) Determining the clustering window for building regions of DSB enrichment

To build the DSB clusters from the sequencing reads, we took the first 30-bp of sequence reads immediately following the proximal label and mapped them using Bowtie to the hg19 or mm9 reference genome, allowing up to 2 mismatches. Following alignment, the reads were grouped using a nearest-neighbor clustering method, hence referred to as "DSB clusters": we determined the genomic coordinate of the 5'-most position (first base) for each read, and grouped reads by applying a sliding window of width x, i.e. within each DSB cluster, the first base of any given read will be no more than x bp from its adjacent reads. We empirically determined that 30-bp windows yielded well-defined DSB clusters.

II) Defining the distribution of pairwise-distances within each DSB cluster

The grouping in the previous section using a sliding window identified all DSB regions, but did not distinguish between the ones induced by Cas9 activity and those from background. To determine properties that could be used to separate the two, we compiled a training data set by extracting the reads mapped to the on-target and a subset of known off-target sites. These offtarget sites, verified by the presence of indels from sequencing, include those predicted based on similarity to the on-target sequence as well as by dCas9-ChIP. We additionally included centromeric regions with DSB

signals observed in both experimental (Cas9 and sgRNA co-transfected) and negative control (pUC19 transfected) samples to further refine the specificity of the algorithm.

Since every DSB generates two open chromosomal ends, the sequencing reads from either end of the DSB align to the + and – strands of the reference genome. The pattern and distribution of the forward (+ strand aligned) and reverse (– strand aligned) reads in a given DSB cluster can help determine whether it is induced by Cas9. Since the DSB site within a centromeric or telomeric region is not consistent from cell to cell, we expect that such a DSB cluster contains forward and reverse reads that are broadly distributed (Fig. S1b)[61]. This contrasts with DSBs induced by Cas9, which typically occur at a well-defined position 3-bp upstream of the PAM[25,28,69] and result in a characteristic distribution of forward and reverse reads that flank a sharp break site (Fig. S1c). However, due to end-resection during DNA repair, there can be reads aligned away from the cleavage site.

Cas9-induced DSBs can be distinguished from background events by the following analysis: first, we calculated the distance between every possible pair of forward and reverse reads in the DSB cluster by subtracting the chromosomal coordinate of the first base on reverse read from that of the forward read. A distance of 1 thus represents the reverse and forward reads directly abutting and facing away from each other. Distances of >1bp indicate reads that are separated by one or more base pairs, and distances of <1bp indicate reads that overlap. Second, we generated a histogram of these distances for each DSB cluster. Histograms of clusters from centromeric, telomeric, and other background regions had broad distributions of pairwise distances (Fig. S1b), while the histograms from Cas9-induced DSB clusters were sharp and asymmetric (Fig. S1c).

Finally, to quantify this difference we calculated the fraction of pairwise reads that overlapped by no more than 7bp (distance ≥ -6bp) within all possible pairwise distances in each cluster. Based on the training dataset, we found that in the majority of Cas9-induced clusters, this fraction was greater than 0.99. In using this metric to filter out background clusters, we required that a candidate Cas9-induced DSB cluster should have a minimal fraction number of 0.95. This relaxed cut-off value of 0.95 was selected to increase the sensitivity for detecting bona fide Cas9-induced clusters, particularly for those with fewer read counts where a small number of outlier reads might significantly reduce the fraction value.

III) Background subtraction using negative controls

Finally, we compared the DSB clusters in the experimental versus the negative control group to locate and remove background signals that should be present in both datasets. The DSB score for a given genomic locus was calculated by comparing the count of pass-filter clusters in the experimental samples with the controls using a maximum likelihood estimate (MLE)[49]. This score describes the number of expected Cas9-induced DSB clusters per 1x10^5 reads and allowed the final ranking of all candidate DSB sites. We have taken the above approach to minimize the use of heuristics and limit the introduction of potential biases during the identification of Cas9-induced DSBs. To assess how effectively the candidate DSB sites from BLESS predict levels of indel formation, we performed targeted deepsequencing on the top ~30 loci that have the

highest DSB scores (Fig. 1, Fig. S2). This revealed a significant linear correlation between BLESS DSB scores and rates of indel mutations (Fig. 1c).

Altogether, these results suggest that BLESS is a sensitive method for detecting Cas9 nuclease activity in an unbiased, genome-wide manner. In particular, BLESS was able to detect off-target sites featuring insertion and deletion mismatches between the RNA guide and the genomic DNA sequence. Furthermore, BLESS is a powerful tool for future studies of designer nuclease specificity in several ways: first, it is a direct measurement of nuclease activity, unaffected by the efficiency of downstream events. This additionally reduces bias resulting from off-target modification of essential genes, which can affect cell viability. Second, BLESS has the potential to be readily applied in vivo without the need to label cleavage events through exogenous markers.

Future implementations of BLESS can be improved by including the use of unique molecular identifiers (UMIs) on the hairpin adaptors to minimize the effects of PCR bias. Additionally, greater sequencing depth can improve the sensitivity for identifying candidate Cas9-induced DSB. Furthermore, taking multiple time points after delivery of Cas9 can shed additional insight on temporal dynamics of its genome-wide activity.

3.2.3 Results: Unbiased characterization of genome editing *inVivo* using BLISS

In collaboration with the lab of Nicola Crosetto, we aimed to develop a more generalizable and quantitative method to directly label CRISPR-Cas induced DSBs. Here, we describe such a method for Breaks Labeling In Situ and Sequencing (BLISS) that builds upon three main features: 1) direct DSB labeling in fixed cells immobilized on a solid surface; 2) T7-mediated in vitro transcription (IVT)[71] to increase the sensitivity of DSB detection in low-quantity samples and control PCR biases by linearly amplifying each labeled DSB event; 3) incorporation of UMIs at break sites to enable exact counting of identical DSB events occurring in different cells. A scheme of BLISS is shown in Fig. 2a. The goal of BLISS is two-fold, 1) to provide sensitive and versatile DSB detection in a wide range of substrates, including low-input specimens, and 2) to scale direct DSB labeling for high-throughput analysis of diverse samples.

We applied BLISS to evaluate the genome editing specificity of Cas9 (Cas9-BLISS) in a 24-well cell culture plate format to assess the sensitivity of BLISS to detect genome-wide Cas9-induced off-target DSBs. We transfected 293T cells with Cas9 as well as one of two single guide RNAs (sgRNAs) targeting either EMX1 or VEGFA, both of which have well-characterized off-target profiles[72,57,73,70]. In situ blunting and ligation steps were performed directly on the plate (Fig. 3a), greatly reducing the amount of input material required and turnaround time per sample in comparison to BLESS (10^5 vs. 10^7 cells/sample; ~12 hands-on man-hours over 5 days vs. at least ~60 hands-on man-hours over 15 days of preparation time for 24 samples). Using an optimized

computational analysis pipeline incorporating unique molecular identifiers for more accurate quantitation of DSB frequency, we were able to clearly identify Cas9-induced DSBs and uncover several off-targets previously identified by GUIDEseq[57] but not BLESS. Additionally, BLISS identifies several low-abundance off targets that had not been previously identified by either method (Fig. 3b-c). We then used rarefaction analysis to assess the complexity of our libraries and estimate the sensitivity of Cas9-induced DSB detection. Labeling of Cas9-induced breaks with UMIs across transfected cells allowed us to accurately quantify the number of unique DSBs observed at Cas9 target locations at multiple read depths ranging from 1 to 18 million reads per sample (Fig. 3d). We fit the data to extrapolate the sequencing depth at which the total population of UMIs is saturated at the on-target locus. Based on the approximate number of cells from which each library was obtained, we estimated that ~0.39–0.78% and ~0.26–0.51% of cells had unrepaired on-target breaks in EMX1 or VEGFA, respectively. Repeating the same analysis for the lowest detectable Cas9 off-targets at the lowest read count of 1 million reads (Fig. 3d), we estimated that BLISS is able to detect off-target DSBs present in as few as ~0.06–0.12% and ~0.08–0.15% of cells. These results demonstrate that BLISS is a sensitive method for the detection of genome-wide DSBs induced by Cas9, even when ~40 times fewer cells are used as input compared to BLESS. Furthermore, the rapid turnaround and scalable nature of BLISS make this method valuable for assessing the specificity of multiple sgRNAs simultaneously.

3.3 Discussion

Compared to other methods for genome-wide unbiased DSB detection, including BLESS[61], BLISS presents a number of key advantages: 1) the assay format is simple and versatile. The ability to carry out all enzymatic reactions and washing steps on a solid surface allows for low-input cell specimens and potentially for tissue sections to be easily and quickly processed with easy buffer exchanges while minimizing sample loss; 2) incorporation of UMIs and selective amplification of labeled DSBs improves the scope, accuracy and sensitivity of BLISS to detect and quantify DSBs; 3) turnaround time is greatly reduced and multiple samples can be easily barcoded and pooled, enabling assay scalability and reducing library preparation costs. Altogether, we believe that BLISS is a powerful method for genome-wide DSBs mapping that opens up the possibility to study DSBs in many more biologically relevant conditions and samples, including precious clinical specimens.

While direct double strand break labeling has a theoretical sensitivity limit of single DSB detection, there are multiple factors that limit the sensitivity of these assays. 1) In contrast to GuideSeq, IDLV insertion, and HTGTS, which successively label NHEJ events over the timecourse of editing and integrate signal over time, BLESS and BLISS only label DSBs with free DNA ends existing at the instant of cell fixation. Hence, BLESS and BLISS capture a snapshot of the DSB landscape at only a single point in time, which limits their sensitivity since they do not integrate the landscape of the cutting events over time. 2) BLESS is also severely limited by the labeling of DSBs introduced mechanically during multiple preceding steps of sample processing prior to break labeling. This problem is partially corrected with BLISS, but both methods also contend

with background from large numbers of DSBs naturally existing in rapidly dividing cells used in our current experiments. It is possible that background DSB levels will be reduced for primary cells or somatic tissues. Although it is desirable to study the specificity of genome editing in endogenous cells and tissues that are the target of editing, significant challenges exist to the sensitivity of BLESS and BLISS for *inSitu* specificity analysis.

3.4 Methods

BLESS for DSB detection

Cells are harvested at 24 h post-transfection, then processed as previously described[61] with the following alterations: a total of 10 million cells are fixed for nuclei isolation and permeabilization, and treated with Proteinase K for 4 min at 37 °C before inactivation with PMSF. All deproteinized nuclei are used for DSB labelling with 100 mM of annealed proximal linkers overnight. After Proteinase K digestion of labelled nuclei, chromatin was mechanically sheared with a 26G needle before sonication (BioRuptor, 20 min on high, 50% duty cycle). 20 µg of sheared chromatin are captured on streptavidin beads, washed, and ligated to 200 mM of distal linker. Linker hairpins are then cleaved off with I-SceI digestion for 1 h at 37 °C, and products PCR-enriched for 18 cycles before proceeding to library preparation with TruSeq Nano LT Kit (Illumina). For the negative control, cells mock transfected with Lipofectamine 2000 and pUC19 DNA were parallel processed through the assay.

BLESS analysis

Fastq files were demultiplexed, and 30-bp genomic sequences were separated from the BLESS ligation handles for alignment. Bowtie was used to map the genomic sequences to hg19 or mm9, allowing for a maximum of 2 mismatches. Following alignment, reads from all bio-replicates for an individual sample were first pooled, and then nearest

neighbour clustering was performed with a 30-bp moving window to identify regions of enrichment across the genome. Within each cluster, the pairwise distance was calculated between all forward and reverse read strand mappings (Extended Data Fig. 7b, c). Pairwise distance distributions were used to filter out wide and poorly defined DSB clusters from the well-defined DSB clusters characteristically found at Cas9-induced cleavage sites (see Supplementary Information). Finally, we adjusted the count of predicted Cas9-induced DSBs at a given locus by using a binomial model to calculate the maximum-likelihood estimate of peak enrichment in the Cas9-sgRNA treated sgRNAs given BLESS measurements from an untreated negative control. After the maximum-likelihood estimate calculation, a list of loci ranked by their DSB scores could be obtained and plotted (Fig. 3b, Extended Data Fig. 8). Additional descriptions can be found in Supplementary Information.

The top-ranking ~30 sites from the list of Cas9 induced DSB clusters were sequenced for indel formation (Extended Data Fig. 8; validated targets in Fig. 3d). Within these loci, PAMs and regions of target homology were identified by first searching all PAM sites within a ± 50 bp window around the DSB cluster, then selecting the adjacent sequence with fewest mismatches to the target sequence.

BLISS transfection for DSB detection

The selected targets for Cas9-BLISS are located within the EMX1 locus (GAGTCCGAGCAGAAGAAGAA gGG) and the VEGFA gene locus (GGTGAGTGAGTGTGTGCGTG tGG). The plasmids used containing the SpCas9 and

the sgRNA cassette were identical to the ones used for Cas9-BLESS[11], where the targets were labeled as EMX1(1) and VEGFA(1). The same targets have also been studied using GUIDEseq[12], where they were labeled as EMX1 and VEGFA_site3. AsCpf1 and LbCpf1 along with their cognate crRNAs were cloned into the same expression vector as Cas9 to enable a direct comparison. Cells were plated before transfection in 24-well plates pre-coated with poly-D-lysine (Merck Millipore, cat. no. A003E) at a density of ~125,000/well, and were let grow for 16-18h until 60–70% confluence. For transfections, we used 2µl of Lipofectamine 2000 (Life Technologies, cat. no. 11668019) and 500ng of Cas9 plasmid in 100µl total of OptiMEM (Gibco, cat. no. 31985062) per each well of a 24-well plate as described previously (Ran et al., 2015).

BLISS adapters

All BLISS adapters were prepared by annealing two complementary oligonucleotides as described below. All oligos were purchased from Integrated DNA Technologies as standard desalted oligos. UMIs were generated by random incorporation of the four standard dNTPs using the "Machine mixing" option. Before annealing, sense oligos diluted at 10µM in nuclease-free water were phosphorylated for 1h at 37°C with 0.2U/µl of T4 Polynucleotide Kinase (NEB, cat. no. M0201). Phosphorylated sense oligos were annealed with the corresponding antisense oligos pre-diluted at 10µM in nuclease-free water, by incubating them for 5min at 95°C, followed by gradual cooling down to 25°C over a period of 45min (1.55°C/min) in a PCR thermo-cycler.

BLISS

A step-by-step BLISS protocol is provided in **Supplementary Information**. For BLISS in cell lines, we typically either grew cells directly onto 13mm coverslips (VWR, cat. no. 631-0148) or we spotted them onto coverslips pre-coated with poly-L-lysine (PLL) (Sigma, cat. no. P8920-100ML). For Cas9 & Cpf1 experiments, we fixed HEK293T cells directly into the 24-well plate used for transfections, and performed all *in situ* reactions done directly inside the plate. For BLISS in mouse liver, we developed two approaches: 1) Tissue cryopreservation and sectioning. Freshly extracted liver biopsies were first fixed in PFA 4% for 1h at rt, and then immersed in a sucrose solution (15% overnight and then 30% until the tissue sank) before embedding in OCT. 30µm-thick tissue sections were mounted onto microscope slides, dried for 60min at rt, and stored at 4°C before further processing. 2) Preparation of nuclei suspensions. Freshly extracted liver biopsies were cut into small pieces and transferred into a 1.5-2ml tube containing nucleus isolation buffer (NaCl 146mM, Tris-HCl 10mM, $CaCl_2$ 1mM, $MgCl_2$ 21mM, Bovine Serum Albumin 0.05%, Nonidet P-40 0.2%, pH 7.8). We typically incubated the samples for 15-40min until the tissue fragments became transparent, after which the nuclei were centrifuged for 5min at 500g and then re-suspended in 200-500µl of 1× PBS. 100µl of nuclei suspension were dispensed onto a 13mm PLL-coated coverslip and incubated for 10min at rt. Afterwards, 100µl of PFA 8% in 1× PBS were gently added and incubated for 10min at rt, followed by two washes in 1× PBS at rt. The samples were stored in 1× PBS at 4°C up to one month before performing BLISS.

Identification of SpCas9, AsCpf1, LbCpf1 on- and off-target DSBs

We updated the original DSB detection pipeline previously described for analyzing Cas9-BLESS data[1,2] to determine whether we could enhance the sensitivity of off-target detection by both BLESS and BLISS. Previously, we demonstrated that a homology search algorithm was capable of separating *bona fide* Cas9 induced DSBs from background DSBs, and performed the analysis on the top 200 DSB loci with the strongest signal after initial filtering[2]. To achieve even greater sensitivity, here we extended this homology search to the top 5,000 DSB locations identified by BLISS. To enable a direct comparison between BLESS and BLISS, we used this updated approach to re-analyze the BLESS data previously obtained with wild-type SpCas9[2] on the same EMX and VEGFA guide targets as studied here. Briefly, a 'Guide Homology Score' was determined using an algorithm that searched for the best matched guide sequence within a region of the genome 50nt on either side of the center of a DSB cluster identified in BLESS/BLISS for all NGG and NAG PAM sequences in the case of SpCas9[3], and all possible PAMs in the case of AsCpf1 and LbCpf1 for maximum sensitivity. A score based on the homology was calculated using the Pairwise2 module in the Biopython Python package with the following weights: a match between the sgRNA and the genomic sequence scores +3, a mismatch is –1, while an insertion or deletion between the sgRNA and genomic sequence costs –5. Thereby, an on-target sequence with the fully matched 20bp guide would have a Guide Homology Score of 60. Previously, we included the PAM match in the scoring, yielding a maximum score of 69, but in order to make the score more versatile and comparable across different

PAMs, we removed the PAM dependence in the scoring. Using this guide homology score, we performed a receiver operating characteristic (ROC) curve analysis based on validated and non-validated off-targets from SpCas9-BLESS[74] which justified our previous choice of a homology score cutoff (41 out of a max score of 60), to maximize the sensitivity and specificity of Cas9-BLISS and Cpf1-BLISS. In practical terms, this score corresponds to ≤4 mismatches or ≤2 gaps, as well as combinations thereof.

Code availability

BLESS analysis code is available at https://github.com/fengzhanglab/BLESS.

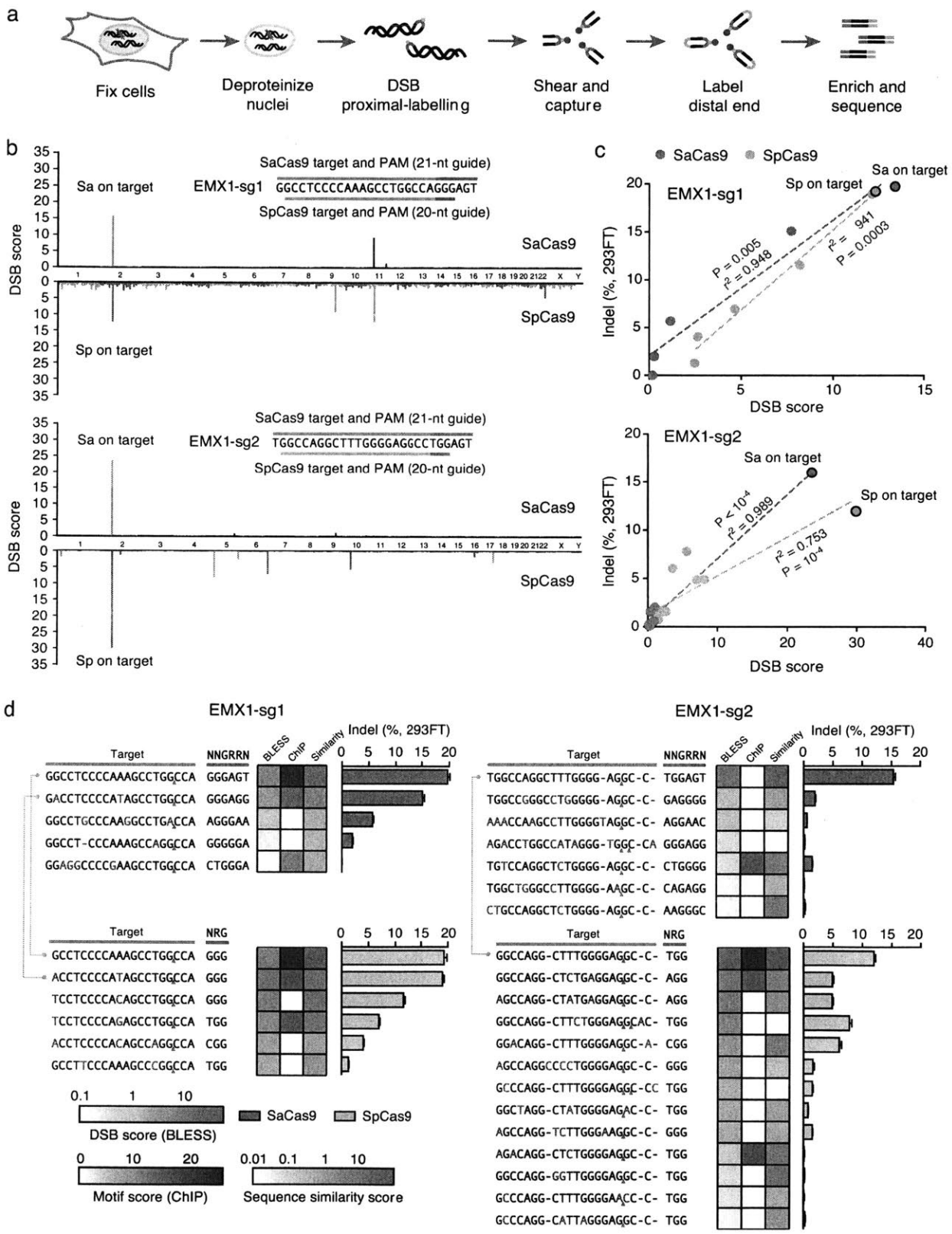

a
Fix cells
Deproteinize nuclei
DSB proximal-labelling
Shear and capture
Label distal end
Enrich and sequence
b
SaCas9 target and PAM (21-nt guide)
Sa on target
EMX1-sg1 GGCCTCCCCAAAGCCTGGCCAGGGAGT
SpCas9 target and PAM (20-nt guide)
SaCas9
SpCas9
Sp on target
DSB score
EMX1-sg2 TGGCCAGGCTTTGGGGAGGCCTGGAGT
c
SaCas9
SpCas9
EMX1-sg1
Sa on target
Sp on target
P = 0.005
r² = 0.948
r² = 941
P = 0.0003
Indel (%, 293FT)
DSB score
EMX1-sg2
P < 10⁻⁴
r² = 0.989
r² = 0.753
P = 10⁻⁴
d
EMX1-sg1
EMX1-sg2
Target
NNGRRN
NRG
BLESS
ChIP
Similarity
Indel (%, 293FT)
GGCCTCCCCAAAGCCTGGCCA GGGAGT
GACCTCCCCATAGCCTGGCCA GGGAGG
GGCCTGCCCAAGGCCTGACCA AGGGAA
GGCCT-CCCAAAGCCAGGCCA GGGGGA
GGAGGCCCCGAAGCCTGGCCA CTGGGA
GCCTCCCCAAAGCCTGGCCA GGG
ACCTCCCCATAGCCTGGCCA GGG
TCCTCCCCACAGCCTGGCCA GGG
TCCTCCCCAGAGCCTGGCCA TGG
ACCTCCCCACAGCCAGGCCA CGG
GCCTTCCCAAAGCCCGGCCA TGG
TGGCCAGGCTTTGGGG-AGGC-C- TGGAGT
TGGCCGGGCCTGGGGG-AGGC-C- GAGGGG
AAACCAAGCCTTGGGGTAGGC-C- AGGAAC
AGACCTGGCCATAGGG-TGGC-CA GGGAGG
TGTCCAGGCTCTGGGG-AGGC-C- CTGGGG
TGGCTGGGCCTTGGGG-AAGC-C- CAGAGG
CTGCCAGGCTCTGGGG-AGGC-C- AAGGGC
GGCCAGG-CTTTGGGGAGGC-C- TGG
GGCCAGG-CTCTGAGGAGGC-C- AGG
AGCCAGG-CTATGAGGAGGC-C- AGG
GGCCAGG-CTTCTGGGAGGCAC- TGG
GGACAGG-CTTTGGGGAGGC-A- CGG
AGCCAGGCCCCTGGGGAGGC-C- GGG
GCCCAGG-CTTTGGGGAGGC-CC TGG
GGCTAGG-CTATGGGGAGAC-C- TGG
AGCCAGG-TCTTGGGAAGGC-C- GGG
AGACAGG-CTCTGGGGAGGC-C- TGG
GGCCAGG-GGTTGGGGAGGC-C- TGG
GCCCAGG-CTTTGGGGAACC-C- TGG
GCCCAGG-CATTAGGGAGGC-C- TGG
0.1 1 10
DSB score (BLESS)
SaCas9
SpCas9
0 10 20
Motif score (ChIP)
0.01 0.1 1 10
Sequence similarity score

Figure 1: Characterization of genome-wide nuclease activity of SaCas9 and SpCas9. a, Schematic of BLESS processing steps. b, Manhattan plots of genome-wide DSB clusters generated by each Cas9 and sgRNA pair, with on-target loci shown above (see Supplementary Discussion). c, Correlation between DSB scores and indel levels for top-scoring DSB clusters. Trend lines, r2 and P values are calculated using ordinary least squares method. d, Off-target loci from BLESS with detectable indels through targeted deep sequencing (n = 3) are shown. Heat maps indicate DSB score (blue), motif score from ChIP (purple), or sequence similarity score (green) for each locus. Blue triangles indicate peak positions of BLESS signal.

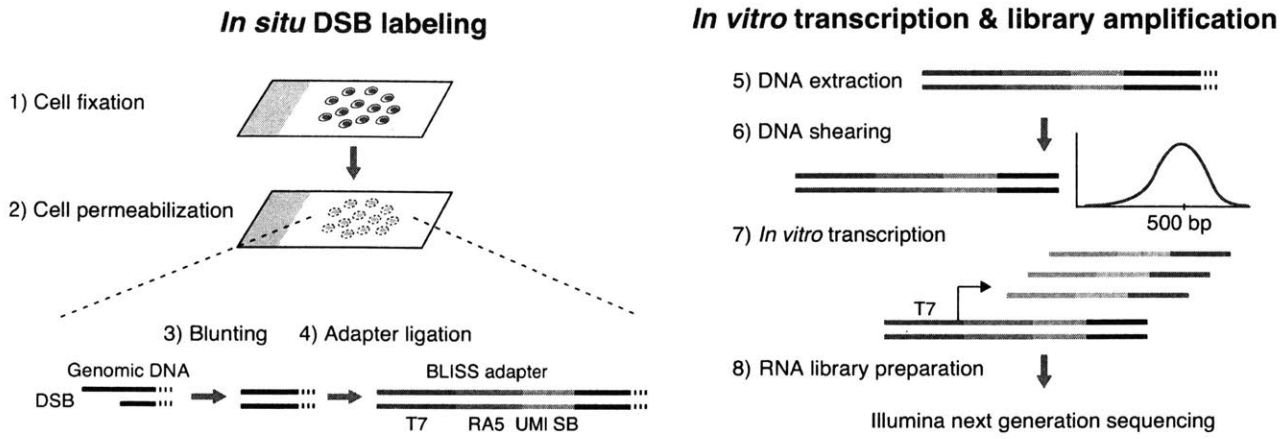


Figure 2: BLISS implementation and validation. (a) Method workflow. Fixed cells are attached onto a coated glass surface (1) and permeabilized to expose DSBs (2). DSBs are then in situ blunted (3) and ligated (4) using double-stranded oligonucleotide adapters containing a sample barcode (SB), a unique molecular identifier (UMI), the Illumina RA5 sequencing adapter, and the T7 promoter sequence. After ligation, cells are detached from the glass and DNA is extracted (5), sonicated (6), and in vitro transcribed (6). Finally, a sequencing library is prepared using a modified Illumina TruSeq Small RNA kit and sequenced (8).

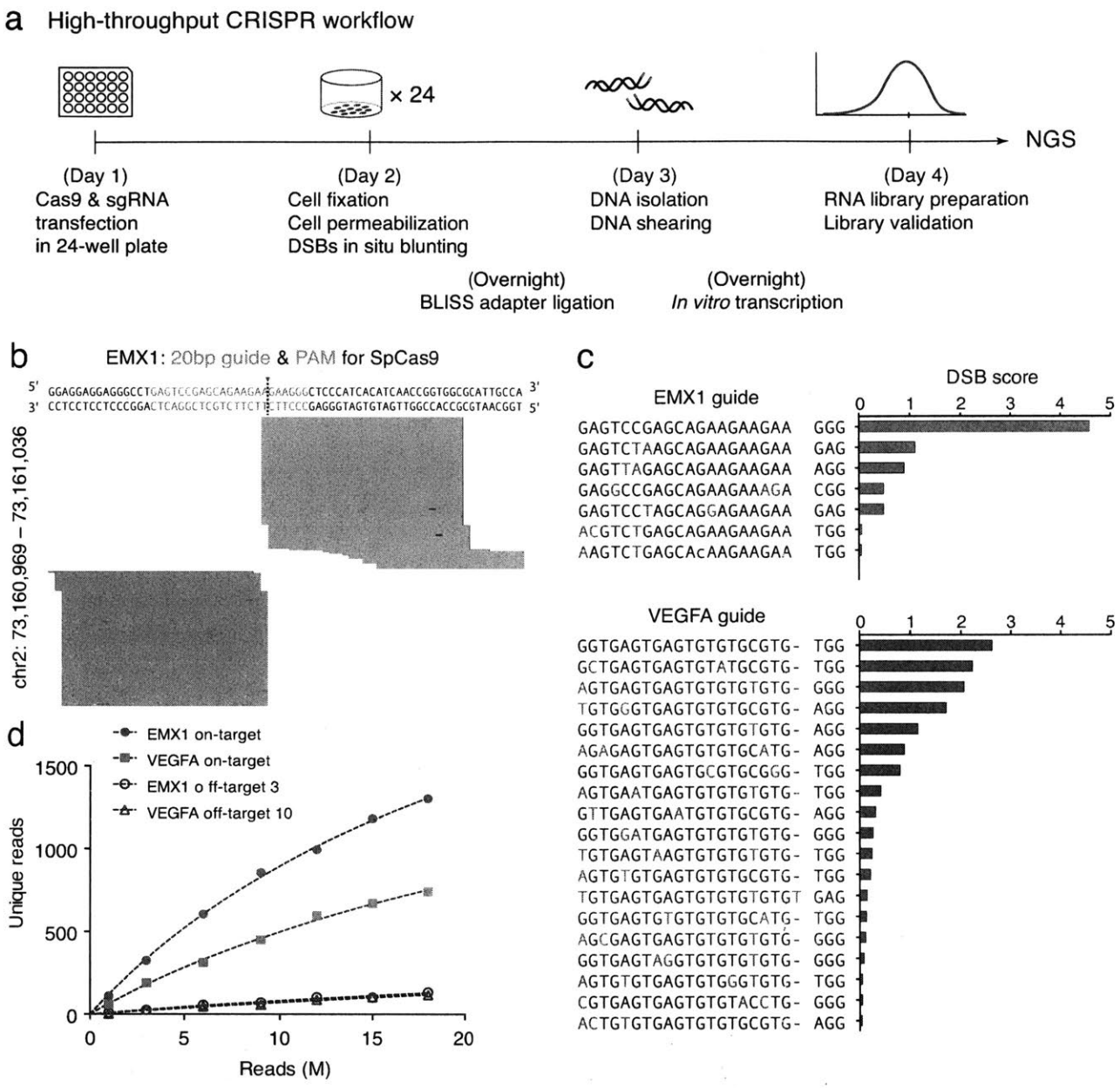
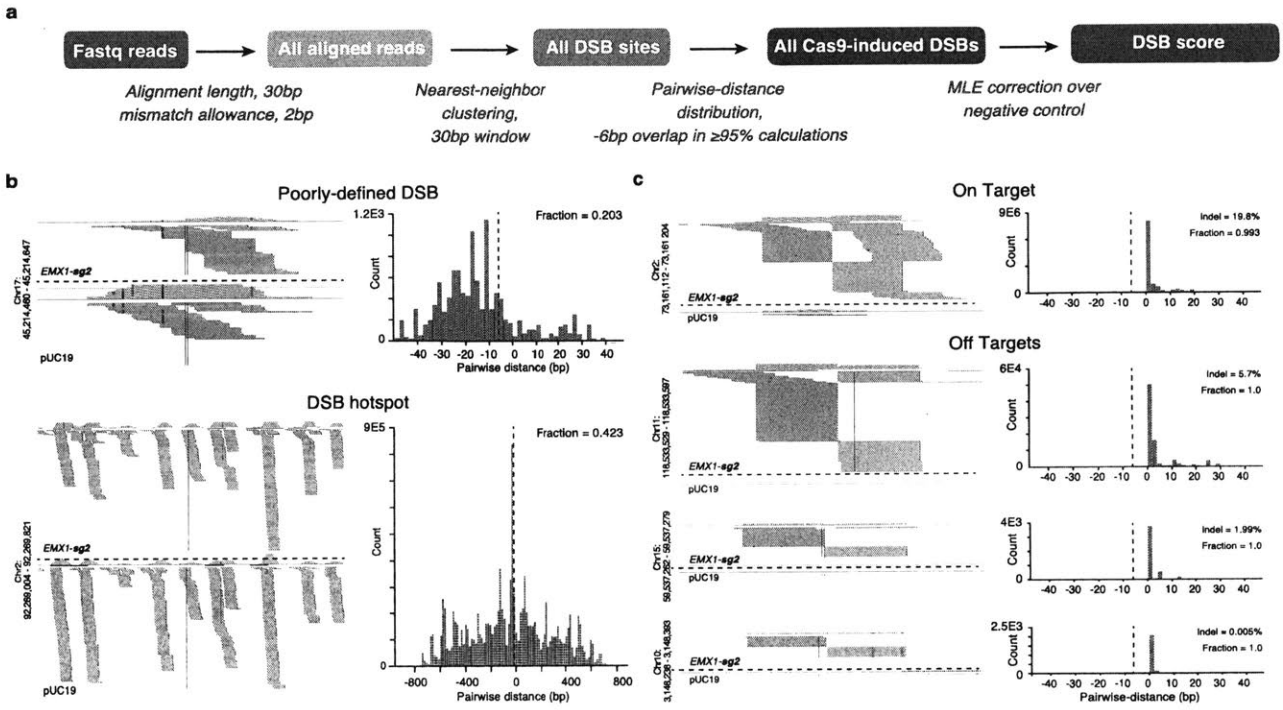
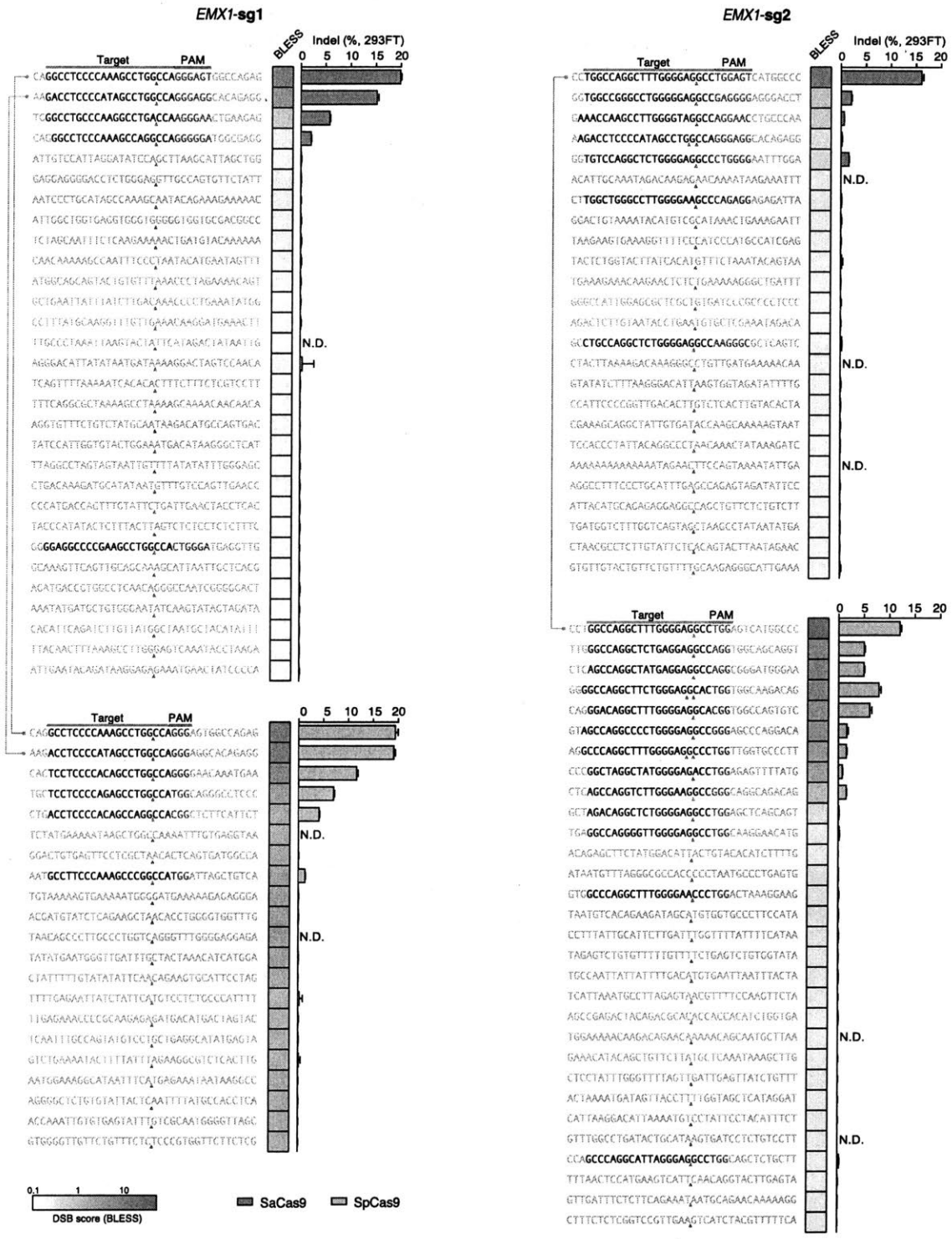


Figure 3: High-throughput Cas9-BLISS for unbiased, genome-wide off-target detection. (a) Overview of the timeline from 293T cell transfection to BLISS library sequencing. The use of poly-D-lysine coated plates allows fixation, permeabilization, and ligation directly in the plate, allowing multiplexing of guides and samples. (b) Genome browser view of BLISS reads mapped to the EMX1 target. Blue, reads mapped to the minus strand. Red, reads mapped to the plus strand. The Cas9-induced DSB (dotted line) occurs 3-4 base pairs upstream of the PAM sequence. (c) On- and off-target loci of guides targeting EMX1 and VEGFA genes identified by Cas9-BLISS. Bases highlighted in red represent mismatches from the on-target sequence. '–' represents an insertion or deletion between the sgRNA and the genomic sequence. (d)

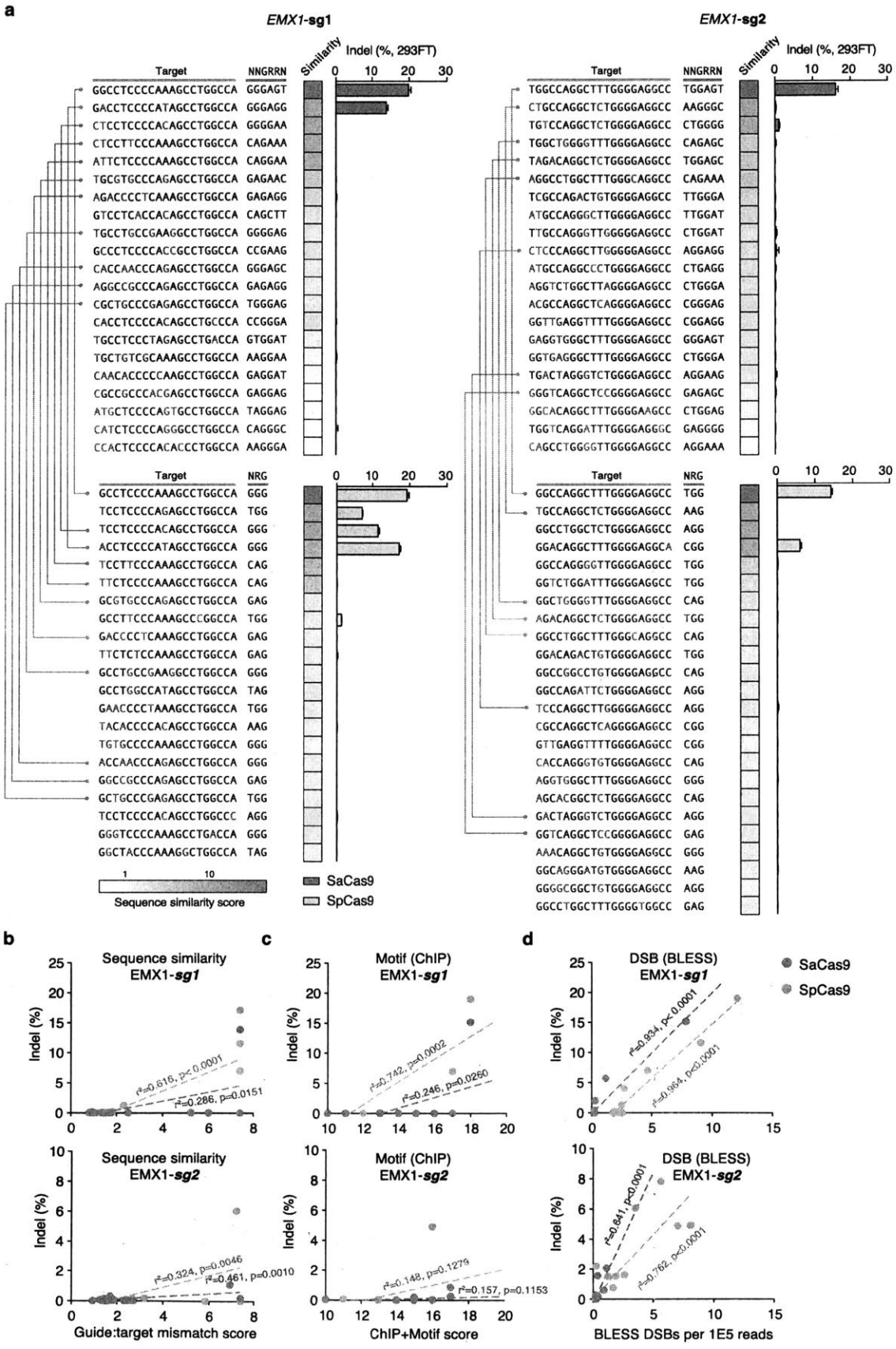
Rarefaction analysis of the number of unique reads (i.e. distinguishable DSBs) mapped to the on-target and the lowest abundance off-target site at 1M total reads for the EMX1 and VEGFA guides as a function of sequencing depth. Dashed lines, regression curves (see Supplementary Information).



Supplementary Figure 1: Analysis pipeline of sequencing data from BLESS. a, Overview of the data analysis pipeline starting from the raw sequencing reads. Representative sequencing read mappings and corresponding histograms of the pairwise distances between all the forward orientation (red) reads and reverse orientation (blue) reads, displayed for representative b, DSB hotspots and poorly defined DSB sites and c, Cas9-induced DSBs with detectable indels. Fraction of pairwise distances between reads overlapping by no more than 6 bp (dashed vertical line) are indicated over histogram plots.



Supplementary Figure 2: Indel measurements at off-target sites based on DSB scores. List of top off-target sites ranked by DSB scores for each Cas9 and sgRNA pair. Indel levels are determined by targeted deep sequencing. Blue triangles indicate positions of peak BLESS signal, and where present, PAMs and targets with sequence homology to the guide are highlighted. Lines connect the common on-targets (EMX1) and off-targets between the two Cas9 enzymes. N.D., not determined.


a
EMX1-sg1
EMX1-sg2
Target
NNGRRN
NRG
Similarity
Indel (%, 293FT)
Sequence similarity score
SaCas9
SpCas9
b
Sequence similarity
EMX1-sg1
EMX1-sg2
Indel (%)
r²=0.616, p<0.0001
r²=0.286, p=0.0151
r²=0.324, p=0.0046
r²=0.461, p=0.0010
Guide:target mismatch score
c
Motif (ChIP)
r²=0.742, p=0.0002
r²=0.246, p=0.0260
r²=0.148, p=0.1279
r²=0.157, p=0.1153
ChIP+Motif score
d
DSB (BLESS)
r²=0.934, p<0.0001
r²=0.964, p<0.0001
r²=0.841, p<0.0001
r²=0.762, p<0.0001
BLESS DSBs per 1E5 reads

Supplementary Figure 3: Indel measurements of top candidate off-target sites based on sequence similarity score. Off-targets are predicted based on sequence similarity to on-target, accounting for number and position of Watson–Crick base-pairing mismatches. NNGRR and NRG are used as potential PAMs for SaCas9 and SpCas9, respectively. Lines connect the common targets (EMX1) and off-targets between the two Cas9 enzymes. Correlation plots between indel percentages and b, prediction based on sequence similarity, c, ChIP peaks ranked by motif similarity, or d, DSB scores for top ranking off-target loci. Trendlines, r^2, and P values are calculated using ordinary least squares.

CHAPTER 4: Implications of Human Genetic Variation for Therapeutic Genome Editing

Adapted from:

Scott, D. A. and Zhang F. Implications of Human Genetic Variation for CRISPR-Based Therapeutic Genome Editing. *Nat. Medicine, Under Review.*

CRISPR-Cas genome editing methods hold immense potential as therapeutic tools, promising to fix disease-causing mutations at the level of DNA. The successful development of such therapies, however, must take into consideration how naturally occurring variation in the human population will affect the targeting specificity of Cas endonucleases. We present here an analysis of the recently released ExAC and 1000 Genomes datasets to investigate how human genetic variation impacts therapeutic genome editing. We find that certain Cas endonucleases are more favorable than others, regions of low variation can be predicted using currently available sequencing datasets, and that in large populations, most off-target candidates for a given RNA guide will be rare and exist in small numbers of patients. We integrate this information to develop a framework (including a compendium of high efficacy RNA guides) to help guide the design of CRISPR-based therapeutics to maximize efficacy and safety across patient populations.

4.1 Introduction

The development of CRISPR-Cas9 RNA-guided endonucleases for eukaryotic genome editing has sparked intense interest in the use of this technology for therapeutic applications[23,24,26]. In contrast to small molecule therapies, which target highly conserved protein active sites, treatment of disease at the genomic level must contend with significant levels of genetic variation in patient populations. Recently, large scale sequencing datasets from the Exome Aggregation Consortium (ExAC) and 1000 Genomes Project have provided an unprecedented view of the landscape of human genetic variation[75,76,77,78]. This variation can affect both the efficacy of a CRISPR-based therapeutic, by disrupting the target site, and its safety, by generating off-target candidate sites. Here we determine the impact of population genetic variation on therapeutic genome editing with *Streptococcus pyogenes* (Sp) Cas9, SpCas9 variants VQR and VRER, *Staphylococcus aureus* (Sa) Cas9, and *Acidaminococcus sp.* (As) Cpf1[23,24,74,79,26]. We find extensive variation likely to substantially alter the efficacy of these enzymes, and we show that unique, patient-specific off-target candidates will be the greatest challenge to safety. These results provide a framework for designing CRISPR-based therapeutics, highlight the need to develop multiple guide RNA-enzyme pairs for each target locus, and suggest that pre-therapeutic whole genome sequencing will be required to ensure uniform efficacy and safety for treatment across patient populations.

4.2.1 Results: Implications of target variation for therapeutic efficacy

Human genetic variation impacts choice of Cas enzyme

To date, two families of Class 2 (single effector) CRISPR nucleases, Cas9 and Cpf1, have been harnessed for eukaryotic genome editing[23,24,80,26]. Both Cas9 and Cpf1 are programmed by RNA guides, which mediate cleavage of DNA targets that are complementary to the guide RNA protospacer sequence and flanked by a short protospacer adjacent motif (PAM) specific to each endonuclease[25,26] (Fig. 1a). Mismatches between the RNA guide and its DNA target have been shown to decrease RNA-guided endonuclease activity, and deviation from the canonical PAM sequence often completely ablates nuclease activity[49,63,52,44]. Currently RNA guides are designed using the reference human genome; however, failing to take into account variation in the human population may confound the therapeutic outcome for a given RNA guide. The recently released ExAC dataset, based on 60,706 individuals, contains on average one variant per eight nucleotides in the human exome[75]. This highlights the potential for genetic variation to impact the efficacy of certain RNA guides across patient populations for CRISPR-based gene therapy[75], due to the presence of mismatches between the RNA guide and variants present in the target site of specific patients. To assess this impact, we use the ExAC dataset to catalog variants present in all possible targets in the human reference exome that either (i) disrupt the target PAM sequence or (ii) introduce mismatches between the RNA guide and the genomic DNA, which we collectively term target variation (Fig. 1a). For treatment of a patient population, avoiding

target variation for RNA guides administered to individual patients will maximize the consistency of outcomes for a genome editing therapeutic.

A number of RNA-guided CRISPR nucleases have now been discovered and engineered as tools for genome editing, each with a different PAM[23,24,74,79,26] (Table 1). For therapeutic design, consideration of multiple enzymes with different PAM requirements is advantageous as it increases the number of available genomic targets for therapeutic loci. We therefore assessed variation at each PAM in the human exome for SpCas9 (PAM = NGG), SpCas9-VQR (NGA), SpCas9-VRER (NGCG), SaCas9 (NNGRRT), and AsCpf1 (TTTN), all of which are currently being considered as candidate enzymes for CRISPR therapeutic development (the recently reported eSpCas9 and SpCas9-HF have the same NGG PAM as SpCas9, and are thus not considered separately here[72,81]). For each nuclease, we determined the fraction of exonic PAMs containing variants that alter PAM recognition. For the ExAC population, the total fraction of targets containing PAM-altering variants was similar for all enzymes (21 – 35%), except for SpCas9-VRER, which is impacted by PAM-altering variants in 80% of targets (Table 1, Fig. S1). The PAM for SpCas9-VRER contains a CpG motif, which has been shown to be highly mutable[75]. Consistent with these results, we find that CG is the most highly mutable 2-nt PAM motif in the human exome, and 66% of cytosine and guanine residues contained in CpG motifs show variation for the 60,706 ExAC individuals[75747372717069686766656463626160595857565756555453] (Fig. 1b, c; Table S3). These results suggest that enzymes using PAMs containing CG motifs are significantly more affected by target variation in the human genome.

Targeting low-variation regions of the human exome enhances therapeutic efficacy

Considering full target variation for all ExAC individuals, we find that 93 – 95% of targets in the human exome for SpCas9, SpCas9-VQR, SaCas9, and AsCpf1 contain variants likely to alter enzymatic activity (Fig. 1d, e; Table S1). Most (88%) of the target variation captured in the ExAC dataset is heterozygous, highlighting the fact that much of this target variation occurs at low frequencies in the population (Fig. 1d,f; Table S1). The ExAC dataset is large enough that it provides near comprehensive coverage of variants in the protein coding genome occurring at allele frequencies of greater than or equal to 0.01% in the population (1 out of 10,000 alleles)[75]. Hence, we used this dataset to compile a compendium of exome-wide target sites for SpCas9, Cas9-VQR, SaCas9, and AsCpf1 that do not contain variants occurring at ≥0.01% allele frequency (referred to as platinum targets; will be made available online) (Fig. 2a). These platinum targets are efficacious in >99.99% of the population (Fig. 2b). For further analysis, we focused on 12 therapeutically relevant genes, including those that are currently the focus of therapeutic development (See Fig. S2 for overview of genes included). For these genes, approximately two-thirds of possible protein coding targets meet our platinum criteria, with *PCSK9* containing the smallest fraction of targets (50%) meeting our platinum criteria (Table S2). While it is preferable to design RNA guides specifically for individual patients this may be challenging from a regulatory standpoint and cost prohibitive. Selecting from these platinum targets during therapeutic design will maximize efficacy across patient populations with the smallest number of RNA guides. When targeting

regions with more than one high frequency haplotypes, it will be necessary to design multiple RNA guides for each independent haplotype.

We find that high variation targets or platinum targets cluster along exons for each of the 12 genes examined. For example, all targets in the 5' half of *PCSK9* exon 4 are platinum, whereas very few platinum targets exist for exon 5 (Fig. 2c). Even for regions in *PCSK9* exons 1-4 with high frequencies of variation, it is still possible to find small numbers of platinum targets for some enzymes (Fig. 2c). This observation for *PCSK9* is representative of the other genes investigated in this study and suggests that considering multiple enzymes with distinct PAM requirements increases the likelihood of finding a platinum target.

4.2.2 Results: Implications of off-target variation for therapeutic safety

Low frequency off-target candidates for a given RNA guide predominate in large populations

When designing RNA guides, in addition to minimizing target variation, it is necessary to ensure safety by minimizing potential off-target activity due to sites in the genome similar to the target. Unbiased investigation of genome-wide CRISPR nuclease activity suggests that most off-target activity occurs at loci with at most three mismatches to the RNA guide[49,57,70,82,74,50,83,84]. Current approaches for Cas9 target selection rank off-target candidates found in the reference human genome by both the number and position of RNA guide mismatches, with the assumption that loci containing less than 3 mismatches or containing PAM distal mismatches are more likely to be cleaved[49,63,52]. However, in a population of individuals, this strategy is complicated by the existence of multiple haplotypes (sets of associated variants), which will contain different positions or numbers of mismatches at candidate off-target sites (Fig. 3a). We used phased single nucleotide variant calls to reconstruct allele-specific whole-genome sequences for each individual in the 1000 Genomes population [85]. For platinum targets in the 12 genes considered here, we quantified off-target candidates (defined as genomic loci with at most three mismatches to a given RNA guide) arising from all 1000 Genomes haplotypes. In this relatively small population (2504 individuals), more than half of the haplotypes containing off-target candidates are common (present in ≥10% of individuals) (Fig. 3b). However, in this population, the number of off-target candidates

for each RNA guide is inversely correlated with haplotype frequency (Fig. 3b). This trend indicates that for large populations the majority of off-target candidates for a given RNA guide will differ between individuals.

Avoiding high-frequency off-target candidates maximizes population safety

For individual RNA guides in these 12 genes, we find that the number of off-target candidates for SpCas9, SpCas9-VQR, SaCas9, and AsCpf1 varies from 0 to greater than 10,000 in the 1000 Genomes population (Fig. 3c). Much of this large variation in the number of off-targets reflects how unique or repetitive an individual target sequence is within the human genome. For instance SaCas9, which has a longer PAM and hence fewer genomic targets, has on average fewer off-target candidates per RNA guide (Fig. 3c). Additionally, in a population, the number of off-target candidates at a given locus is further compounded by multiple haplotypes, such that as the size of a population increases so does the number of haplotypes for an individual off-target locus. Hence, for each off-target candidate present in a high frequency haplotype, in a large population, multiple lower frequency haplotypes are likely to exist with reduced numbers of RNA guide mismatches. These data indicate that minimizing the number of off-target candidates occurring in high frequency haplotypes is of critical importance for the selection of therapeutic RNA guides. By minimizing these off-target candidates in high frequency haplotypes, off-target candidates occurring in low frequency haplotypes that uniquely impact individual or small numbers of patients will also be minimized The current 1000 genomes dataset provides comprehensive coverage of alleles occurring at

up to 0.1% in the population (considered to be the lower bound of high frequency variants), allowing us to identify platinum targets with minimal off-target candidates occurring in high frequency haplotypes in the human population[85,75].

Of the 12 genes we considered, some are more repetitive relative to the rest of the human genome, which impacts the specificity of the underlying RNA guides for each gene (Fig. 4a). For example, within *PCSK9* exons 2 – 5, we observed that platinum targets with high or low numbers of off-target candidates tend to cluster in regions of sequence that are either repetitive or unique within the genome, respectively (Fig. 4b) This pattern holds true for all 12 genes studied. Interestingly, within repetitive regions of exons, we identified platinum targets with significantly reduced quantities of off-target candidates. These findings further support the notion that utilizing multiple enzymes with distinct PAM requirements will enhance both safety and efficacy. Use of the enhanced specificity enzymes eSpCas9 and Cas9-HF1 will further reduce the likelihood of cleavage at off-target candidate sites, but it will still remain important to avoid repetitive therapeutic targets with large numbers of off-target candidates even with these enzymes[72,81].

Population demographics can be used to further improve therapeutic design.

Because the 1000 Genomes project provides demographic information for each individual, we used this data to explore how much off-target candidate variation for a given individual is explained by population demographics. For all off-target candidates for RNA guides targeting the 12 genes considered here, we performed principle

component analysis (PCA) and find that the first five principle components separate individuals very effectively by continent, sub-continent, and sex (Fig. 4c, Fig. S3 – 5). Cumulatively, population demographics account for 12% of the off-target candidates for a given individual, indicating that safety and efficacy of therapeutics can be enhanced by designing therapeutic targets for specific geographical or genotypic patient subpopulations.

4.3 Discussion

Ideally, personalized genomic medicine would tailor RNA-guided endonuclease therapeutics for each patient. However, it would likely be cost-prohibitive and infeasible from a regulatory standpoint to design an individual RNA guide for each patient receiving a genome editing therapy. Our analysis of the impact of genetic variation on the efficacy and safety of RNA-guided endonucleases motivates the following framework to streamline the design and testing of genome editing therapeutics (Fig. 4d). First, use of RNA guides for platinum targets would ensure perfect targeting for 99.99% of patients. Second, these RNA guides need to be further selected to minimize the number of off-target candidates occurring on high frequency haplotypes in the patient population. Third, low frequency variation captured in large scale sequencing datasets can be used to estimate the number of guide RNA-enzyme combinations required to effectively and safely treat different sizes of patient populations. Growth of large scale sequencing datasets will improve the accuracy of these estimates. Fourth, pre-therapeutic whole genome sequencing of individual patients will be needed to select a single approved guide RNA-enzyme combination for treatment. This combination should be a perfect match to the patient's genome and be free of patient-specific off-target candidates. This framework, in combination with rapidly accumulating human sequencing data, which will further refine these selection criteria, will enable the design and validation of genome editing therapeutics minimizing both the number of guide RNA-enzyme combinations necessary for approval and the cost of delivering effective and safe gene therapies to patients.

4.4 Methods

Human Variation Datasets

Our target variation analysis was performed using the Exome Aggregation Consortium (ExAC) dataset from 60,706 globally diverse individuals[75]. Our investigation of off-target candidates was performed using the 1000 Genomes Project phase 3 dataset containing phased whole genome sequences from 2504 globally diverse individuals[85].

Whole-exome target variation analysis

We included all targets for CRISPR enzymes SpCas9-WT, SpCas9-VQR, SpCas9-VRER, SaCas9, and AsCpf1 in the human exome that map to protein coding regions of exons with an average coverage of at least 20 reads per ExAC sample. For analysis of variation in these targets, we included all missense or synonymous variants passing quality filtering in the ExAC dataset as described previously[75]. Because the publicly available ExAC dataset includes only summary information for each variant, it was not possible to determine if multiple variants occurring in a single genomic target occur on different haplotypes. Hence, we calculated target variation frequency as the maximum frequency of variants in an individual target. While accurately approximating the variation of most targets in the population, this approach does underestimate the variation frequency for rare targets containing multiple high frequency variants existing

on separate haplotypes. Platinum targets were defined as those with a maximum variant frequency of less than 0.01% in the ExAC population.

Off-target candidate analysis

Phased haplotypes included in the 1000 Genomes phase 3 dataset were used to create whole genome allele-specific references for 2504 individuals. We included in our analysis all single nucleotide polymorphisms passing quality filtering in the 1000 Genomes phase 3 dataset as described previously[85]. Up to 100 protein-coding platinum targets for each therapeutically relevant gene, CEP290, CFTR, DMD, G6PC, HBB, IDUA, IL2RG, PCSK9, PDCD1, SERPINA1, TTR, VEGFA were selected for proteins SpCas9-WT, SpCas9-VQR, SaCas9, and AsCpf1. Targets for each gene were searched against the references for each of the 2504 1000 genomes individuals to profile off-target candidates specific to each individual. For the purpose of this study, off-target candidates are defined as unintended genome-wide targets for a specific guide RNA-enzyme combination with less than or equal to 3-mismatches with the guide RNA protospacer. We performed principle component analysis (PCA) taking into account all off-target candidates present in less than 100% of the 1000 Genomes individuals.

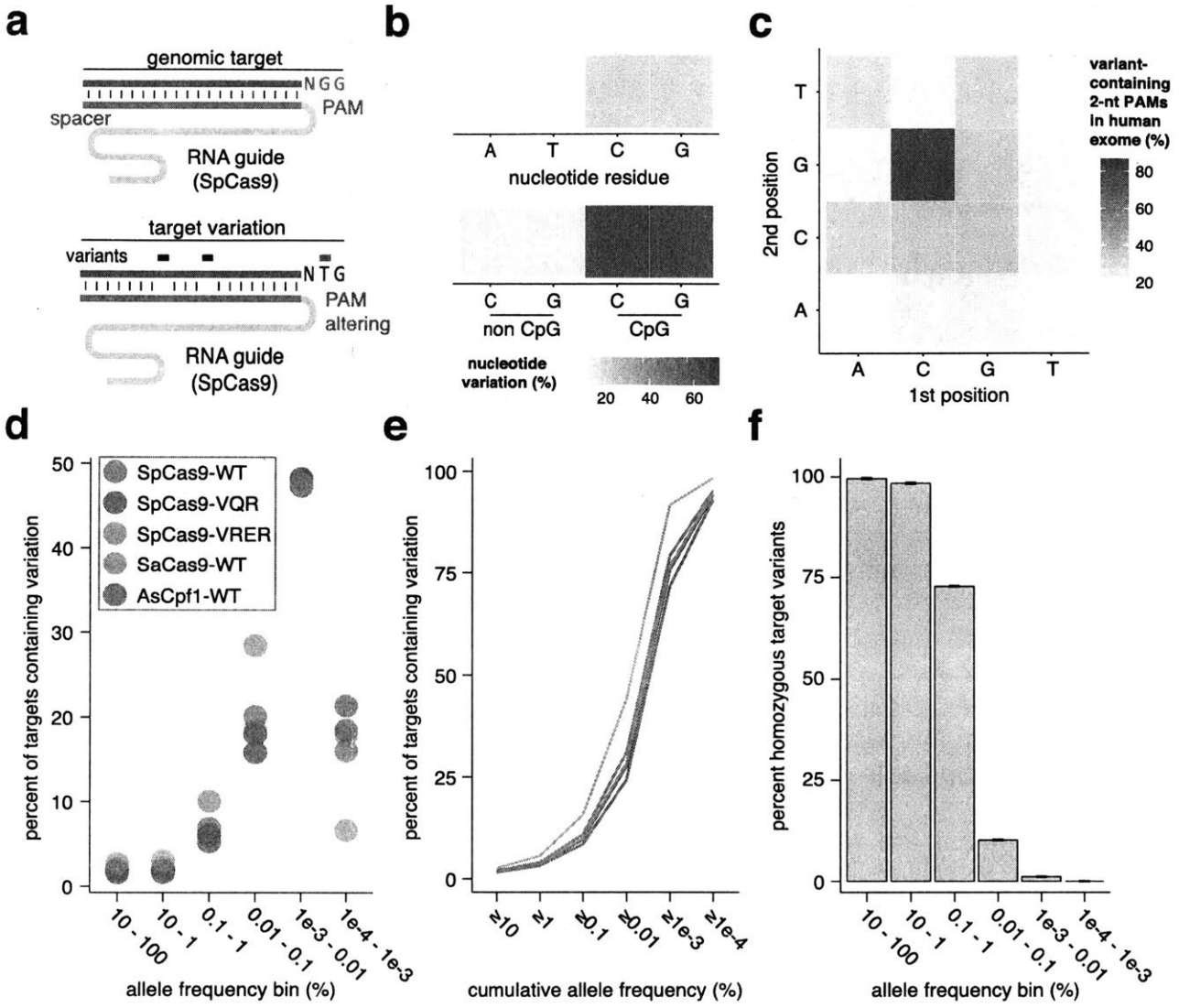


Figure 1: Human genetic variation significantly impacts the efficacy of RNA-guided endonucleases. a, Schematic illustrating the genomic target, RNA guide, and target variation. **b,** Fraction of residues for individual nucleotides containing variation in the ExAC dataset. **c,** Fraction of 2-nt PAM motifs altered by variants in the ExAC dataset. **d,** Percent of targets variants at different allele frequencies for each CRISPR endonuclease. **e,** Cumulative percent of targets containing variants for each enzyme. **f,** Fraction of targets containing homozygous variants at different allele frequencies. The mean and standard deviation for all enzymes is shown.

			Whole-exome PAM variation by allele frequency (%)						
protein	**PAM**	**orientation**	**≥10%**	**≥1**	**≥0.1**	**≥0.01**	**≥0.001**	**total**	**n**
AsCpf1-WT	TTTN	left	0.15	0.26	0.61	1.81	8.91	21.04	2702056
SpCas9-VQR	NGA	right	0.11	0.25	0.69	2.28	11.39	23.19	9838603
SpCas9-WT	NGG	right	0.16	0.37	1.13	3.82	17.46	32.61	10286445
SaCas9-WT	NNGRRT	right	0.23	0.44	1.16	3.68	17.29	34.68	1938911
SpCas9-VRER	NGCG	right	0.77	1.86	5.79	20.72	66.67	80.16	981524

Table 1: Fraction of targets containing PAM altering variants for each CRISPR endonuclease (n specifies the number of protein coding targets in the human exome for each enzyme).

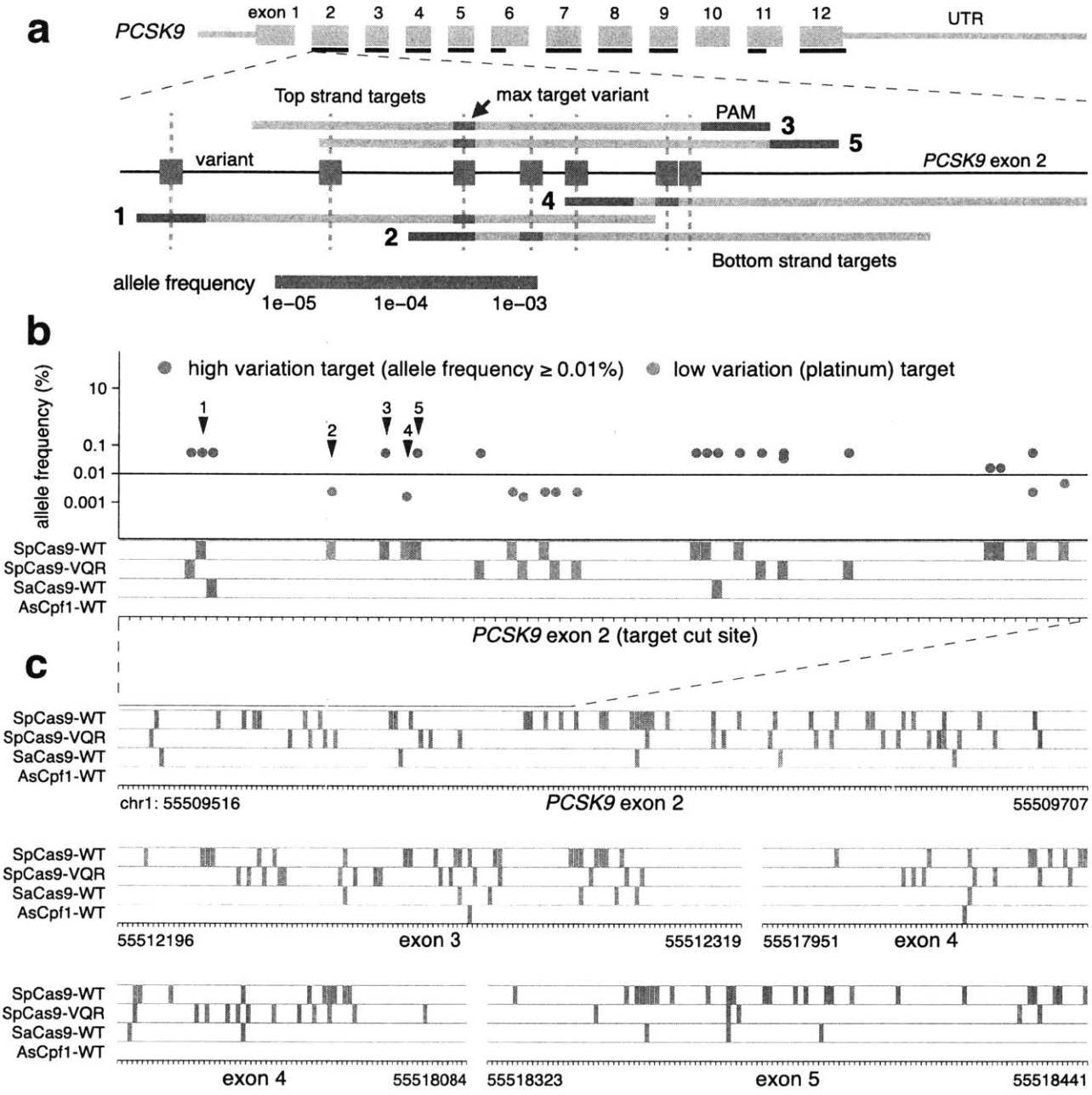


Figure 2: Selection of platinum targets maximizes population efficacy. a, Schematic showing target variation within exon 2 of *PCSK9-001*, with regions containing high coverage in the ExAC dataset indicated (black lines below exons). **b,** Frequency of target variation plotted by cut site position for targets spanning the start of *PCSK9-001* exon 2, with targets shown in (a) indicated by arrows. The horizontal line at 0.01% separates platinum targets (grey) from targets with high variation (red). The classification for each target is depicted below for each enzyme (grey or red boxes). **c,** Classification of targets for each enzyme spanning exons 2 – 5 of *PCSK9-001*.

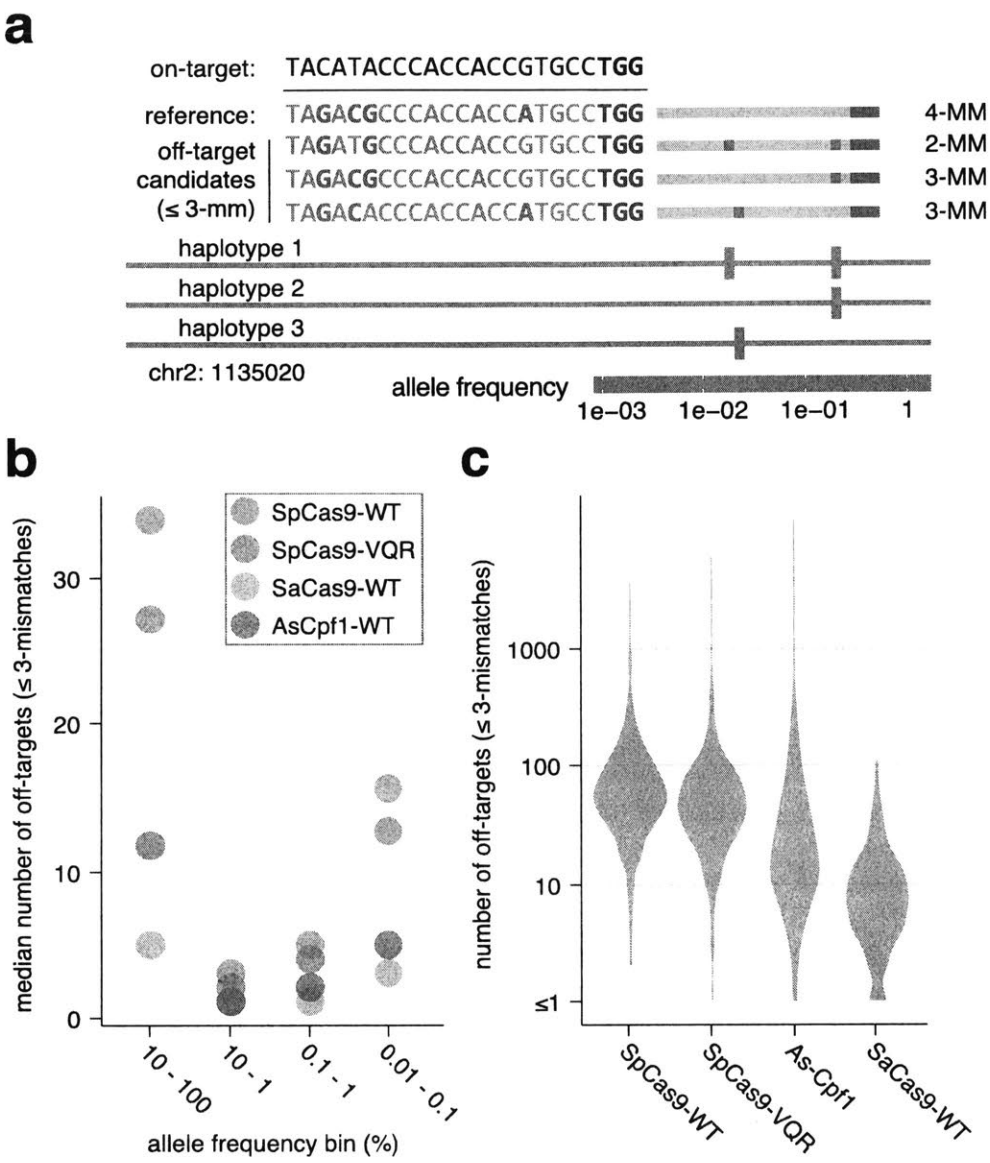


Figure 3: Human genetic variation significantly impacts CRISPR endonuclease therapeutic safety. a, Schematic illustrating off-target candidates arising due to multiple different haplotypes. **b,** Number of off-target candidates for each CRISPR endonuclease at different allele frequencies. **c,** Distribution of the number of off-target candidates per platinum target for each CRISPR endonuclease.

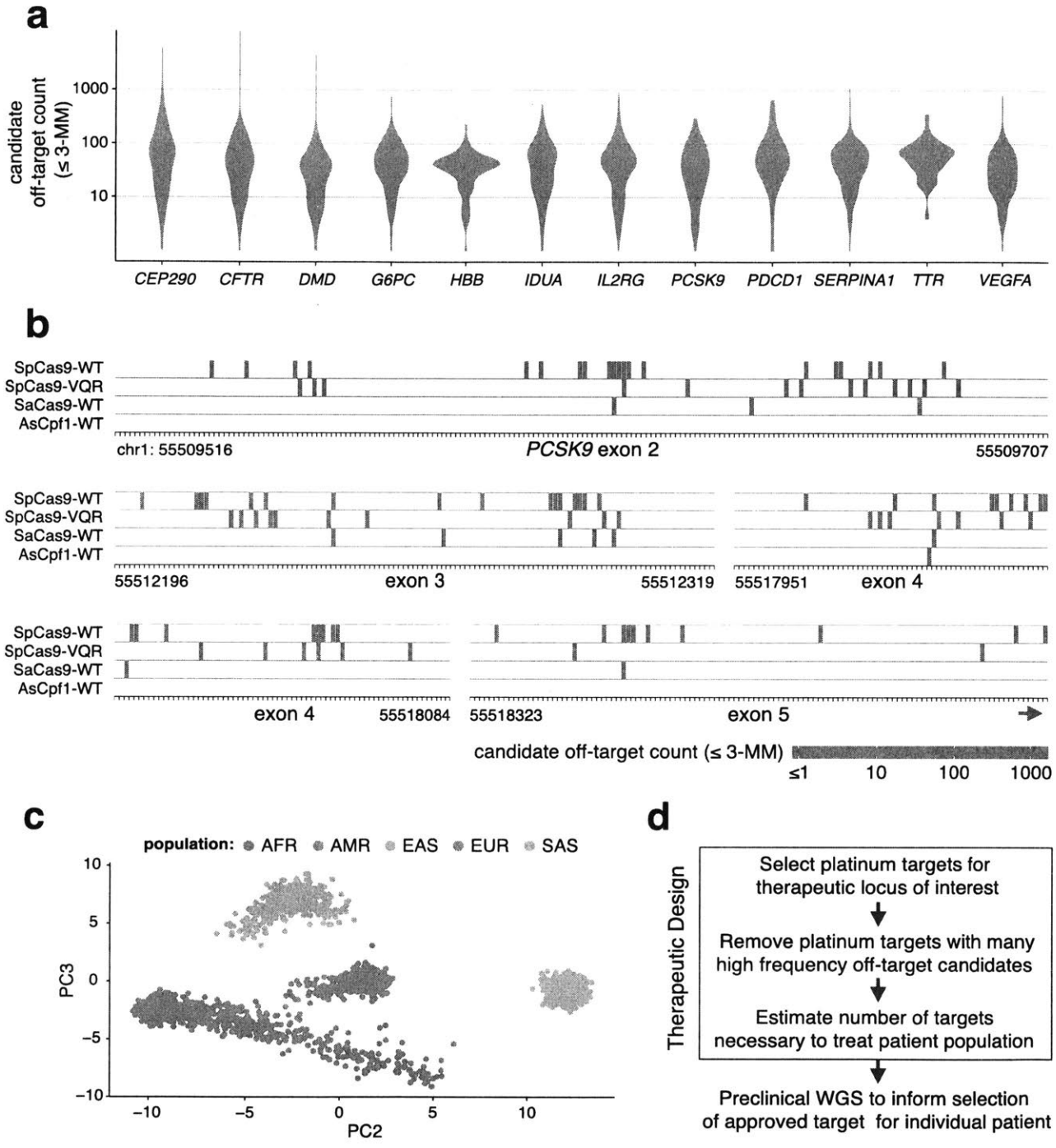


Figure 4: Gene- and population-specific variation informs therapeutic design. a, Distribution of the number of off-target candidates per platinum target for 12 therapeutically relevant genes. **b,** Total off-target candidates for platinum targets spanning exons 2 – 5 of *PCSK9-001* are shown for each enzyme. **c,** Principal component analysis (PCA) separating 1000 Genomes individuals into super populations based on patient-specific off-target profiles for platinum targets spanning 12 therapeutically relevant genes. PC2 and PC3 are shown. AFR, African; AMR, Ad mixed

American; EAS, East Asian; EUR, European; SAS, South Asian. **d,** Proposed therapeutic design framework.

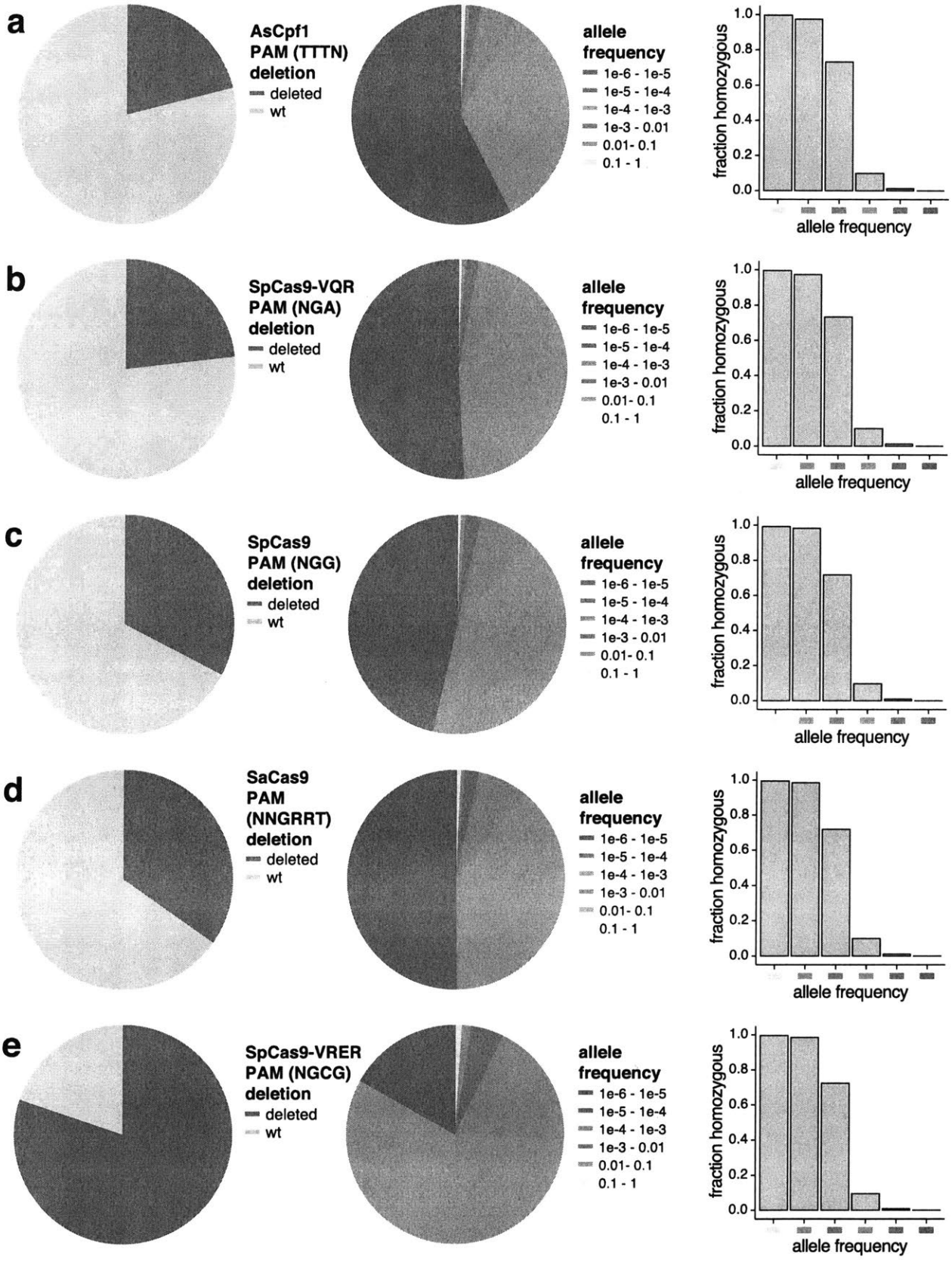

a
AsCpf1
PAM (TTTN)
deletion
deleted
wt
allele
frequency
1e-6 - 1e-5
1e-5 - 1e-4
1e-4 - 1e-3
1e-3 - 0.01
0.01- 0.1
0.1 - 1
fraction homozygous
1.0
0.8
0.6
0.4
0.2
0.0
allele frequency
b
SpCas9-VQR
PAM (NGA)
deletion
deleted
wt
allele
frequency
1e-6 - 1e-5
1e-5 - 1e-4
1e-4 - 1e-3
1e-3 - 0.01
0.01- 0.1
0.1 - 1
fraction homozygous
1.0
0.8
0.6
0.4
0.2
0.0
allele frequency
c
SpCas9
PAM (NGG)
deletion
deleted
wt
allele
frequency
1e-6 - 1e-5
1e-5 - 1e-4
1e-4 - 1e-3
1e-3 - 0.01
0.01- 0.1
0.1 - 1
fraction homozygous
1.0
0.8
0.6
0.4
0.2
0.0
allele frequency
d
SaCas9
PAM
(NNGRRT)
deletion
deleted
wt
allele
frequency
1e-6 - 1e-5
1e-5 - 1e-4
1e-4 - 1e-3
1e-3 - 0.01
0.01- 0.1
0.1 - 1
fraction homozygous
1.0
0.8
0.6
0.4
0.2
0.0
allele frequency
e
SpCas9-VRER
PAM (NGCG)
deletion
deleted
wt
allele
frequency
1e-6 - 1e-5
1e-5 - 1e-4
1e-4 - 1e-3
1e-3 - 0.01
0.01- 0.1
0.1 - 1
fraction homozygous
1.0
0.8
0.6
0.4
0.2
0.0
allele frequency

Figure S1: a – e, Left, fraction of PAMs altered by variants in the ExAC dataset; center, distribution of PAM-altering variant frequencies; right, fraction of homozygous variants by frequency. Data shown for AsCpf1 (a), SpCas9-VQR (b), SpCas9 (c), SaCas9 (d), and SpCas9-VRER (e).

			Whole-exome target variation by allele frequency (%)						
protein	**PAM**	**orientation**	**≥10%**	**≥1**	**≥0.1**	**≥0.01**	**≥0.001**	**total**	**n**
SpCas9-WT	NGG	right	2.03	4.16	11.03	31.14	79.39	95.44	10286445
SpCas9-VQR	NGA	right	1.81	3.66	9.65	27.66	75.77	94.27	9838603
AsCpf1-WT	TTTN	left	1.61	3.25	8.52	24.35	71.71	93.09	2702056
SaCas9-WT	NNGRRT	right	1.94	3.85	10.05	28.53	76.81	94.78	1938911
SpCas9-VRER	NGCG	right	2.70	5.72	15.77	44.21	91.82	98.44	981524

Table S1: Fraction of targets containing target variation for each CRISPR endonuclease (n specifies the number of protein coding targets in the human exome for each enzyme).

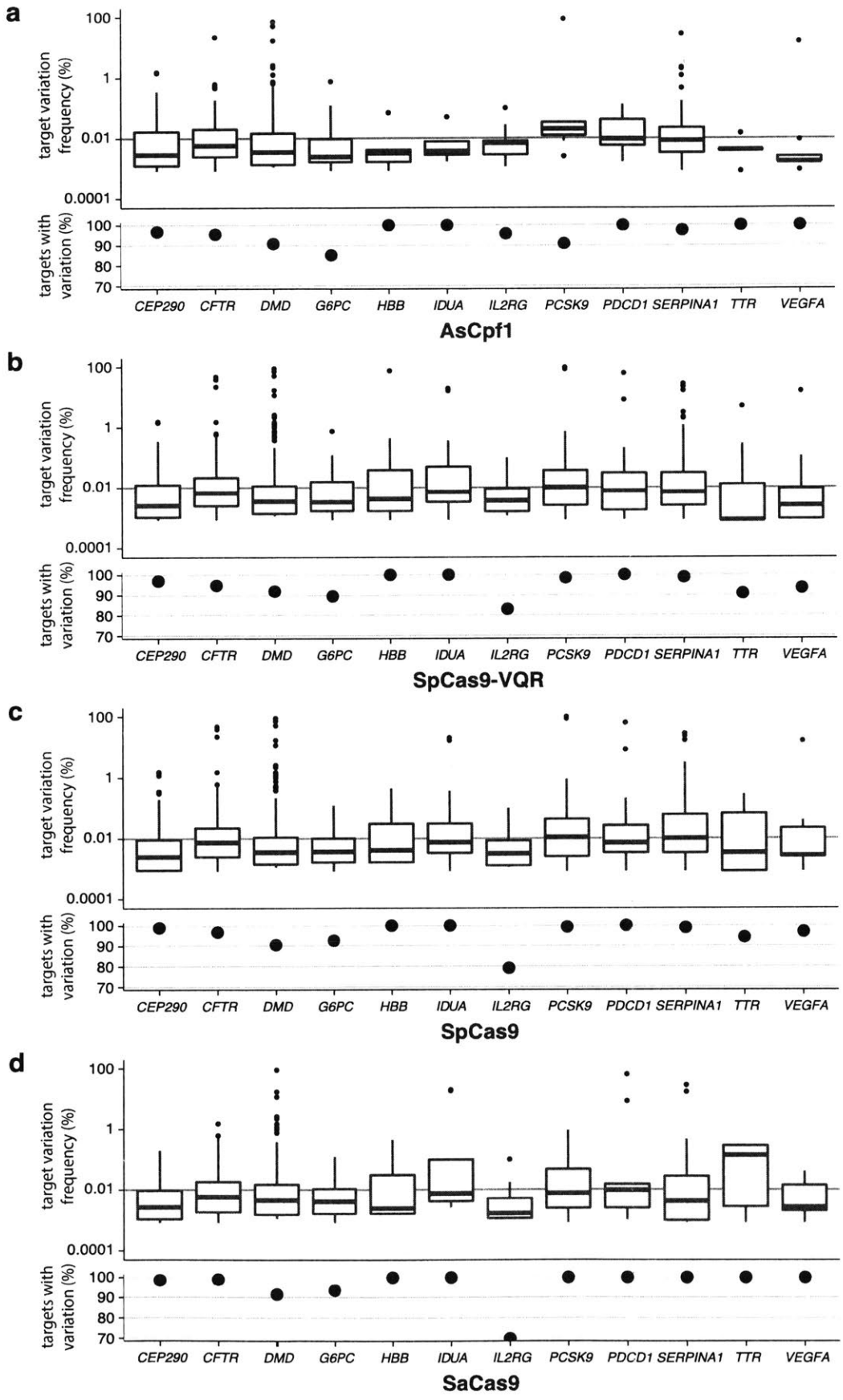

a
target variation frequency (%)
100
1
0.01
0.0001
targets with variation (%)
100
90
80
70
CEP290 CFTR DMD G6PC HBB IDUA IL2RG PCSK9 PDCD1 SERPINA1 TTR VEGFA
AsCpf1
b
target variation frequency (%)
targets with variation (%)
CEP290 CFTR DMD G6PC HBB IDUA IL2RG PCSK9 PDCD1 SERPINA1 TTR VEGFA
SpCas9-VQR
c
target variation frequency (%)
targets with variation (%)
CEP290 CFTR DMD G6PC HBB IDUA IL2RG PCSK9 PDCD1 SERPINA1 TTR VEGFA
SpCas9
d
target variation frequency (%)
targets with variation (%)
CEP290 CFTR DMD G6PC HBB IDUA IL2RG PCSK9 PDCD1 SERPINA1 TTR VEGFA
SaCas9

Figure S2: a – d, Top, distribution of target variation for therapeutically relevant genes. Targets with frequencies of variation less than 0.01% (red line) are considered platinum. Bottom, fraction of all targets in these genes containing variation. Data shown for AsCpf1 (a), SpCas9-VWR (b), SpCas9-WT (c), SaCas9-WT (d).

number of platinum targets and non-platinum targets for 12 therapeutically relevant genes

enzyme	target classification	CEP290	CFTR	DMD	G6PC	HBB	IDUA	IL2RG	PCSK9	PDCD1	SERPINA1	TTR	VEGFA	total
AsCpf1	platinum	218	135	356	22	7	3	21	3	3	22	5	10	805
AsCpf1	not platinum	97	90	133	5	1	1	3	8	2	18	1	1	360
SpCas9	platinum	154	185	670	106	40	78	127	135	96	96	34	46	1767
SpCas9	not platinum	45	134	213	32	22	40	7	137	55	93	18	19	815
SpCas9-VQR	platinum	348	320	948	99	22	45	94	70	37	94	31	47	2155
SpCas9-VQR	not platinum	132	185	303	34	15	29	12	66	29	73	12	14	904
SaCas9	platinum	53	67	160	23	9	5	29	20	9	21	3	11	410
SaCas9	not platinum	16	35	65	8	4	4	4	12	4	13	5	4	174
all	platinum	773	707	2134	250	78	131	271	228	145	233	73	114	5137
all	not platinum	290	444	714	79	42	74	26	223	90	197	36	38	2253
all	platinum/total (%)	72.7	61.4	74.9	76.0	65.0	63.9	91.2	50.6	61.7	54.2	67.0	75.0	69.5

Table S2: Number of platinum and high variation frequency targets for 12 therapeutically relevant genes for AsCpf1, SpCas9, SpCas9-VQR, or SaCas9.

Source	nucleotide	A	T	C	G
ExAC	exome nt fraction (%)	24.64	24.73	25.45	25.18
	nt variantion fraction (%)	10.28	10.18	20.12	20.16
	exome nt fraction (%)	22.70	22.29	2.75	2.89
	nt variantion fraction (%)	14.53	14.51	66.24	63.82
	nucleotide	C (non CpG)	C (non CpG)	C (CpG)	G (CpG)

Table S3: Fraction of total residues and fraction of residues containing variation for individual nucleotides (nt) in the human exome.

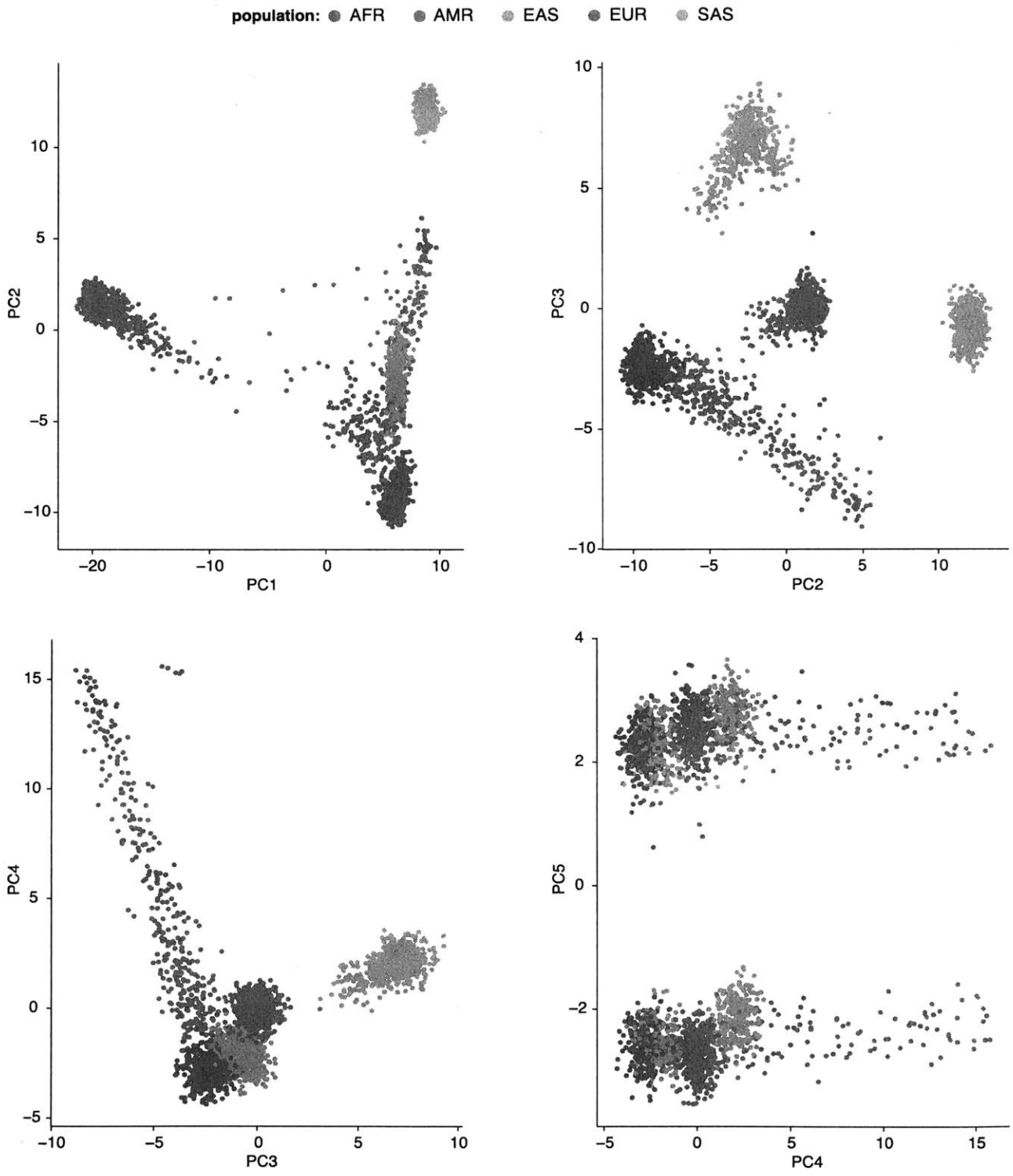


Figure S3: Separation of 1000 Genomes individuals into super populations based on patient specific off-target profiles for targets spanning 12 therapeutically relevant genes. Principle components 1 – 5 shown. AFR, African; AMR, Ad mixed American; EAS, East Asian; EUR, European; SAS, South Asian.

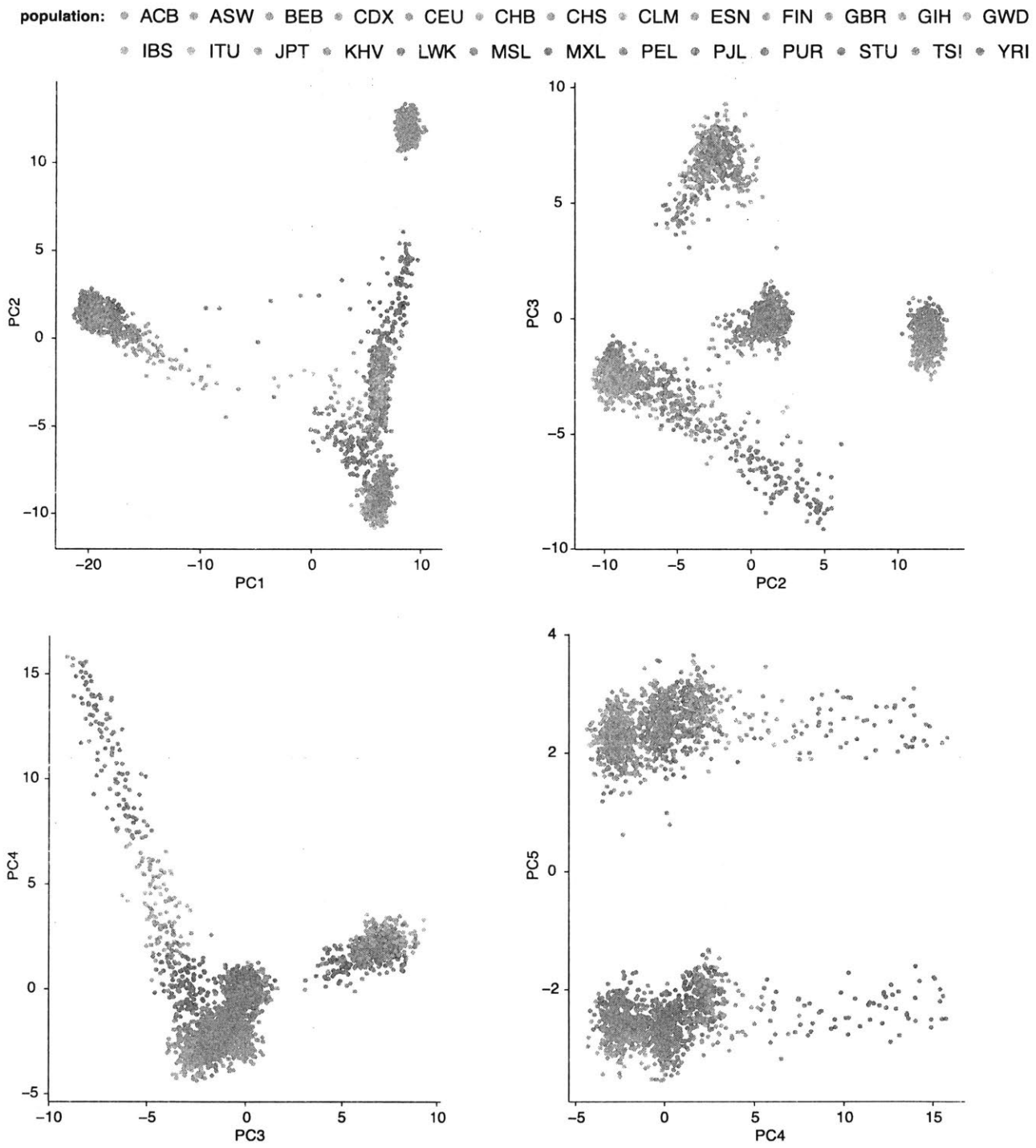


Figure S4: Separation of 1000 Genomes individuals into populations based on patient specific off-target profiles for targets spanning 12 therapeutically relevant genes. Principle components 1 – 5 shown. CHB, Han Chinese in Bejing, China; JPT, Japanese in Tokyo, Japan; CHS, Southern Han Chinese; CDX, Chinese Dai in Xishuangbanna, China; KHV, Kinh in Ho Chi Minh City, Vietnam; CEU, Utah Residents (CEPH) with Northern and Western Ancestry; TSI, Toscani in Italia; FIN, Finnish in Finland; GBR, British in England and Scotland; IBS, Iberian Population in Spain; YRI, Yoruba in

Ibadan, Nigeria; LWK, Luhya in Webuye, Kenya; GWD, Gambian in Western Divisions in the Gambia; MSL, Mende in Sierra Leone; ESN, Esan in Nigeria; ASW, Americans of African Ancestry in SW USA; ACB, African Caribbeans in Barbados; MXL, Mexican Ancestry from Los Angeles USA; PUR, Puerto Ricans from Puerto Rico; CLM, Colombians from Medellin, Colombia; PEL, Peruvians from Lima, Peru; GIH, Gujarati Indian from Houston, Texas; PJL, Punjabi from Lahore, Pakistan; BEB, Bengali from Bangladesh; STU, Sri Lankan Tamil from the UK; ITU, Indian Telugu from the UK.

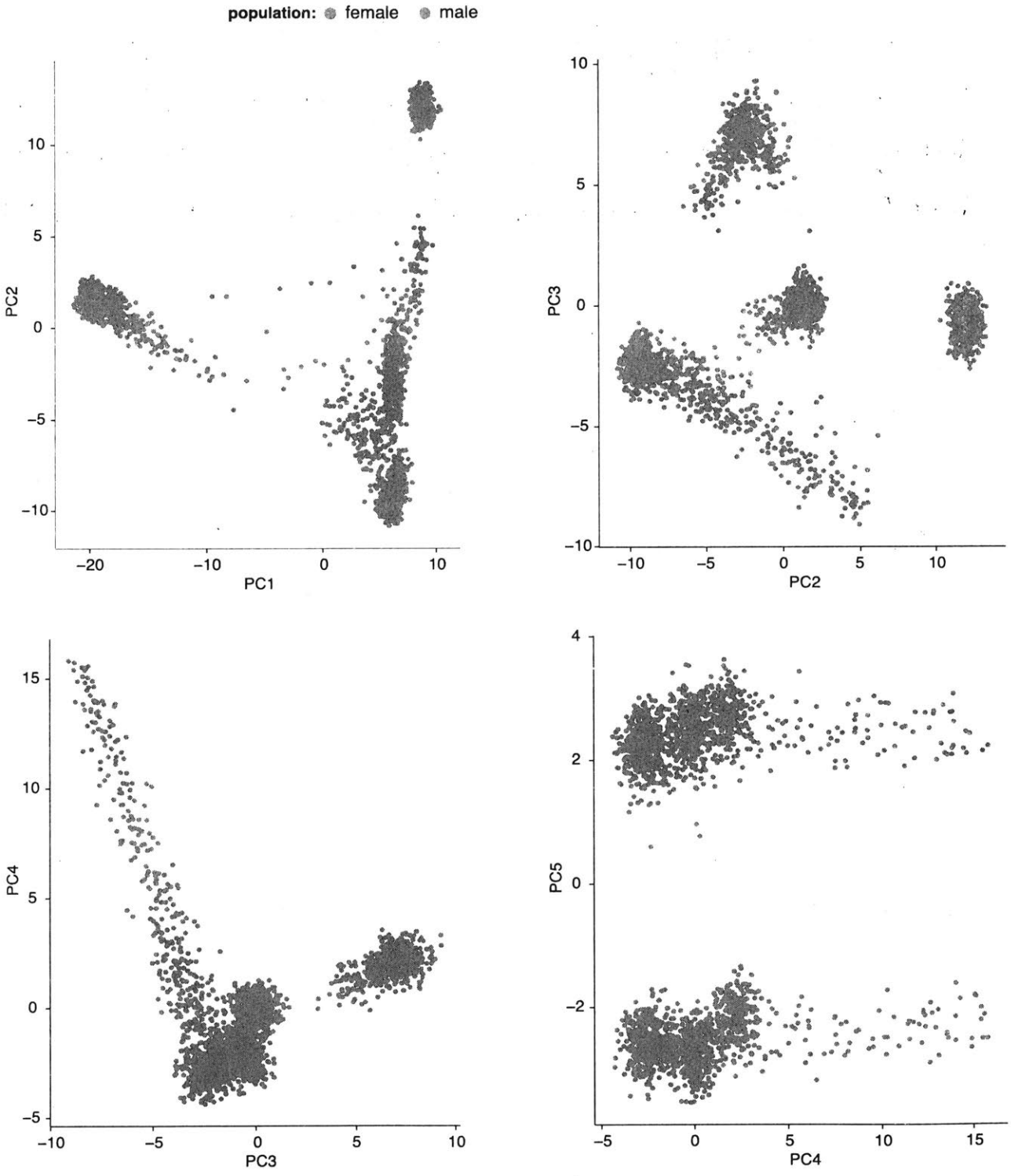


Figure S5: Separation of 1000 Genomes individuals by sex based on patient specific off-target profiles for targets spanning 12 therapeutically relevant genes. Principle components 1 – 5 shown.

CONCLUSION

Type II CRISPR-Cas single effector endonucleases enable high efficiency double stranded DNA cleavage in the mammalian genome. We and others have demonstrated that Cas9 has robust and uniform cleavage activity across genomic targets in regions of varying chromatin structure and epigenetic context. While robust Cas9 enzyme activity across genomic targets with varying chromatin complexity and epigenetic marks is desirable, CRISPR-Cas9 mismatch tolerance is not. Our work as well as studies from other groups demonstrates that SpCas9 and SaCas9 orthologs show tolerance of both mismatches and bulges between the RNA guide and DNA targets[49,52,63,50,74]. The complexity of mismatch and bulge tolerance complicate computational prediction of Cas9 off-targets and motivate unbiased approaches to interrogate the genome wide specificity of individual Cas9 targets. High specificity engineered SpCas9 variants, eSpCas9[72] and Cas9-HF1[81], now exist and show decreased tolerance of mismatches between DNA targets and the RNA guide (relevant to mismatches in approximately the PAM distal 12-14 nucleotides of the guide RNA given 20nt of guide RNA target complementarity). Additionally, specificity characterization of the recently discovered CRISPR endonuclease Cpf1 shows decreased mismatch tolerance. Nonetheless, the specificity of Cpf1 and high-fidelity Cas9 variants is not perfect, and it remains critical to comprehensively validate the specificity of CRISPR-Cas targets for therapeutic applications.

Multiple experimental methods have emerged from our group and others to assay off-target activity in a reportedly unbiased, genome-wide manner. These include GUIDEseq, integrase deficient lentiviral (IDLV) integration, HTGTS, Digenome-seq,

BLESS, and BLISS[57,22,70,82,74]. Although all these methods have proven very useful to elucidate the genome-wide landscape of double stranded breaks (DSBs) accompanying CRISPR-Cas genome editing, they all present a number of drawbacks. GUIDEseq, IDLV integration, and HTGTS methods all rely errors in NHEJ to capture the location of DSBs in the genome, limiting the sensitivity of these techniques. Furthermore, GUIDEseq and IDLV are limited in their versatility due to the necessity of delivering high concentration exogenous DNA fragments to cells for capture at DSB sites. While BLESS and BLISS directly label double stranded breaks and are amenable to *inVivo* applications, these methods are severely limited by high levels of background DSB detection and require very deep sequencing to achieve high sensitivity. Additionally, all of these methods occur in the context of a cell, where diverse chromatin structure and epigenetic modification of the genome introduce contextual biases.

Although it has long been considered advantageous to investigate the specificity of CRISPR-Cas in cells, *inVitro* approaches to genome-wide specificity offer enhanced sensitivity for the detection of off-target cleavage activity by CRISPR-Cas enzymes. The general goal of such techniques is to capture the superset of all possible off-target activity associated with particular CRISPR-Cas target. For therapeutic genome editing, CRISPR endonucleases are likely to contact large numbers of heterogeneous cell types, and *inVitro* approaches enable profiling of specificity invariant to differences in chromatin structure and epigenetic states in diverse cell types.

One method to detect CRISPR-Cas off targets *inVitro* is Digenome-seq, where the genomic DNA (gDNA) from a cell is purified and *inVitro* digested using a CRISPR-Cas enzyme-RNA guide pair[82]. All of the digested gDNA is then prepped for next generation

sequencing. The extraction of gDNA prior to CRISPR-Cas DNA cleavage removes all cellular context, with the goal of discovering the superset of all the off targets for different cell types. However, because there is no ability to enrich for CRISPR-Cas9 induced breaks using this method, whole genome sequencing (WGS) and rigorous computational filtering are required to identify CRISPR-Cas induced cleavage sites using Digenome-seq. Despite steady reduction in the cost of WGS, this is still an expensive proposition that additionally may result in a loss of sensitivity due the limited sequencing depth as well as biases in library preparation, WGS readout, and analysis.

Further complicating assessment of the efficiency and safety of CRISPR-Cas therapeutics is human genetic diversity. Current approaches for Cas9 target selection rank off-target candidates found in the reference human genome by both the number and position of RNA guide mismatches, with the assumption that loci containing less than 3 mismatches or containing PAM distal mismatches are more likely to be cleaved. However, in a population of individuals, this strategy is complicated by the existence of multiple haplotypes (sets of associated variants), which will contain different positions or numbers of mismatches at candidate off-target sites. Ultimately, our results suggest that it will be necessary to screen individual patients for perfect target guide RNA complementarity and rare off-target candidates.

Thus, there remains a need for an efficient, versatile and comprehensive method to evaluate the specificity of genome engineering technology. We are engaged in ongoing efforts to address this need with the creation of efficient, unbiased, and comprehensive *inVitro* assays to evaluate the specificity of programmable endonucleases.

REFERENCES

1. Lee, M. S., Gippert, G. P., Soman, K. V., Case, D. A. & Wright, P. E. Three-dimensional solution structure of a single zinc finger DNA-binding domain. *Science* **245,** 635–637 (1989).
2. Pavletich, N. P. & Pabo, C. O. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science* **252,** 809–817 (1991).
3. Moore, M., Klug, A. & Choo, Y. Improved DNA binding specificity from polyzinc finger peptides by using strings of two-finger units. *Proc. Natl. Acad. Sci.* **98,** 1437–1441 (2001).
4. Kim, Y. G., Cha, J. & Chandrasegaran, S. Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 1156–1160 (1996).
5. Bitinaite, J., Wah, D. A., Aggarwal, A. K. & Schildkraut, I. FokI dimerization is required for DNA cleavage. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 10570–10575 (1998).
6. Smith, J., Berg, J. M. & Chandrasegaran, S. A detailed study of the substrate specificity of a chimeric restriction enzyme. *Nucleic Acids Res.* **27,** 674–681 (1999).
7. Urnov, F. D. *et al.* Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature* **435,** 646–651 (2005).
8. Ciccia, A. & Elledge, S. J. The DNA Damage Response: Making it safe to play with knives. *Mol. Cell* **40,** 179–204 (2010).
9. Smithies, O., Gregg, R. G., Boggs, S. S., Koralewski, M. A. & Kucherlapati, R. S. Insertion of DNA sequences into the human chromosomal β-globin locus by homologous recombination. *Nature* **317,** 230–234 (1985).
10. Thomas, K. R., Folger, K. R. & Capecchi, M. R. High frequency targeting of genes to specific sites in the mammalian genome. *Cell* **44,** 419–428 (1986).
11. Bibikova, M. *et al.* Stimulation of Homologous Recombination through Targeted Cleavage by Chimeric Nucleases. *Mol. Cell. Biol.* **21,** 289–297 (2001).
12. Porteus, M. H. & Baltimore, D. Chimeric nucleases stimulate gene targeting in human cells. *Science* **300,** 763 (2003).
13. Bibikova, M., Beumer, K., Trautman, J. K. & Carroll, D. Enhancing Gene Targeting with Designed Zinc Finger Nucleases. *Science* **300,** 764–764 (2003).
14. Moore, J. K. & Haber, J. E. Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in Saccharomyces cerevisiae. *Mol. Cell. Biol.* **16,** 2164–2173 (1996).
15. Szczepek, M. *et al.* Structure-based redesign of the dimerization interface reduces the toxicity of zinc-finger nucleases. *Nat. Biotechnol.* **25,** 786–793 (2007).
16. Miller, J. C. *et al.* An improved zinc-finger nuclease architecture for highly specific genome editing. *Nat. Biotechnol.* **25,** 778–785 (2007).
17. Boch, J. *et al.* Breaking the Code of DNA Binding Specificity of TAL-Type III Effectors. *Science* **326,** 1509–1512 (2009).
18. Moscou, M. J. & Bogdanove, A. J. A Simple Cipher Governs DNA Recognition by TAL Effectors. *Science* **326,** 1501–1501 (2009).
19. Christian, M. *et al.* Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics* **186,** 757–761 (2010).

20. Wood, A. J. *et al.* Targeted genome editing across species using ZFNs and TALENs. *Science* **333,** 307 (2011).
21. Yuan, J. *et al.* Zinc-finger Nuclease Editing of Human cxcr4 Promotes HIV-1 CD4+ T Cell Resistance and Enrichment. *Mol. Ther.* **20,** 849–859 (2012).
22. Gabriel, R. *et al.* An unbiased genome-wide analysis of zinc-finger nuclease specificity. *Nat. Biotechnol.* **29,** 816–823 (2011).
23. Cong, L. *et al.* Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science* **339,** 819–823 (2013).
24. Mali, P. *et al.* RNA-Guided Human Genome Engineering via Cas9. *Science* **339,** 823–826 (2013).
25. Garneau, J. E. *et al.* The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468,** 67–71 (2010).
26. Zetsche, B. *et al.* Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. *Cell* **163,** 759–771 (2015).
27. Deltcheva, E. *et al.* CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471,** 602–607 (2011).
28. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337,** 816–821 (2012).
29. Wang, H. *et al.* One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* **153,** 910–918 (2013).
30. Marinus, M. G. & Morris, N. R. Isolation of deoxyribonucleic acid methylase mutants of Escherichia coli K-12. *J. Bacteriol.* **114,** 1143–1150 (1973).
31. May, M. S. & Hattman, S. Analysis of bacteriophage deoxyribonucleic acid sequences methylated by host- and R-factor-controlled enzymes. *J. Bacteriol.* **123,** 768–770 (1975).
32. Kelleher, J. E. & Raleigh, E. A. A novel activity in Escherichia coli K-12 that directs restriction of DNA modified at CG dinucleotides. *J. Bacteriol.* **173,** 5220–5223 (1991).
33. Guschin, D. Y. *et al.* A rapid and general assay for monitoring endogenous gene modification. *Methods Mol. Biol. Clifton NJ* **649,** 247–256 (2010).
34. Bogenhagen, D. F. & Brown, D. D. Nucleotide sequences in Xenopus 5S DNA required for transcription termination. *Cell* **24,** 261–270 (1981).
35. Bultmann, S. *et al.* Targeted transcriptional activation of silent oct4 pluripotency gene by combining designer TALEs and inhibition of epigenetic modifiers. *Nucleic Acids Res.* **40,** 5368–5377 (2012).
36. Valton, J. *et al.* Overcoming transcription activator-like effector (TALE) DNA binding domain sensitivity to cytosine methylation. *J. Biol. Chem.* **287,** 38427–38432 (2012).
37. Miller, J. C. *et al.* A TALE nuclease architecture for efficient genome editing. *Nat. Biotechnol.* **29,** 143–148 (2011).
38. Mussolino, C. *et al.* A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. *Nucleic Acids Res.* **39,** 9283–9293 (2011).
39. Hsu, P. D. & Zhang, F. Dissecting neural function using targeted genome engineering technologies. *ACS Chem. Neurosci.* **3,** 603–610 (2012).
40. Sanjana, N. E. *et al.* A transcription activator-like effector toolbox for genome engineering. *Nat. Protoc.* **7,** 171–192 (2012).
41. Sander, J. D. *et al.* Selection-free zinc-finger-nuclease engineering by context-dependent assembly (CoDA). *Nat. Methods* **8,** 67–69 (2011).

42. Bobis-Wozowicz, S., Osiak, A., Rahman, S. H. & Cathomen, T. Targeted genome editing in pluripotent stem cells using zinc-finger nucleases. *Methods San Diego Calif* **53,** 339–346 (2011).
43. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152,** 1173–1183 (2013).
44. 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).
45. Cho, S. W., Kim, S., Kim, J. M. & Kim, J.-S. Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nat. Biotechnol.* **31,** 230–232 (2013).
46. Michaelis, L. & Menten, M. L. Die Kinetik der Invertinwirkung.
47. Mahfouz, M. M. *et al.* De novo-engineered transcription activator-like effector (TALE) hybrid nuclease with novel DNA binding specificity creates double-strand breaks. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 2623–2628 (2011).
48. Wilson, E. B. Probable Inference, the Law of Succession, and Statistical Inference. *J. Am. Stat. Assoc.* **22,** 209–212 (1927).
49. Hsu, P. D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31,** 827–832 (2013).
50. Lin, Y. *et al.* CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. *Nucleic Acids Res.* **42,** 7473–7485 (2014).
51. Paruzynski, A. *et al.* Genome-wide high-throughput integrome analyses by nrLAM-PCR and next-generation sequencing. *Nat. Protoc.* **5,** 1379–1395 (2010).
52. Fu, Y. *et al.* High frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.* **31,** 822–826 (2013).
53. Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell* **136,** 215–233 (2009).
54. Jinek, M. & Doudna, J. A. A three-dimensional view of the molecular machinery of RNA interference. *Nature* **457,** 405–412 (2009).
55. Jinek, M. *et al.* Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science* **343,** 1247997 (2014).
56. Nishimasu, H. *et al.* Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156,** 935–949 (2014).
57. 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).
58. Wu, X. *et al.* Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nat. Biotechnol.* **32,** 670–676 (2014).
59. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).
60. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14,** 1188–1190 (2004).
61. Crosetto, N. *et al.* Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nat. Methods* **10,** 361–365 (2013).
62. Mali, P. *et al.* CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat. Biotechnol.* **31,** 833–838 (2013).
63. Pattanayak, V. *et al.* High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat. Biotechnol.* **31,** 839–843 (2013).

64. Bae, S., Park, J. & Kim, J.-S. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinforma. Oxf. Engl.* **30,** 1473–1475 (2014).
65. Kuscu, C., Arslan, S., Singh, R., Thorpe, J. & Adli, M. Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. *Nat. Biotechnol.* **32,** 677–683 (2014).
66. Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S. D. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151,** 2551–2561 (2005).
67. Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315,** 1709–1712 (2007).
68. Sapranauskas, R. *et al.* The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. *Nucleic Acids Res.* **39,** 9275–9282 (2011).
69. Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc. Natl. Acad. Sci. U. S. A.* **109,** E2579-2586 (2012).
70. Frock, R. L. *et al.* Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases. *Nat. Biotechnol.* **33,** 179–186 (2015).
71. Van Gelder, R. N. *et al.* Amplified RNA synthesized from limited quantities of heterogeneous cDNA. *Proc. Natl. Acad. Sci. U. S. A.* **87,** 1663–1667 (1990).
72. Slaymaker, I. M. *et al.* Rationally engineered Cas9 nucleases with improved specificity. *Science* **351,** 84–88 (2016).
73. Wang, X. *et al.* Unbiased detection of off-target cleavage by CRISPR-Cas9 and TALENs using integrase-defective lentiviral vectors. *Nat. Biotechnol.* **33,** 175–178 (2015).
74. Ran, F. A. *et al.* In vivo genome editing using Staphylococcus aureus Cas9. *Nature* **520,** 186–191 (2015).
75. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).
76. Consortium, T. 1000 G. P. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
77. Consortium, T. 1000 G. P. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
78. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
79. Kleinstiver, B. P. *et al.* Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature* **523,** 481–485 (2015).
80. Makarova, K. S. *et al.* An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol.* **13,** 722–736 (2015).
81. Kleinstiver, B. P. *et al.* High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529,** 490–495 (2016).
82. Kim, D. *et al.* Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nat. Methods* **12,** 237–243 (2015).
83. Kleinstiver, B. P. *et al.* Genome-wide specificities of CRISPR-Cas Cpf1 nucleases in human cells. *Nat. Biotechnol.* **34,** 869–874 (2016).
84. Kim, D. *et al.* Genome-wide analysis reveals specificities of Cpf1 endonucleases in human cells. *Nat. Biotechnol.* **34,** 863–868 (2016).

85. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).