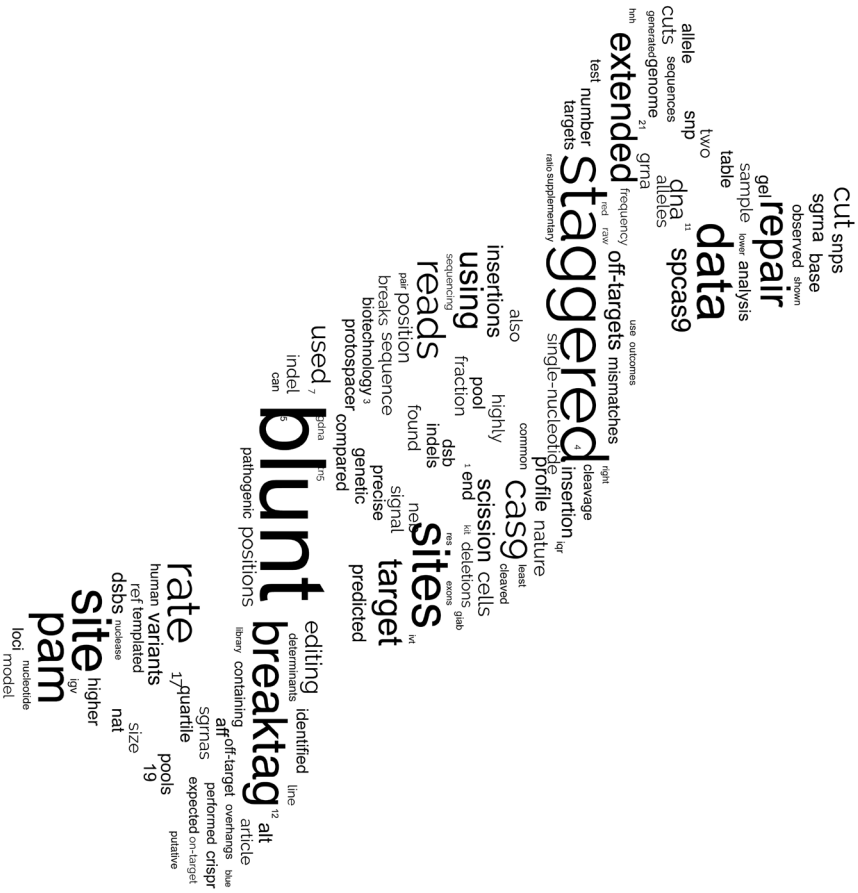
Illuminating Cas9 Scission Profile for Precise Genome Editing



Dissertation

Zur Erlangung des Grades

Doktor der Naturwissenschaften

Am Fachbereich Biologie

Der Johannes Gutenberg-Universität Mainz

Gabriel Mello da Cunha Longo

geboren am 13.07.1996

in Rio de Janeiro, Brasilien

Mainz, 2024

Zusammenfassung

CRISPR/Cas9 ist eine leistungsstarke Plattform zur Genom-Editierung und birgt enormes Potenzial für die erfolgreiche Gentherapie verschiedener genetischer Erkrankungen. Cas9 kann unterschiedliche Arten von DNA-Doppelstrangbrüchen (DSBs) erzeugen – stumpfe und versetzte; jedoch ist weitgehend unbekannt, was die Entscheidung von Cas9 für einen bestimmten Schnitt beeinflusst. Darüber hinaus wurde berichtet, dass die Struktur der DSB-Enden des von Cas9 induzierten Schnitts einen direkten Einfluss auf das Reparaturergebnis der für die Genbearbeitung anvisierten Loci hat. Das Schnittprofil bleibt ein übersehenes Merkmal von Cas-Nukleasen, und die Häufigkeiten und Determinanten der DSB-Endstrukturen sind nach wie vor unklar.

Hier haben wir BreakTag entwickelt, eine vielseitige, hochparallele und schnittbewusste Methodik zur Profilierung von Cas9-induzierten DSBs, um molekulare Determinanten zu identifizieren, die Cas9-Schnitte beeinflussen. Über die Prüfung der Genauigkeit von gRNAs und Nukleasen hinaus ist BreakTag eine unkomplizierte Methode zur Charakterisierung von CRISPR-Cas-Nukleasen. Untersuchung der Endstruktur von mehr als 150.000 einzigartig von Cas9 geschnittenen Loci im gesamten menschlichen Genom zeigte, dass zusätzlich zu den häufigen stumpfen DSBs etwa 34 % der SpCas9-Schnittstellen versetzte Enden mit 1 bis 3 Nukleotid-5'-Überhängen bilden. Fehlpaarungen zwischen der gRNA und Ziel-DNA beeinflussten das Schnittprofil, und die Verhältnisse von stumpfen und versetzten Schnitten waren zielabhängig. Trainieren eines maschinellen Lernmodells, das das Cas9-Schnittprofil vorhersagt, ergab, dass der Nuklease-Schnitt stark von der Protospacer-Sequenz abhängt, mit starken Sequenzdeterminanten. Wir zeigten weiterhin, dass genetische Variation die Cas9-Schnittkonfiguration beeinflusst und somit das DNA-Reparaturergebnis. Mit BreakTag identifizierten wir hochfidele Cas9-Varianten mit veränderten Schnittprofileigenschaften, die die Anzahl der für stark versetzte Schnitte geeigneten Loci erweitern. Durch den Vergleich übereinstimmender Datensätze von Cas9-Schnitten und Reparaturergebnissen stellten wir fest, dass Cas9-induzierte Brüche mit präzisen, vorlagenbasierten und vorhersagbaren Einzel-Nukleotid-Insertionen verbunden sind, was darauf hinweist, dass die Kontrolle des Cas9-versetzten Schnittprofils die Vorhersage von Reparatur-Genotypen mit wünschenswerten Insertionen und Deletionen ermöglichen könnte. In einem *Proof-of-Principle*-Experiment demonstrierten wir, dass ein schnittbewusstes gRNA-Design genutzt werden kann, um pathogene Einzel-Nukleotid-Deletionen zu korrigieren, Diese Ergebnisse unterstreichen die klinische Relevanz des Cas9-induzierten versetzten Schnittprofils.

Unsere Arbeit beleuchtet grundlegende Eigenschaften der Cas9-Nuklease und legt den Grundstein für die Nutzung des flexiblen Cas9-Schnittprofils und neu entwickelter Varianten für präzise, vorlagenunabhängige Genom-Editierung.

Summary

CRISPR/Cas9 is a powerful genome editing platform, holding immense potential for successful gene therapy of various genetic diseases. Cas9 can generate different types of DNA double-strand breaks (DSBs) – blunt and staggered; however, what dictates Cas9 scission decision is largely unknown. Furthermore, it has been reported that the DSB end structure of the Cas9-induced cut has a direct effect on the repair outcome of loci targeted for gene editing. The scission profile remains an overlooked feature of Cas nucleases, and frequencies and determinants of the DSB end structures remain elusive.

Here we developed BreakTag, a versatile, highly parallel and scission-aware methodology for the profiling of Cas9-induced DSBs to identify molecular determinants influencing Cas9 incisions. Beyond testing the fidelity of gRNAs and nucleases, BreakTag is a straightforward framework for the characterization of CRISPR-Cas nucleases. Assessing the end-structure of more than 150,000 uniquely cleaved by Cas9 loci across the human genome, we found that in addition to frequent blunt DSBs, approximately 34% of SpCas9 cut sites form staggered ends displaying 1 to 3 nucleotide 5' overhangs. The presence of mismatches between the gRNA and the target DNA influenced the scission profile, and the ratios of blunt and staggered cuts were target-dependent. Training a machine learning model that predicts Cas9 scission profile revealed that the nuclease incision is highly dependent on the protospacer sequence, with strong sequence determinants. We further demonstrated that genetic variation impacts Cas9 cut configuration and therefore, the DNA repair outcome. Using BreakTag, we identified high-fidelity Cas9 variants with altered scission profile properties, expanding the loci amenable for highly staggered cleavage. Comparing matched datasets of Cas9 incisions and repair outcome, we established that Cas9 staggered breaks are linked with precise, templated and predictable single-nucleotide insertions, indicating that controlling Cas9 staggered cut profile could allow prediction of repair genotypes with desirable indels. We demonstrated in a proof-of-principle experiment that a scission-aware gRNA design can be leveraged for correcting pathogenic single-nucleotide deletions, demonstrating the clinical application of staggered cleavage.

Our work illuminates fundamental characteristics of the Cas9 nuclease and lays the foundation for harnessing the flexible Cas9 cut profile and engineered variants for precise template-free genome editing.

Preface

The research presented in this dissertation was generated in the research group of Dr. Vassilis Roukos at the Institute of Molecular Biology (IMB) in Mainz, Germany. Additional contributions were made by the IMB Bioinformatics Core Facility and IMB Protein Production Core Facility. A patent covering part of this work was registered:

Patent Number: WO2024033378A1. Application Number: PCT/EP2023/071967 – Filled on August 9th of 2022.

Parts of the work presented here were published in the following research article:

***Longo GMC**, Sayols S, Kotini AG, Heinen S, Möckel MM, Beli P, Roukos V. Linking CRISPR-Cas9 double-strand break profiles to gene editing precision with BreakTag. Nat Biotechnol. 2024 May 13. doi: 10.1038/s41587-024-02238-8. PMID: 38740992.*

List of publications

Longo GMC*, Sayols S*, Kotini AG, Heinen S, Möckel MM, Beli P, Roukos V. Linking CRISPR-Cas9 double-strand break profiles to gene editing precision with BreakTag. **Nat Biotechnol**. 2024 May 13. doi: 10.1038/s41587-024-02238-8. PMID: 38740992.

Longo GMC, Roukos V. Territories or spaghetti? Chromosome organization exposed. **Nat Rev Mol Cell Biol.** 2021 Aug;22(8):508. doi: 10.1038/s41580-021-00372-8. PMID: 33854243.

Mosler T, Conte F, **Longo GMC**, Mikicic I, Kreim N, Möckel MM, Petrosino G, Flach J, Barau J, Luke B, Roukos V, Beli P. R-loop proximity proteomics identifies a role of DDX41 in transcription-associated genomic instability. **Nat Commun**. 2021 Dec 16;12(1):7314. doi: 10.1038/s41467-021-27530-y. PMID: 34916496; PMCID: PMC8677849.

Zhang N*, Harbers L*, Simonetti M*, Diekmann C, Verron Q, Berrino E, Bellomo SE, **Longo GMC**, Ratz M, Schultz N, Tarish F, Su P, Han B, Wang W, Onorato S, Grassini D, Ballarino R, Giordano S, Yang Q, Sapino A, Frisén J, Alkass K, Druid H, Roukos V, Helleday T, Marchiò C, Bienko M, Crosetto N. High clonal diversity and spatial genetic admixture in early prostate cancer and surrounding normal tissue. **Nat Commun**. 2024 Apr 24;15(1):3475. doi: 10.1038/s41467-024-47664-z. PMID: 38658552; PMCID: PMC11043350.

Sanders AD*, Meiers S*, Ghareghani M*, Porubsky D*, Jeong H, van Vliet MACC, Rausch T, Richter-Pechańska P, Kunz JB, Jenni S, Bolognini D, **Longo GMC**, Raeder B, Kinanen V, Zimmermann J, Benes V, Schrappe M, Mardin BR, Kulozik AE, Bornhauser B, Bourquin JP, Marschall T, Korbel JO. Single-cell analysis of structural variations and complex rearrangements with tri-channel processing. **Nat Biotechnol**. 2020 Mar;38(3):343-354. doi: 10.1038/s41587-019-0366-x. Epub 2019 Dec 23. PMID: 31873213; PMCID: PMC7612647.

Abu-Libdeh B*, Jhujh SS*, Dhar S*, Sommers JA,* Datta A, **Longo GM**, Grange LJ, Reynolds JJ, Cooke SL, McNee GS, Hollingworth R, Woodward BL, Ganesh AN, Smerdon SJ, Nicolae CM, Durlacher-Betzer K, Molho-Pessach V, Abu-Libdeh A, Meiner V, Moldovan GL, Roukos V, Harel T, Brosh RM Jr, Stewart GS. RECON syndrome is a genome instability disorder caused by mutations in the DNA helicase RECQL1. **J Clin Invest.** 2022 Mar 1;132(5):e147301. doi: 10.1172/JCI147301. PMID: 35025765; PMCID: PMC8884905.

Lemos LGT, **Longo GMDC**, Mendonça BDS, Robaina MC, Brum MCM, Cirilo CA, Gimba ERP, Costa PRR, Buarque CD, Nestal de Moraes G, Maia RC. The LQB-223 Compound Modulates Antiapoptotic

Proteins and Impairs Breast Cancer Cell Growth and Migration. **Int J Mol Sci**. 2019 Oct 12;20(20):5063. doi: 10.3390/ijms20205063. PMID: 31614718; PMCID: PMC6834317.

Leonel AV, Alisson-Silva F, Santos RCM, Silva-Aguiar RP, Gomes JC, **Longo GMC**, Faria BM, Siqueira MS, Pereira MG, Vasconcelos-Dos-Santos A, Chiarini LB, Slawson C, Caruso-Neves C, Romão L, Travassos LH, Carneiro K, Todeschini AR, Dias WB. Inhibition of O-GlcNAcylation Reduces Cell Viability and Autophagy and Increases Sensitivity to Chemotherapeutic Temozolomide in Glioblastoma. **Cancers (Basel).** 2023 Sep 27;15(19):4740. doi: 10.3390/cancers15194740. PMID: 37835434; PMCID: PMC10571858.

Bernardo PS, Guimarães GHC, De Faria FCC, **Longo GMDC**, Lopes GPF, Netto CD, Costa PRR, Maia RC. LQB-118 compound inhibits migration and induces cell death in glioblastoma cells. **Oncol Rep.** 2020 Jan;43(1):346-357. doi: 10.3892/or.2019.7402. Epub 2019 Nov 6. PMID: 31746438.

Chapter 1. Introduction

1.1.Precision medicine and gene editing

The first two decades of the 21st century have witnessed a rapid acceleration in our understanding of the genetic basis of many diseases. Fueled by the advent of genomics, the complex molecular genetics of several human diseases has been detangled, slowly shifting the "one-size-fits-all" model of clinical therapy to more tailored approaches. The principle of *precision medicine* accounts for an understanding of disease at a deeper level, considering the genetics of the pathology in order to develop more targeted therapy (Ashley, 2016). In the context of genetic diseases, the molecular root lies in the DNA sequence, in the form of mutations that can range from few bases to entire chromosomes. What once seemed like an unachievable goal – customizing the genome of an organism – has become increasingly possible with the latest progress in the genome engineering field. The ability to manipulate genomes is highly desirable for understanding the molecular basis of biological traits, with far-reaching impacts on society.

The field of genetic manipulation started with *transgenesis,* the intentional transfer of genetic information from one organism to another. Pioneer work achieved integration of viral DNA into mice embryos, that were able to develop into an adult animal (Jaenisch & Mintz, 1974). Although the integration of viral genome was uncontrolled in terms of location and copy number, this model is considered to be the first transgenic organism. Continuous progress in human genetics efforts led to the development of different strategies for nucleic acid delivery. Among the most notable advancements is gene therapy using Adeno-associated Virus (AAV), which employs the dependovirus as a vector for gene insertion into the human genome (Naso et al., 2017). This strategy has been successfully adapted for delivering functional copies of defective genes, showing success in treating diseases such as spinal muscular atrophy (SMA) and β-thalassemia (Burdett & Nuseibeh, 2023). However, the AAV vector delivers its cargo to a predefined integration site considered a safe harbor and cannot be programmed to integrate into new sequences.

The field of programmable genome editing in mammalian cells began with early demonstrations that homologous recombination can be applied to mutate specific genes in mouse embryonic cells (Doetschman et al., 1987; Thomas & Capecchi, 1987). The seminal works elegantly demonstrated that DNA segments can be integrated to/swapped with specific regions of the mouse genome by exploring homology between a donor template and the endogenous sequence, and it was one of

the first demonstrations that editing specific locations in a genome is feasible. However, the strategy had relatively low frequency, obtaining recombination frequencies of $x10^{-3}$ to $x10^{-6}$ (Thomas & Capecchi, 1987). A subsequent groundbreaking discovery was that DNA double-strand breaks (DSBs) are highly recombinogenic and can be utilized for increased gene exchange. Early work devised a strategy where a restriction site of a rare-cutting endonuclease, I-*Sce*I, was inserted into a cassette containing a selective marker. Co-delivery of an I-*Sce*I expression vector and a replacement homologous construct containing a proficient copy of the marker showed increased rates of recombination between the genome and the homologous template compared to when no DNA cut was induced (Cohen-Tannoudji et al., 1998; Rouet et al., 1994). Moreover, DNA DSBs might be repaired via end-joining pathways that scars the cleavage site with small mutations that can induce gene knock out. Collectively, the early body of work demonstrated that precise gene editing is viable by inducing controlled DNA DSBs, initially perceived as highly deleterious and unwanted assaults, can be harnessed for genome engineering.

1.2.First-generation genome editors

The role of DNA DSBs in increasing recombination frequencies prompted the field to discover and develop molecular scissors for site-specific DNA cleavage. Given the site-specific nature of restriction enzymes, homing endonucleases (HEs) or meganucleases are a candidate for such task. HEs are rare cutting enzymes that recognize large DNA sequences of up to ~30 base pairs (Pingoud & Silva, 2007). Due to the large size of the human genome, matching the restriction site of an HE with a region of interest is unlikely, and therefore, customizable nucleases are desirable tools that can expand the universe of targetable regions.

The ideal site-specific DNA cutter would allow for programmable recognition and cleavage of new sequences. The type IIS restriction enzymes are a class of DNA cutters characterized by a two-domain organization. These enzymes contain a DNA binding domain that recognizes the target sequence, and a DNA cleavage domain that has no sequence specificity, such as *Fok*I (Pingoud & Silva, 2007). Because of the lack of sequence-specificity of the DNA cleavage domain, *Fok*I offers an attractive framework for customizing DNA cutters by substituting the DNA binding domain with alternative DNA binding sequences.

1.2.1.Zinc-Finger Nucleases (ZFNs)

Zinc-finger proteins (ZFPs) are a large family of transcription factors with finger-like DNA binding domains. Each finger domain consists of ~30 amino acids, recognizing 3-4 base pairs in a target DNA sequence (Gersbach et al., 2014). Because each finger binds as one independent module, they

can be combined and designed to bind a predetermined DNA sequence. Fusion of the sequence-agnostic *Fok*I DNA cleavage site domain with different zinc fingers allows binding to new target sequences. These new enzymes were termed Zinc-Finger Nucleases (ZFNs) (Y. G. Kim et al., 1996). For successful DNA DSB, a right ZFN and a left ZFN nuclease are designed for a given target sequence of 9-12bp, as *Fok*I nicks only one of the strands of the DNA (Figure 1.1A). The target sequence is called spacer, and is composed of 5-7 base pairs. Since the early *in vitro* demonstrations of site-specific DNA cleavage using ZFNs, many ZFN pairs have been designed and successfully used to target individual genes in several organisms (Carroll, 2011).

In primary human cells, ZFNs have been tested against an X-linked severe combined immune deficiency (SCID) mutation in the IL2Ry gene, and high frequencies of HDR were observed (Urnov et al., 2005). The first clinical trial using ZFNs aimed to knockout the CCR5 coreceptor for human immunodeficiency virus (HIV) in patients. CD4+ T-cells were isolated from the blood of the patients, and ZFNs were delivered *ex vivo*, following by infusion of the modified cells. Encouragingly, one of four patients who could be tested for HIV viral load was undetectable, and the study redeemed CCR5-modified autologous CD4+ T-cells as a safe strategy (Tebas Pablo et al., 2014). Despite the promising use of ZFNs for precise gene editing, the assembly of the zinc finger domains is difficult, and the workflow from design to validation of ZFNs is cumbersome and not straightforward, requiring multiple rounds of optimization.

1.2.2.Transcription Activator-Like Nucleases (TALENs)

Transcription Activator-Like Effectors (TALEs) are proteins found in bacteria that infect plant species. TALEs work by binding to specific DNA sequence in the host genome, activating the expression of genes important for bacterial infection (Moore et al., 2014). Like ZFPs, TALEs have a modular DNA-binding domain composed of arrays of 34 amino acid repeats. The nucleotide specificity is determined by the positions 12 and 13 of each repeat region, and changing these amino acid sequences allow binding of different DNA sequences. Compared to ZFPs, TALEs have a simpler genetic code, and pose as an attractive framework for the design of novel site-specific nucleases.

Transcription Activator-Like Nucleases (TALENs) are a fusion of the modular DNA binding domain of TALEs with the *Fok*I DNA cleavage domain. Similar to ZFP, successful DNA cleavage is achieved by designing a right TALEN and a left TALEN flanking a desired spacer sequence (Christian et al., 2010) (Figure 1.1B). Due to its relatively simple design, TALENs were widely employed in genome editing endeavors across various organisms, leading to notable progress in TALEN-edited crops (Becker & Boch, 2021), as well as TALEN gene-edited CAR T cells for B cell

acute lymphoblastic (Qasim et al., 2017). However, production of TALENs is not straightforward and scaling is both cost- and time-inefficient.

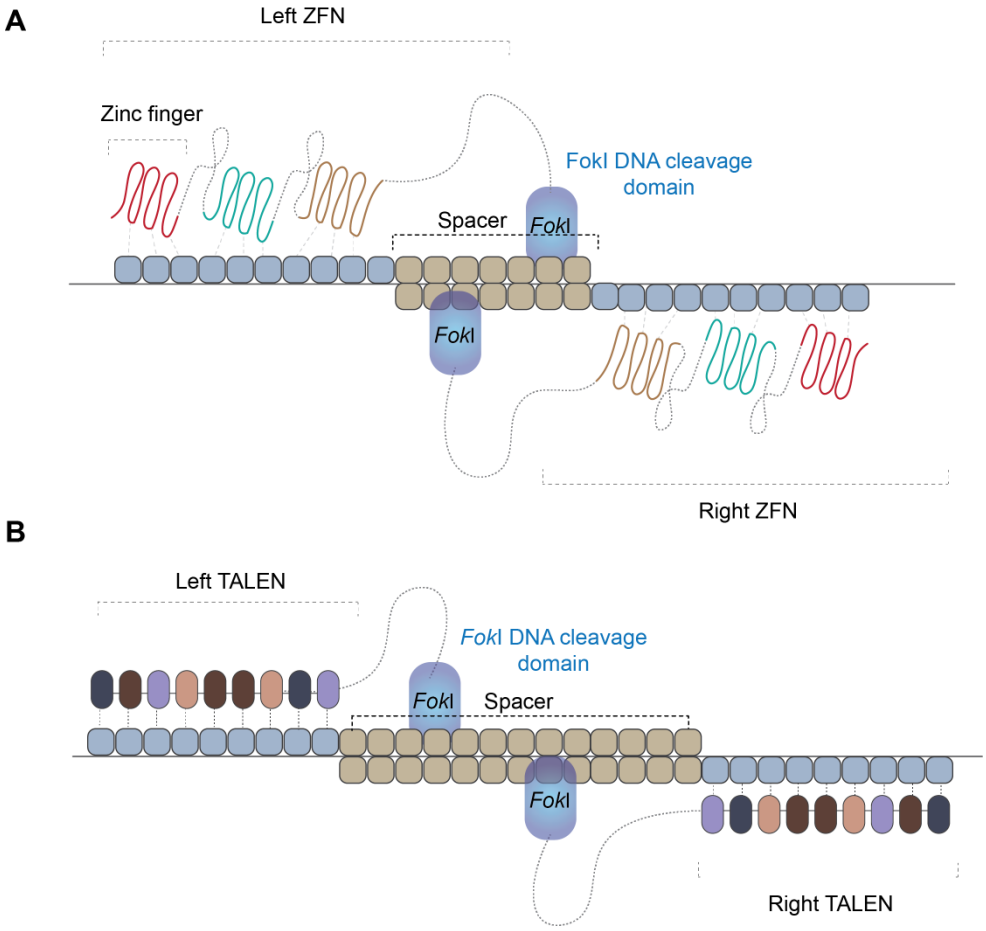


Figure 1.1. Schematic representation of first generation genome editors. (A) Representation of ZFNs. (B) Representation of TALENs.

1.3.The CRISPR revolution

The Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) are arrayed DNA sequences found in the genomes of prokaryotic organisms. The CRISPR-Cas system uses an RNA-guided nuclease, such as Cas9, to cleave exogenous DNA sequences in a target-specific manner. The sophisticated system of adaptive immunity has been repurposed for the easy and highly customizable targeting of specific DNA sequences in several species. Compared to ZFNs and TALENs, the CRISPR-Cas system is a "plug-and-play" mechanism that allows easy customization of target sequences by substituting the guide RNA molecule. Since the initial *in vitro* reconstitution of the CRISPR-Cas9 system, the field of genome engineering has seen rapid advancements, which

have had a profound impact on society. The discoveries that supported these advancements will be discussed in detail in this section.

1.3.1.CRISPR biology

The first observation of DNA repeats in the genome of prokaryotes that would later be known as CRISPR, was made in the late 80's in gram-negative bacteria (Ishino et al., 1987). Similar observations of palindromic repeats in the genome of closely related bacteria as well as unrelated species was reported (F. J. M. Mojica et al., 1993), indicating that this unknown locus might have an important function since it is widely conserved among bacteria species (Lander, 2016; F. j. m. Mojica et al., 1995; F. J. M. Mojica & Rodriguez-Valera, 2016). Analysis of the spacer sequence revealed that the origin of the fragments is from exogenous sources, and matched with the genome of bacteriophages (F. J. M. Mojica et al., 2005). These findings led to the hypothesis that the CRISPR locus is a recorder of past viral infections, a prokaryote adaptive immune response against bacteriophage infection.

The CRISPR-Cas system can be divided into 3 major stages: *adaptation, expression* and *interference* (Amitai & Sorek, 2016). During the adaptation phase, Cas (CRISPR-associated) proteins identify the target DNA and inserts a snippet of the viral genome into the CRISPR locus as a spacer (Figure 1.2A). The CRISPR array is transcribed in the form of a precursor RNA that is further cleaved into smaller units by RNAse III known as CRISPR RNAs (crRNAs) (Deltcheva et al., 2011), containing one spacer sequence and part of the repetitive portion of the array (Figure 1.2B). The crRNAs are assembled into an effector complex with a Cas protein, for example, Cas9. During the interference stage, the effector complex scans the bacterial cell space for exogenous sequences that are recognized via base complementarity with the crRNA sequences, leading to cleavage of the target DNA sequence, impeding viral infection (Figure 1.2C). The Cas nuclease first scans for a Protospacer Adjacent Motif (PAM), a short nucleotide sequence that triggers the effector complex to initiate DNA:crRNA complementarity test via base pairing (Jinek et al., 2012). The PAM motif is an important mechanism to differentiate self from exogenous. The bacterial CRISPR locus does not contain a PAM sequence integrated in the spacer sequence, and therefore, the nuclease cannot cleave the bacterium's own genome (F. J. M. Mojica & Rodriguez-Valera, 2016).

Due to its importance in prokaryote evolution, the CRISPR system shows a remarkable diversity of Cas proteins, array composition and target substrate, being broadly grouped into 2 classes. Class 1 CRISPR-Cas systems have effector modules composed of several Cas protein acting at the interference stage, and include types I, III and IV (Makarova et al., 2020). Class 2 CRISPR-Cas systems contain a single large multidomain protein, such as Cas9. This class includes types II, V

and VI, and its simplicity makes it attractive for the development of genome-editing tools. The Class 2 type VI effectors, such as Cas13, cleave the transcript and therefore target RNA molecules. Class 2 type II and type V both target DNA substrate, but type V contains a single RuvC-like domain that cleaves both strands (for example, Cas12a (Swarts, 2019; Zetsche et al., 2015)). The type II systems are composed of two nuclease domains, HNH and RuvC, as seen in Cas9 (Jiang et al., 2016; Jinek et al., 2012; Makarova et al., 2020).

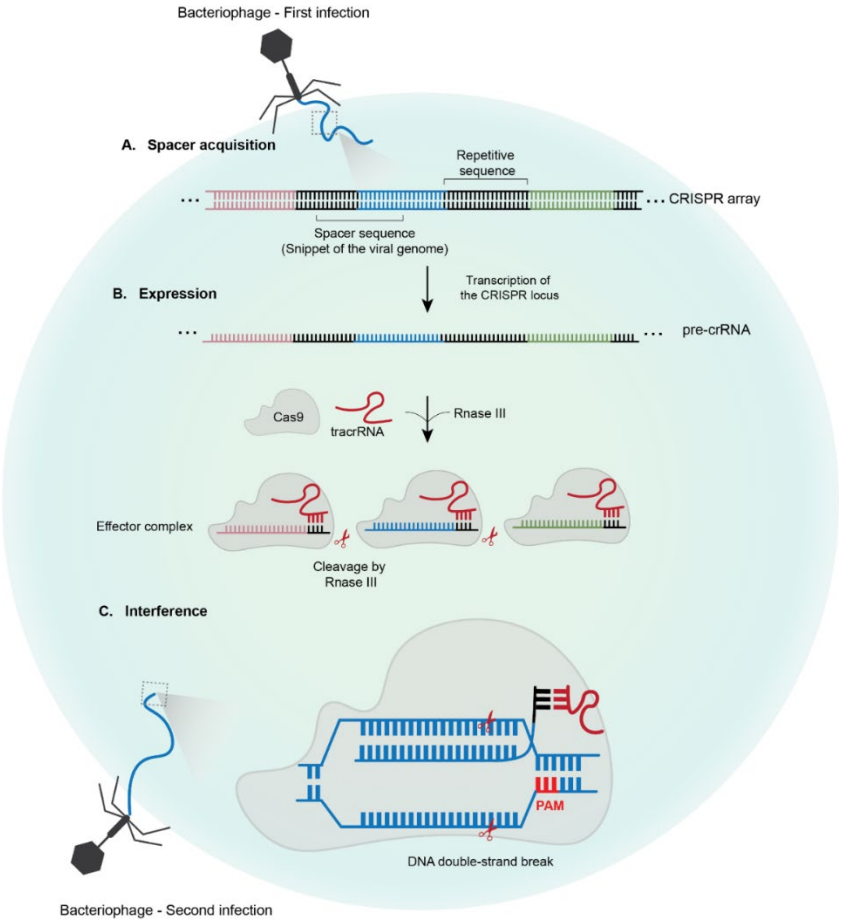


Figure 1.2. Steps of CRISPR defense mechanism, using the *Streptoccocus pyogenes* system as an example. (A) Snippets of viral genome are incorporated into the CRISPR locus. (B) The pre-crRNA is transcribed, a pre-effector complex is formed with the Cas9 and tracrRNA, and Rnase III cleaves the crRNA portion into individual effector complexes. (C) On a second infection, Cas9 recognizes a PAM-endowed viral sequence and cleaves the exogenous DNA, impeding the infection.

1.3.2.Harnessing CRISPR for precise gene editing

Following the discovery and characterization of the CRISPR-Cas system as a prokaryotic adaptive immune system, several studies set out to reconstitute the effector complexes type II system for

programmable DNA cleavage. The pioneer work from the Doudna and Charpentier groups successfully reconstituted the *Streptoccocus pyogenes* Cas9 (SpCas9) system *in vitro,* and demonstrated that SpCas9 utilizes the HNH and the RuvC domains to cleave both strands of a target DNA. They further stablished that Cas9 requires a crRNA, a tracrRNA and 5'-NGG-3' PAM sequence for successful cleavage, and programmable DNA targeting can be achieved by changing the crRNA portion of the system (Jinek et al., 2012). Similar findings were reported by another group around the same time using another typeII system from the *Streptococcus thermophilus* bacteria (Gasiunas et al., 2012).

The *in vitro* characterization of the CRISPR-Cas9 system prompted the field to reconstitute it in human cells for gene editing. Pioneer work used plasmids encoding the SpCas9, SpRNAseIII, tracrRNA and pre-crRNA components for transfection into 293FT cells. Using the surveyor assay, the authors identified that the system had successfully cleaved the target sequence as measured by the percentage of indels. Moreover, the authors demonstrated a D10A mutation in the RuvC domain converts the nuclease into a nickase, and integration of exogenous sequences via Homology Directed Repair (HDR) can be achieved by co-delivery of CRISPR and a donor DNA template (Cong et al., 2013). Similar findings were reported by the another study, where the AAVS1 locus was successfully engineered in human cells using the CRISPR-Cas9 system (Mali et al., 2013)

These seminal scientific papers started a revolution on the gene editing field, demonstrating that the simple yet effective CRISPR-Cas9 system can be customizable for programmable gene editing. Customization of target sequencing can be achieved by simply changing the crRNA portion of the gRNA, eliminating labor-intensive protein design and production as for ZFNs and TALENs. In 2020, Dr. Jennifer Doudna and Dr. Emmanuelle Charpentier were awarded the Nobel Prize in Chemistry for discovering the CRISPR-Cas9 system (Westermann et al., 2021).

1.3.3.The Cas9 protein

The SpCas9 protein is a large multidomain protein (1,368 amino acids) with DNA endonuclease activity (Figure 1.3A). Its cleavage activity is mediated by an HNH domain that cleaves the target strand (TS), and the RuvC domain cleaves the nontarget strand (NTS), generating a DNA DSB (Jinek et al., 2012). The HNH domain cleaves the TS between positions 17 and 18 of the target sequence, whereas the RuvC domain can cleave the NTS at multiple positions in addition to 17|18 (Jinek et al., 2012; Jr et al., 2021; Molla & Yang, 2020; Shen et al., 2018; Stephenson et al., 2018). SpCas9 has a bilobal organization, with the helical recognition lobe (REC) and a nuclease (NUC) lobe containing the HNH and split RuvC domains. The lobes are connected by an arginine-rich

bridge with intrinsic disorder properties. The HNH and RuvC domains are hinged by the Linker 1 (L-I) and Linker 2 (L-II) domains. A C-terminal domain mediates the Cas9-PAM interaction necessary for cleavage (Figure 1.3B).

The large size of the SpCas9 protein poses a problem for protein delivery in therapeutic settings. For example, Adeno-associated virus (AAV) vectors are commonly used as a mechanism for gene therapy, but the AAV cargo of 4.7 kb is a limitation for packaging the 4.2 kb SpCas9 in a single vector (Cebrian-Serrano & Davies, 2017). One route to solving this problem is to utilize alternative CRISPR effectors with smaller sizes. Many Cas9 orthologues of smaller size have been characterized and applied in mammalian systems (Cebrian-Serrano & Davies, 2017), but the *Streptoccocus pyogenes* Cas9 remains the most widely characterized and used nuclease.

SpCas9 requires binding to a crRNA:tracrRNA (gRNA) or a chimeric single gRNA (sgRNA) molecule for the activation of DNA-interrogation activity. By fusing the 3' end of the crRNA to the 5' end of the tracrRNA, the chimeric RNA retains full activity with the advantage of being produced by a single transcript (Jinek et al., 2012). Structural data demonstrated that the sgRNA is a key regulator of SpCas9 activity, with major structural changes upon the complex formation necessary for DNA binding activity (Jiang & Doudna, 2017; Jinek et al., 2014). The sgRNA molecule starts at the 5' prime with a 20-nucleotide region named *spacer* that is complementary to the target sequence of interest (Figure 1.3C). The 2D structure of the sgRNA is composed of three Stem Loops, where Stem loop 1 and Stem loop 2 are connected by a linker domain (Figure 1.3C).

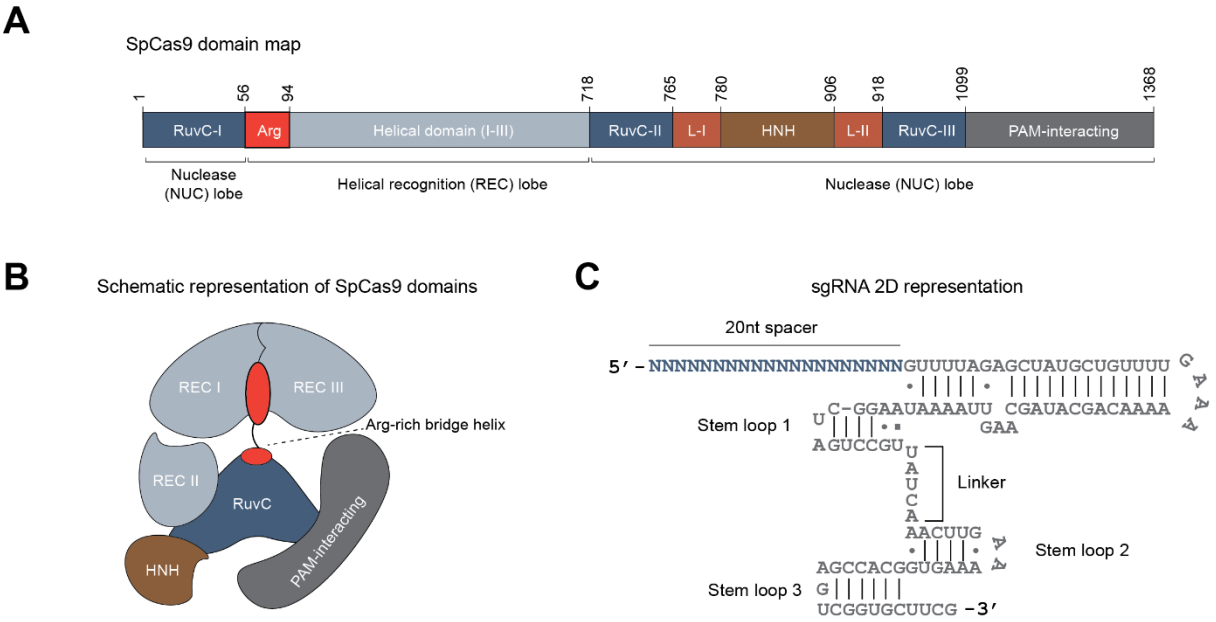


Figure 1.3. Cas9 protein domains and sgRNA representation. (A) SpCas9 domain map. (B) Schematic representation of SpCas9 domains. (C) Two-dimensional representation of a sgRNA,

1.3.4.Cas9 mechanism of action

The Cas9 protein undergoes a series of discrete steps from activation until DNA cleavage. In the apo (inactive) state (Figure 1.4A), Cas9 cannot bind to DNA and requires a crRNA:tracrRNA or sgRNA binding for DNA recognition (Figure 1.4B). The enzyme makes contacts with the Stem Loop 1 of the sgRNA, and this region is indispensable for Cas9 activity (Jiang & Doudna, 2017). Target recognition requires complementary base pairing between the crRNA and the DNA sequence, as well as a 5'-NGG-3' PAM sequence. Cas9 initiates target DNA search by probing for a PAM sequence before interrogating the crRNA:DNA complementarity. If no PAM is found, Cas9 rapidly dissociates from the DNA molecule (Figure 1.4C). Upon PAM recognition, Cas9 triggers local DNA melting followed by the formation of the RNA invasion and formation of the R-loop (crRNA:DNA hybrid) from the PAM-proximal to the PAM-distal side of the target (Figure 1.4D-F). Mismatches between crRNA:DNA in the seed portion (positions 10-20) abrogate R-loop formation (Ivanov et al., 2020), but mismatches at PAM-distal regions are accepted (Jr et al., 2021).

Upon PAM recognition and R-loop formation, Cas9 is activated for target DNA cleavage (Figure 1.4G). The HNH domain cleaves the TSS between the 3rd and the 4th bases from the PAM, while the three split RuvC domain cleaves the NTS mostly at the same positions, generating a blunt DNA DSB. Staggered cleavage has been reported, with 5' ssDNA overhangs at the cut site mediated by the RuvC domain (Chauhan et al., 2023; Jinek et al., 2014; Jr et al., 2021; Lemos et al., 2018; Shou et al., 2018; Slaman et al., 2023). After cleavage, Cas9 stays bound to the DSB site and releases the TS on the PAM-distal side of the break (Wienert et al., 2020).

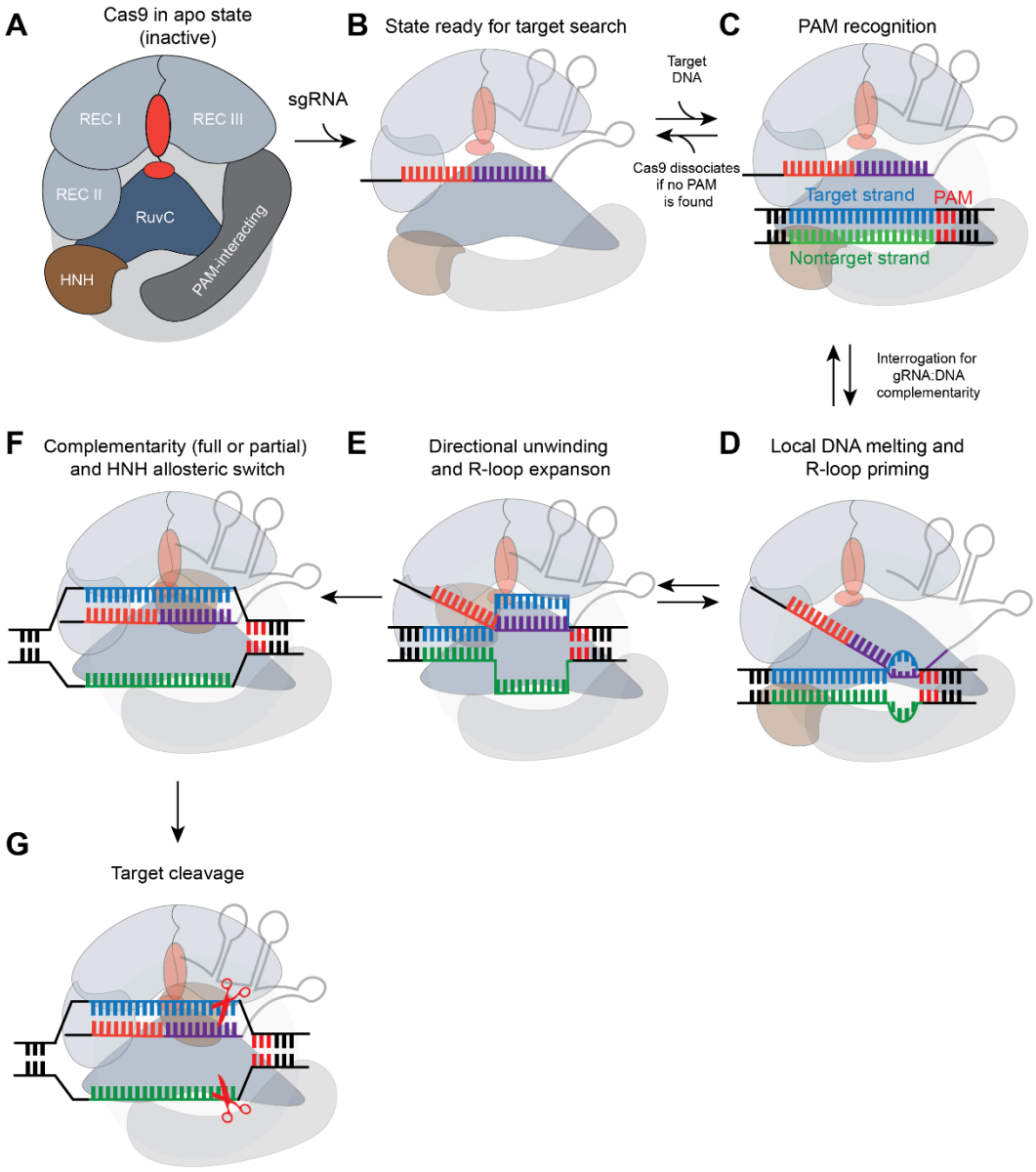


Figure 1.4. Discrete steps of Cas9 R-loop formation and cleavage. (A) Representation of Cas9 in apo state (inactive). (B) Cas9 becomes active once loaded with a gRNA. (C) Cas9 recognizes a 5'-NGG-3' PAM sequence and (D) tests for gRNA complementarity. (E) If no mismatches are found, R-loop is expanded towards the 5' region of the spacer. (F) Enough gRNA:DNA complementarity activates target cleavage. (G) Cas9 cleaves both strands of a target DNA.

1.3.5.DSB-independent gene editing

Despite the explosive growth in the gene editing field owing to the discovery of the CRISPR-Cas9 system, safety is paramount for the translation of CRISPR-nuclease based therapy in the clinic (see 1.4). Nuclease-mediated gene editing relies on DNA cleavage and the DNA damage response in cells for the installment of mutations (see 1.5.2). Mutagenic pathways might be elicited for the

repair of DNA DSBs, and translocations, large deletions and *chromothripsis* have been reported as a consequence of Cas9 on- and off-target activity (Brunet & Jasin, 2018; Leibowitz et al., 2021; Stadtmauer et al., 2020). Therefore, a direct editing of a target locus without relying on DNA DSBs and repair pathways is a desirable strategy to mitigate gene editing genotoxicity.

Base editors (BEs) install precise and targeted point mutations without requiring DSBs or donor templates (Gaudelli et al., 2017; Komor et al., 2016). Two main classes of base editors have been developed: Cytosine Base Editors (CBEs) can convert C:G base pairs to T:A (Komor et al., 2016), and Adenine Base Editors (ABEs) can convert A:T pairings to G:C (Gaudelli et al., 2017). In BEs, the catalytically dead Cas9 (dCas9) localizes an ssDNA deaminase enzyme to a target sequence (Figure 1.5A). Upon dCas9 binding, the crRNA:DNA R-loop is formed, and the bases in the loop are modified. CBEs uses a cytidine deaminase that converts cytosines into uracils that are read by thymines during DNA replication. ABEs convert adenines into inosines that are read as guanines (Anzalone et al., 2020). For CBEs and ABEs the editing window are positions 4-8 in the protospacer.

Several modifications in the BE design have been done during the years to increase editing efficiency. For example, fusion of a Uracil Glycosylase Inhibitor (UGI) with a CBE increases editing efficiency by inhibiting base excision repair (Figure 1.5B). The use of an nCas9 instead of fully catalytically impaired Cas9 increases editing efficiency by nicking the non-deaminated strand, directing DNA repair to install the editing in the other strand. To date, ABEs and CBEs have been used in a variety of organisms, and clinical trials using a ABE targeting the PCSK9 gene in patients with atherosclerotic cardiovascular disease (ASCVD) (R. G. Lee et al., 2023), and is estimated that BEs can correct approximately 30% of annotated human pathogenic variants (Anzalone et al., 2020). One drawback of BEs is that only base conversions can be catalyzed, and other mutation classes such as insertions and deletions are not possible with this method.

Prime Editors (PEs) are a novel class of genome editor that can install edits of up to approximately 100bps as well, as all 6 possible base pair conversions (Anzalone et al., 2019). PEs are a fusion of nCas9 with an engineered reverse transcriptase domain (Figure 1.5C). An engineered Prime Editing gRNA (pegRNA) directs PE to the target sequence via base complementary, and encodes the desire edit in a RNA template (Figure 1.5C). Upon binding, the active RuvC domain of PE nicks the NTS, and the editor uses the 3' end of the pegRNA to prime a reverse transcription (RT) reaction (Figure 1.5D). The RT reaction uses the template with the desired DNA mutation encoded in the pegRNA, as well as homology to the target site to increase repair rates. After the RT reaction, the newly synthesized DNA forms a 3' DNA flap that is redundant with a 5' flap that contains the original WT sequence. DNA repair processes cleave off the 5' flap, favoring the incorporation of

the 3’ flap at the target site. This generates a heteroduplex where one strand is edited and the other is not. Permanent installation of the edit is completed by DNA repair, where the edited strand will be copied into the newly synthesized DNA, or by forcing DNA repair of the non-edited DNA strand with a nick (Figure 1.5D). Prime editing has tested in multiple cell lines, but with great variation in editing efficiency in cell-type and target-dependent manner (Anzalone et al., 2019; Koeppel et al., 2023).

Although PE is extremely versatile and shows great potential for installing nearly any mutation, the technology is still in its infancy, and further research is necessary to increase PE editing efficiencies. Despite bypassing the need for DNA DSBs, genotoxic in the form of off-target and chromosome rearrangements have been reported for BEs and PEs (Fiumara et al., 2023; Huang et al., 2024).

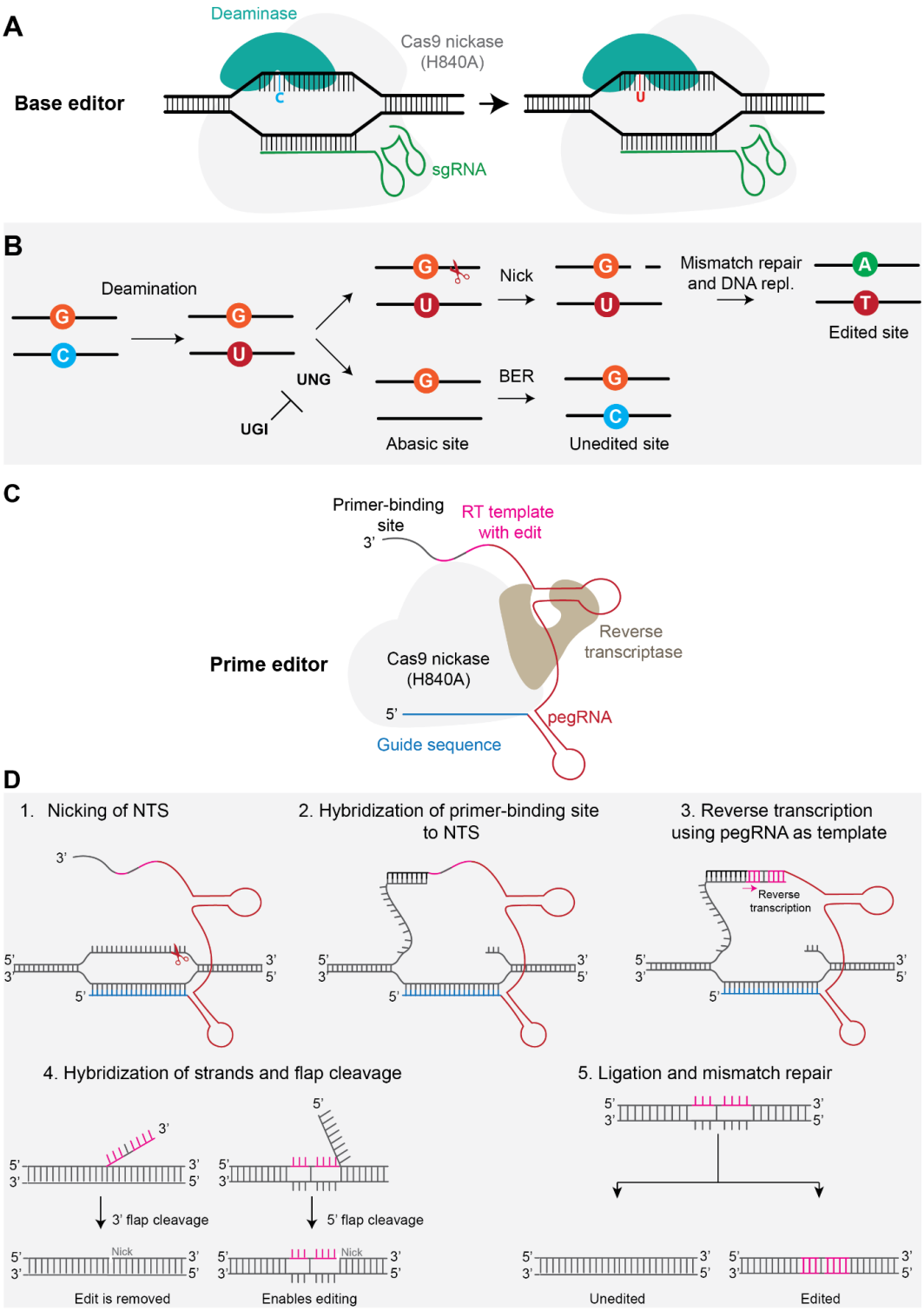


Figure 1.5. Schematic representation of base editors and primer editors. (A) An example of a base editor, composed of a Cas9 nickase fused with a deaminase. (B) Deamination reaction induced by a cytosine base editor as an example. The CBE converts G:C to A:T pairs. (C) Schematic representation of prime editor, composed of a Cas9 nickase fused with a reverse transcriptase. The pegRNA contains the guide sequence, the Cas9 scaffolding loops, a reverse transcription template containing the edit and a primer-binding site with homology to the region adjacent to the guide sequence at the 5' end. (D) Basic steps of a prime editing reaction.

1.4.Gene editing safety

Safety is paramount for the success of CRISPR-based therapy, and challenges need to be addressed before widespread clinical implementation of gene editing. One main concern is immunogenicity. Cas9 is a large protein from the *Streptococcus pyogenes* species that causes many infections in humans. It has been reported that the human immune system recognizes this protein as antigen, and therefore, SpCas9 can trigger an immunoresponse *in vivo* (Ewaisha & Anderson, 2023). Another major concern is the off-target effect. When applying the CRISPR-Cas9 in complex genomes, such as mammalians, the gRNA might bring the nuclease to regions that share some degree of homology to a given on-target. These off-target regions can be cleaved, and therefore an unintended mutation might arise. If the off-target region falls in tumor-suppressor genes, the mutation might give a fitness advantage to a population of cells by disrupting important cell proliferation control pathways, and cancer could arise from gene editing therapy. Moreover, because Cas9 generates DNA DSBs, illegitimate repair can generate gross rearrangements, such as translocations (Brunet & Jasin, 2018; Stadtmauer et al., 2020) and *chromothripsis*, a type of catastrophic chromosomal rearrangement that can initiate tumorigenesis (Leibowitz et al., 2021). Therefore, it is important to understand the molecular mechanisms underlying off-target activity, and ensure that sensitive tools available for Cas9 off-target study.

1.4.1.CRISPR off-target activity

Given the dangerous threat that off-target activity pose for the success of gene editing, it is paramount that tools for studying CRISPR-mediated cleavage are available to the scientific community. The evaluation of off-target activity becomes increasingly important as CRISPR moves closer to clinical translation. To this end, several methods were developed throughout the years with the goal of mapping the cleavage landscape of nucleases, allowing a genome-wide nomination of off-target activity. Off-target discovery is comprised of two steps: nomination and validation (Atkins et al., 2021). The nomination step is a broad survey of potential off-targets and can be done in three major formats: *in silico, in vitro* and *in cellulo.* Following the nomination step, putative off-targets are selected for validation typically via amplicon sequencing, in order to confirm the unintended edit of the off-target locus. Each off-target nomination approach has advantages and disadvantages and serve different purposes depending on the research question, and they build on they improve upon earlier methods in different capacities, such as throughput, feasibility or sensitivity.

1.4.2.*In silico* methods for predicting off-target cleavage

Computational, or *in silico*, methods predict off-target effect based on homology to a given target sequence. Early computational methods do not account for the complex cellular nuclear organization that can protect from off-target activity, usually providing users with a large list of putative off-target that requires extensive validation (Guo et al., 2023). A second group of *in silico* predictors devises machine learning strategies to predict on- and off-target. Algorithms can be trained on *in vitro* or *in cellulo* off-target nomination data (see 1.4.3 and 1.4.4) to predict cleavage, and outperform tools that predict off-targets purely by sequence homology (Sherkatghanad et al., 2023). Some of the widely used *in* silico tools are CHOPCHOP (Montague et al., 2014), Cas-OFFinder (Bae et al., 2014) and FlashFry (McKenna & Shendure, 2018), with different strategies to predict off-target activity (reviewed in Bao et al., 2021). Nonetheless, computation predictions are usually accompanied by orthogonal analysis with an *in vitro* or *in vivo* assay, as well validation of *bona fide* off-targets using amplicon sequencing.

1.4.3.*In vitro* methods for investigating off-target activity

Biochemical methods rely on the *in vitro* digestion of genomic DNA templates with RNPs, followed by sequencing of DSB ends. Because this mode eliminates the need of CRISPR delivery to living cells, *in vitro* methods are sensitive and provide a straightforward approach for off-target investigation (Figure 1.6). They rely on the digestion of deproteinated gDNA, and because there is no ongoing DNA repair, these methods are a sensitive tool for mapping nuclease cleavage activity (Atkins et al., 2021).

Digenome-seq was the first method to be developed and tailored to sequencing of nuclease-mediated DSBs (D. Kim et al., 2015). In the method, deproteinated gDNA is cleaved by RGENs (RNA-guided engineered nucleases) and the sample is submitted to deep whole-genome sequencing (WGS) to identify DSB ends. Digenome-seq lacks a DSB enrichment step, and putative off-targets are called based on blunt alignment of reads at the expected cut site over the staggered alignment of WGS reads (D. Kim et al., 2021). Because of the lack of enrichment of cleaved molecules before DNA sequencing, the method is impractical because of its lower sensitivity compared to newer tools and the need for WGS increases the relative cost per sample. CIRCLE-seq (Tsai et al., 2017) and SITE-seq (Cameron et al., 2017) build on Digenome-seq by introducing an enrichment of cleaved molecules step, increasing sensitivity to 0.1% (*i.e* can map off-targets that generate as low as 0.1% indel frequency) and reducing the number of reads needed per sample. CIRCLE-seq relies on the circularization of gDNA via a ligation reaction, forming circles that can be selectively linearized if they contain an off-target sequence of a given gRNA. The linearized molecules represent the off-target sequences, and can be selectively prepared for illumina sequencing (Lazzarotto et al., 2018; Tsai et al., 2017). This elegant strategy allows for

highly sensitive mapping of cleavage activity, but the circularization of gDNA fragments in CIRCLE-seq via a ligation reaction is not efficient and requires large amounts of starting material, precluding investigation of the off-target landscape in rare populations of cells or when limited amounts of material are available (Lazzarotto et al., 2018; Tsai et al., 2017). On the other hand, SITE-seq relies on the *in vitro* digestion of high molecular DNA, followed by the selective enrichment of DSB ends by ligation of a biotinylated adapter at the free DSB ends (Cameron et al., 2017). Although technically straightforward, the method is low throughput and requires large amounts of starting material (Cameron et al., 2017).

As CRISPR-Cas9 gene editing becomes widely accessible in laboratories, as well as in the form of therapies, off-target nomination and nuclease activity becomes routine in CRISPR-based biology. Therefore, it is important that such nominating tools allow scalability, reducing the experiment cost and throughput. CIRCLE-seq and SITE-seq cannot be performed in a high-throughput manner, precluding massive investigation of off-target effect. Building on CIRCLE-seq, CHANGE-seq (Lazzarotto et al., 2020) was developed with the goal of eliminating time-consuming steps in the CIRCLE-seq workflow, such as the DNA fragmentation via sonication and the circularization of DNA via ligation (Lazzarotto et al., 2018; Tsai et al., 2017). CHANGE-seq uses a Transposase-assisted strategy for DNA circularization, reducing the amount of input material by 10-fold and the number of enzymatic reactions in the workflow, allowing increase in scalability and throughput (Lazzarotto et al., 2020). Of note, Digenome-seq, CIRCLE-seq, SITE-seq and CHANGE-seq are inheritably biased towards blunt cuts, and cannot access the off-target landscape of staggered-cleaving nucleases such as Cas12a (Zetsche et al., 2015), and Cas9 staggered breaks.

1.4.4.*In cellulo* methods for off-target nomination

Although *in vitro* methods provide a sensitive map of nuclease activity, they all share one common drawback: a high "false positive" rate. Because the *in vitro* digestions are carried in deproteinated gDNA, no histones or 3D organization are present during gDNA digestion with the RNPs. It has been shown that chromatin accessibility and DNA folding can protect from off-target activity, where low accessibility reduce Cas9 cleavage (Lazzarotto et al., 2020) and a nucleosome positioned at the target site can impede Cas9 cleavage (Horlbeck et al., 2016; Yarrington et al., 2018). Therefore, mapping the off-target landscape with *in cellulo* methods in a cell physiological context provides a refined list of sites that can be cleaved when chromatin is present.

In cellulo tools are based on the delivery of CRISPR to the living cells, followed by the identification of cleaved regions (Figure 1.6). BLESS and BLISS are unbiased methods for the detection of DNA DSBs (Crosetto et al., 2013; Gothe et al., 2019; Yan et al., 2017) that can be applied to the

identification of CRISPR off-targets. BLISS relies on the labeling of unrepaired of DSB end with DNA adapters containing a T7 promoter sequence, in fixed cells. DSBs are then enriched via an *in vitro* transcription reaction, and aRNA products are selected for library preparation (Bouwman et al., 2020). INDUCE-seq leverages a similar strategy of labeling DSB ends with via a ligation reaction, but the enrichment of DNA breaks occurs in the sequencing flow, increasing sensitivity and quantification power (Dobbs et al., 2022). BLESS employs a biotin-streptavidin pulldown of DNA DSB ends (Crosetto et al., 2013). The methods are highly sensitive and have been applied to different types of DSBs (Biernacka et al., 2018; Crosetto et al., 2013; Dziubańska-Kusibab et al., 2020; Gothe et al., 2019; Mosler et al., 2021), but because these tools map unrepaired DNA breaks at the time of fixation, off-targets that are repaired with a fast kinetic might be missed, since living cells actively repair DSBs. Another method, DISCOVER-seq (Wienert et al., 2020), uses ChIP-seq for the repair factor MRE11 as a surrogate for sites of DSBs, but suffers from similar drawbacks as MRE11 only binds to DSBs primed for repair and therefore provides a snapshot of unrepaired breaks at the time of harvesting.

GUIDE-seq is a NGS-method that allows off-target nomination via tagging of DSB ends (Malinin et al., 2021; Tsai et al., 2015). The method relies on the co-delivery of the CRISPR system along with a short dsDNA donor molecule (dsODN) that is integrated at the DSB cleavage site upon sealing of DNA ends, and the tag serves as a PCR handle for library preparation. TTISS-seq (Schmid-Burgk et al., 2020) builds on GUIDE-seq by adding a Tn5-assisted tagmentation for library preparation, making the sample processing more streamlined. Both methods are extremely sensitive (~0.03%) and have an excellent signal-to-noise ratio because the integration of the dsODN into the DSB site is extremely precise (Cromer et al., 2023). Moreover, integration-based methods provide a recording of the entire cleavage as the cells are exposed with the donor tag for ~3 days - enough time to saturate the potential cleavage sites as opposed to mapping unrepaired DSB ends. One potential drawback is that the conditions for dsODN delivery must be optimized on a cell-type basis, and toxicity of the short donor sequence to stem cells has been reported (Wienert et al., 2020).

Given the different set of pros and cons for each method, combining the strengths of both *in vitro* and *in cellulo* methods is an advantageous strategy for nominating putative off-targets before the validation step.

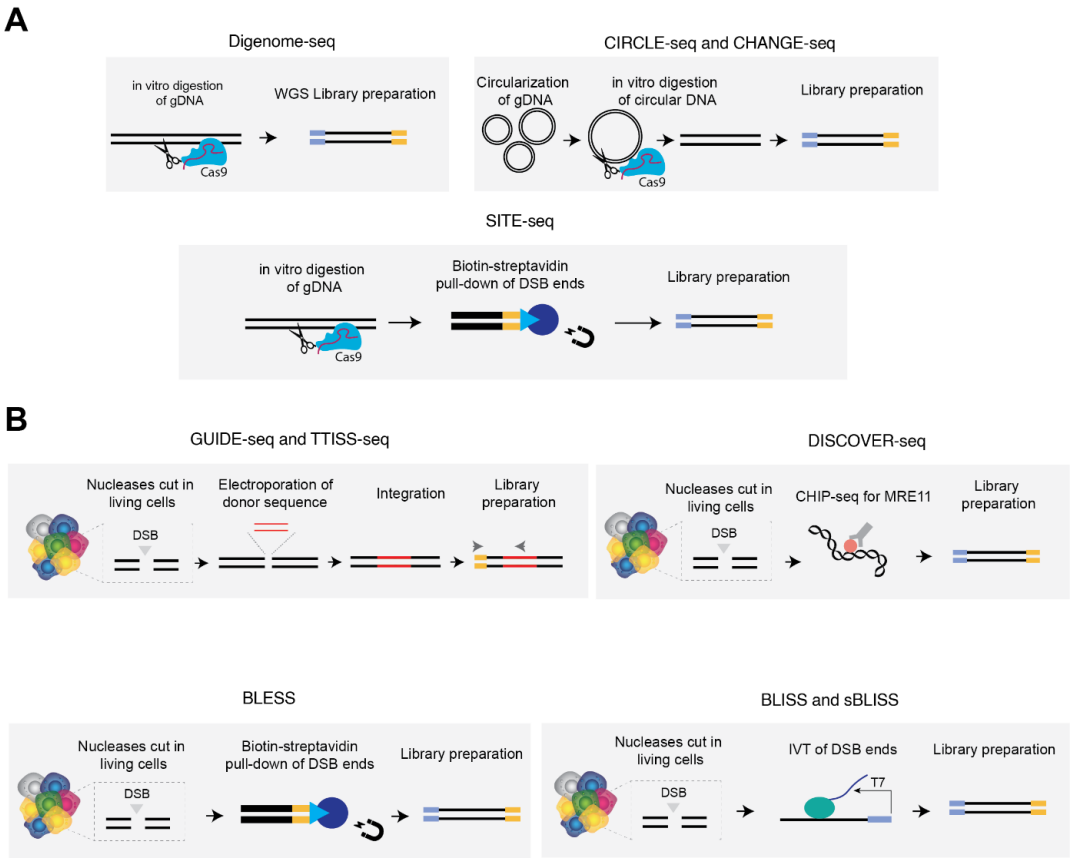


Figure 1.6. Schematic representation of off-target nominating tools. (A) DNA DSB enrichment strategies of most commonly used *in vitro* methods for off-target nomination. (B) DNA DSB enrichment strategies of *in cellulo* methods.

1.4.5.Improving CRISPR fidelity

Several strategies have been developed to mitigate CRISPR-Cas9 off-target activity, ranging from engineering of nuclease variants to gRNA modifications. In this context, *in vitro* and *ex vivo* off-target nominating tools are attractive methods to characterize and accelerate the discovery of strategies for the mitigation of promiscuity.

It has been reported that the gRNA sequence is an important factor for target specificity. Highly repetitive gRNAs are more likely to find semi-complementary homologous regions in the human genome, and increasing gRNA sequence complexity is an important strategy to mitigate off-target effect (Lazzarotto et al., 2020). Incorporation of 2'-O-methyl-3'-phosphonoacetate at the 5' of the gRNA has been reported to cause 40- to 120-fold reduction in off-target effect, without affecting on-target cleavage (Naeem et al., 2020). Another study reported that chimeric gRNA sequences with DNA bases induced 100-fold lower off-target activity while retaining on-target levels

(Donohoue et al., 2021). The addition of 2'-deoxyrubonucleotides in the gRNA molecule caused distortions in the guide-target complex, disfavoring binding at mismatched loci.

Given the relative gRNA-dependent SpCas9 promiscuity, several groups sought to develop improved Cas9 variants with higher fidelity. Engineering approaches range from a rational design, where one or few mutations are introduced in the SpCas9 backbone (Vakulskas et al., 2018), randomizing screens, where portions of the SpCas9 are mutated using error-prone PCR and selection of variants (Casini et al., 2018), or continuous evolution-based approaches that automate protein evolution (Hu et al., 2018). The overarching goal is to reduce crRNA:gDNA mismatch tolerance, allowing a better discrimination of on- from off-target sites.

Using a rational design approach, a study identified that the REC3 domain mediated mismatched target recognition, and that clustered mutations in this region reduced off-target activity without affecting on-target efficiency. The new variant was named HypaCas9, and showed a significant reduction of mismatched-substrate cleavage *in vitro* and *in cellulo* compared to SpCas9 (J. S. Chen et al., 2017). A similar rational approach was used in the design and discovery of LZ3, a high fidelity variant (Schmid-Burgk et al., 2020). Another study devised a positive selection screen using a library of ~250,000 variants in *E. coli.* The variants with strong on-target activity were able to cleave a toxic gene, ablating its expression. An off-target sequence was placed in an antibiotic resistance marker, and promiscuous variants were selected out due to cleavage of the resistance gene. The authors identified HiFiCas9, a high fidelity variant with a single point mutation R691A that significantly decreased off-target cleavage (Vakulskas et al., 2018). A similar directed evolution approach in *E. coli* was used for the discovery of the high fidelity variant SniperCas9 (J. K. Lee et al., 2018). Another selection strategy in yeast used error-prone PCR to introduce random mutations in the REC3 domain. Selection was carried out with reporters for on- and off-target activity, and after a single round of directed evolution the authors discovered EvoCas9 containing four amino acid substitutions in the REC3 domain (Casini et al., 2018). Finally, the PAM-flexible variant xCas9 was developed using Phage Assisted Continuous Evolution (PACE). The elegant method leverages the rapid bacteriophage cycle to evolve a protein of interest. The authors designed a circuit that favored selection of Cas9 variants with expanded PAM compatibility along lower off-target cleavage. The xCas9 recognizes an NG PAM instead of the canonical NGG, expanding the scope of Cas9 targetability (Hu et al., 2018).

Despite the growing body of work on novel engineered Cas variants one common drawback of introducing mutations in SpCas9 is a global reduction in cleavage activity of engineered variants (Kulcsár et al., 2022; Schmid-Burgk et al., 2020; Vakulskas et al., 2018). Therefore, methods that

facilitate the characterization of engineered nucleases can accelerate the discovery and validation of new gene editors.

1.5. Repair of DNA DSBs

DNA double-strand breaks are highly deleterious assaults that must be counteracted to ensure cell homeostasis. DSBs can arise from endogenous sources such as transcription, replication or the action of topoisomerases. Exogenous sources of DNA DSBs include UV radiation, chemotherapeutics or CRISPR-Cas9 (Hoeijmakers, 2009). If not faithfully repaired, DNA DSBs can cause consequences that can range from small mutations to gross chromosome rearrangements. Regardless of the source of DNA breakage, cells have evolved complex mechanisms known as the DNA Damage Response (DDR) to ensure proper DNA replication and chromosome segregation during cellular division.

The eukaryotic DDR response can be conceptualized as a series of "decision trees" where each branch represents a repair pathway, and the nodes consist of points of transitions between the branches (Scully et al., 2019). The two major branches are canonical Non-Homologous End Joining (c-NHEJ) or Homologous Recombination (HR). The branch decision is influenced by the competition of factors of the different pathways for recruitment at DSBs, the DSB end structure (Scully et al., 2019), the transcriptional status of the locus (Aymard et al., 2014), the presence of chromatin marks (Aymard et al., 2017), cell cycle phase (Xue & Greene, 2021), accessibility of the locus (Lemaître & Soutoglou, 2014), radial nuclear positioning (Lemaître et al., 2014), DSB end structure (Scully et al., 2019), among other factors.

1.5.1. Major DNA DSB repair pathways

The cNHEJ is a fast repair pathway that often leads to small indels (insertions/deletions) at the DSB site upon repair. This pathway is active throughout the cell cycle ad DSB ends are promptly religated with minimum DNA end processing. cNHEJ is initiated by the binding of the ring-shaped Ku70/Ku80 dimer to the DSB ends, blocking DNA resection and serving as a scaffold to recruit downstream factors (Figure 1.7A). DNA-dependent protein kinase (DNA-PK) is recruited at the DSB end, further recruiting the DNA ligase IV-XRCC4-XLF complex to ligate the DSB ends (Figure 1.7). Further processing might be required if DNA DSB ends are not directly ligatable, for example by Artemis (Biehs et al., 2017), PNKP and DNA polymerases Polμ and Polλ (Chatterjee & Walker, 2017; Xue & Greene, 2021).

The Microhomology-Mediated End Joining (MMEJ) is dependent on a short-range resection of the DNA DSB by the MRN/CtIP complex, exposing short patches of homology around the DSB site (Figure 1.7B) (Sfeir & Symington, 2015). Regions of homology on each side of the break will anneal, and the 3' ssDNA flaps formed are cleaved off, generating a small deletion as a repair outcome (Figure 1.7C). The remaining gaps are filled-in and the ends are sealed by DNA ligases, completing the repair cycle (Sfeir & Symington, 2015). Although initially perceive as a backup pathway, recent evidence suggests that MMEJ is the pathway of choice for DNA breaks in mitosis (Brambati et al., 2023), and are responsible Cas9-induced deletions (Shen et al., 2018).

The Single-Strand Annealing (SSA) (Figure 1.7E) and Homologous Recombination (HR) pathways (Figure 1.7H) are stimulated if long-range resection by EXO1 or BLM-WRN-DNA2 occurs (Sfeir & Symington, 2015) (Figure 1.7D). In SSA, The long 3' ssDNA overhangs are coated by RAD52, promoting the annealing of homologous sequences. The long 3' flaps are removed by XPF-ERCC1, ultimately generating large deletions as a DNA repair product (Figure 1.7E) (Sfeir & Symington, 2015). Alternatively, providing the cells with a DNA donor template containing homology to the ssDNA portion of the resected DNA stimulates Homology-Directed Repair (HDR), and can be leveraged for the controlled insertion of novel sequences (Figure 1.7F).

HR is a template-dependent and mostly non-mutagenic repair pathway. In somatic cells, the activation of HR factors is strictly regulated by the cell cycle, restricted mainly to S and G2 phases in order to ensure that a sister chromatid is available for templated repair (Heyer et al., 2010). As an exception to this rule, activation of HR in G1 has been recently described exclusively in centromeres, despite the absence of a sister chromatid (Yilmaz et al., 2021). HR requires long-range resection to form 3' ssDNA overhangs that are coated by RPA (Figure 1.7G). RPA is replaced by RAD51 filaments that aligns and pairs the ssDNA with a homologous dsDNA sequence, for example, from a sister chromatid. Furthermore, HR ensures proper DNA replication by being the pathway of choice for stalled replication forks (Scully et al., 2019).

In the context of gene editing, accuracy and efficiency are directly influenced by the DNA repair pathway of choice, and modulating the DDR response afford an opportunity to favor specific repair outcomes.

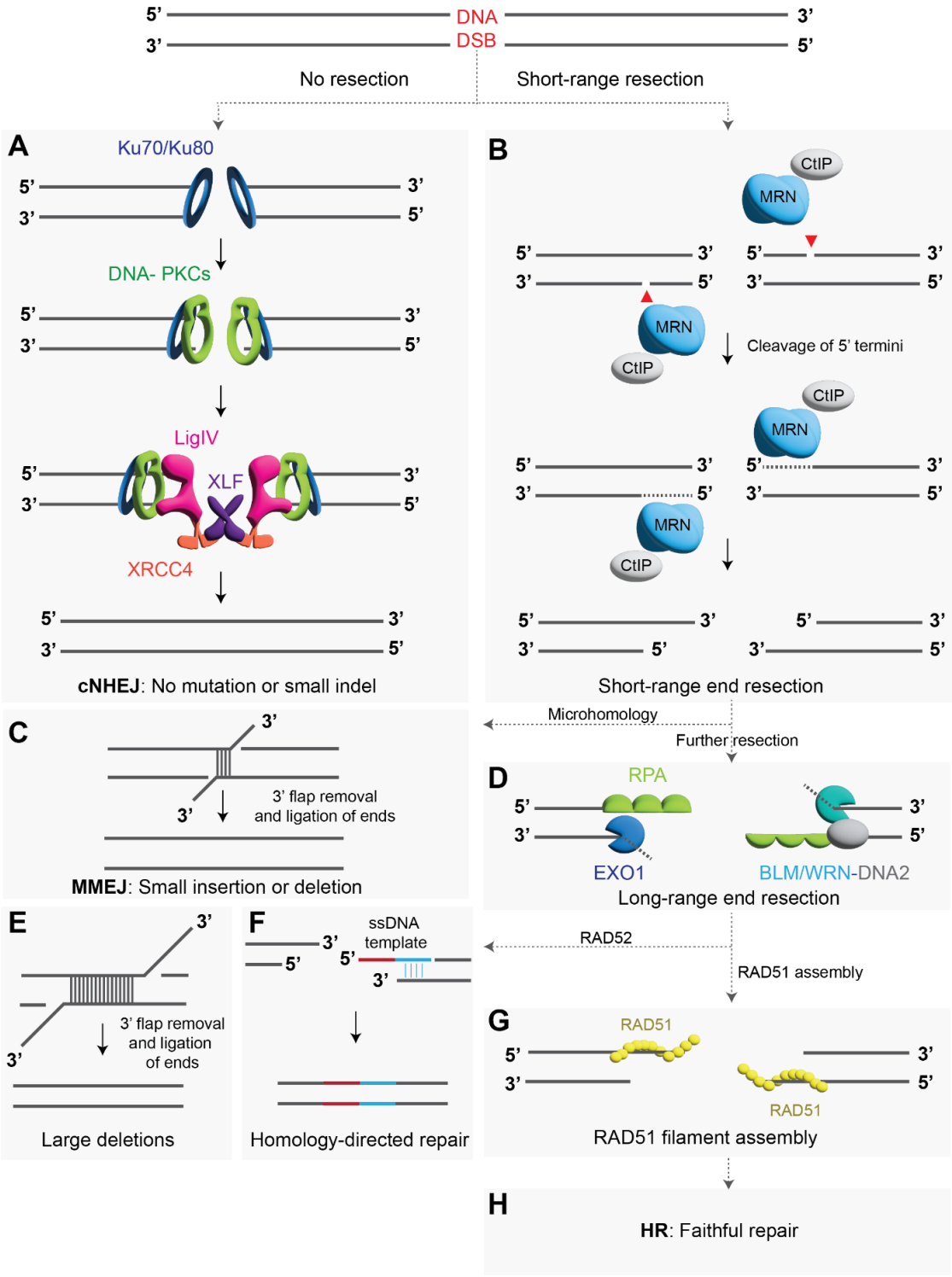


Figure 1.7. Major repair pathways in the eukaryotic DDR. (A) c-NHEJ ligates DSB ends with minimal processing. DNA-PKcs: DNA-dependent protein kinase catalytic subunit, LigIV: Ligase IV, XRCC4: X-ray cross complementing Group 4. XLF: XRCC4-like factor. (B) Short-range resection is mediated by MRN and CtIP. MRN: Mre11-RAD50-NBS1. CtIP: C-terminal binding protein interacting protein. (C) MMEJ uses homology surrounding the DSB site. (D) Long-range resection is necessary for SSA and HR. RPA: Replication Protein A. BLM: Bloom Helicase. WRN: Werner Helicase. EXO1: Exonuclease 1. (E) SSA mediates large deletions. (F) HDR is stimulated by a donor template containing homology arms to regions surround the DSB site. (G) RAD51 filaments coat the ssDNA portion of resected DNA. (H) HR promotes faithful repair of DSBs. Adapted from Xue and Greene, 2021.

1.5.2.DDR response and Cas9-induced DSBs

Like other types of DNA DSBs, cleavage by Cas9 induces a DDR response that is affected by several factors, including the chromatin context. The Cas9 cleavage activity itself is impacted by accessibility of the target sequence, with sites at open chromatin displaying higher indel frequencies (Lazzarotto et al., 2020; Schep et al., 2021). Moreover, a nucleosome positioned at the target site can impede cleavage by Cas9 (Horlbeck et al., 2016; Yarrington et al., 2018). The two major pathways utilized for template-free DNA repair are NHEJ and MMEJ, mediating small insertion and small deletions, respectively (Xue & Greene, 2021) (Figure 1.8B). However, the relative contribution of each pathway is affected by the chromatin context. A study devised a reporter assay named TRIP (Thousands of Reporters Integrated in Parallel) to investigate the contribution of DNA repair pathways across different chromatin domains (Schep et al., 2021). The reporter is a fixed target sequence endowed by a NGG PAM, and the contribution of MMEJ or NHEJ can be investigated by assessing the frequency of indel sizes. In this target sequence, single-nucleotide insertions are mediated by NHEJ, whereas 7bp deletions are mediated by MMEJ due to the presence of microhomology at the target sequencing (Brinkman et al., 2018). The reporter was integrated at over 1,000 random genomic locations with known chromatin context, and the integration sites were mapped providing a catalogue of different chromatin domains, Upon the DNA cleavage and repair, the relative contribution of MMEJ and NHEJ can be measured by the frequency of -7 and +1 indels. The authors found an over 5-fold variation in repair pathway usage across the chromatin domains, with MMEJ being preferably used at heterochromatin markers such as lamin associating domains, H3K9me2, H3K9me3 and late replicating regions. For open chromatin and actively transcribed regions, NHEJ was preferably elicited for DNA repair (Figure 1.8A) (Schep et al., 2021).

Template-based Cas9-mediated DSB repair is an advantageous strategy that allows installation of controlled mutations. By co-delivering CRISPR with a DNA donor template with homology arms to the target sequence, knock-in of sequences can be achieved. However, the proteins necessary for HDR are tightly cell cycle-regulate, and NHEJ or MMEJ are the pathway of choice in G1 cells (Xue & Greene, 2021). The donor template is the most important component of HDR, and can be designed as a plasmid, linear single-stranded or double-stranded DNA or synthetic oligonucleotides (Figure 1.8C). A common strategy for enhancing HDR is the chemical inhibition of NHEJ proteins, favoring HDR usage (M. Liu et al., 2019). Recently, a study reported that inhibition of MMEJ and NHEJ by chemical inhibition of DNA-PKcs, Polθ and RAD52 achieved HDR rates of up to 93% in human cells (Riesenberg et al., 2023).

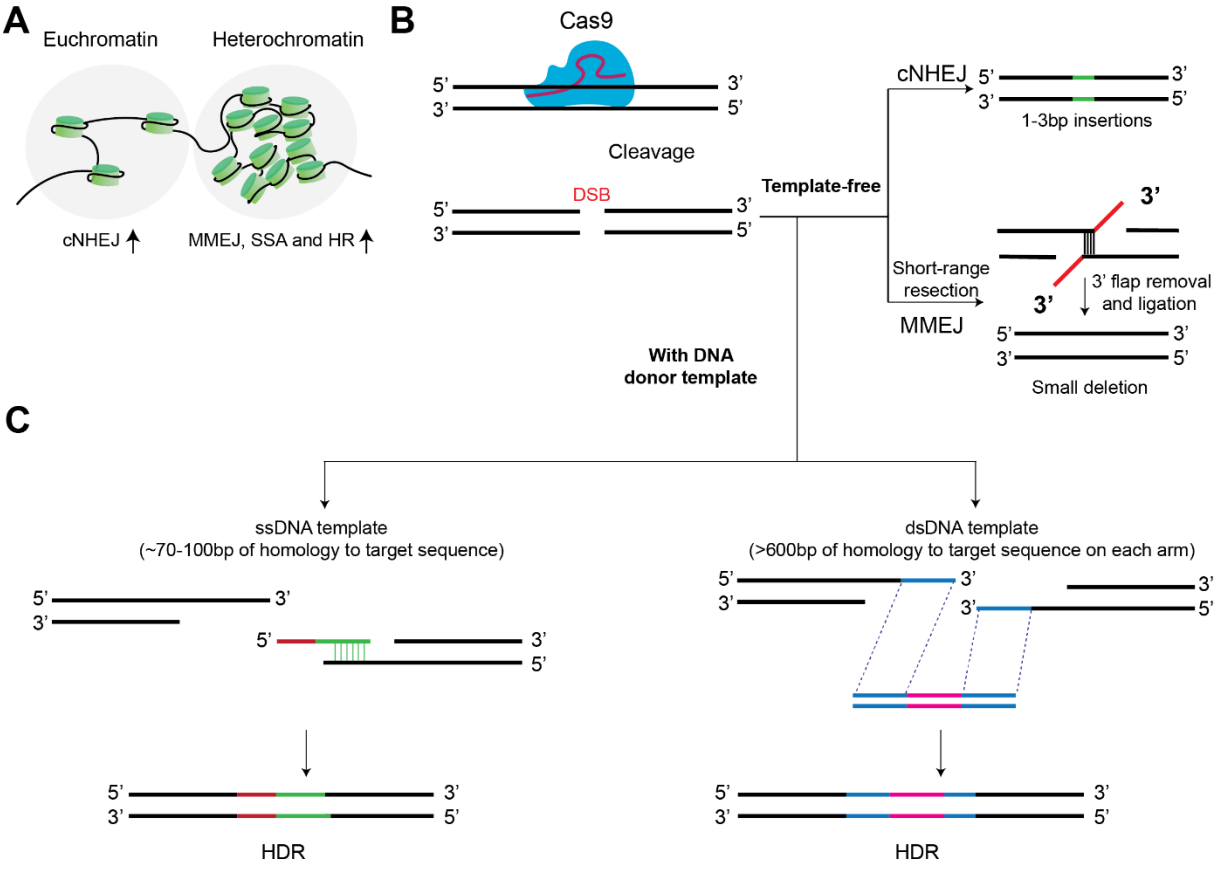


Figure 1.8. Schematics of Cas9-mediated DSB repair outcomes. (A) The effect of chromatin accessibility on DNA repair of choice. (B) Template-free repair outcome of Cas9-mediated DSBs. (C) Cas9-mediated DSB repair outcomes by HDR.

1.5.3.A link between DSB-end structure and predictable indels

The Cas9 nuclease contains two nuclease domains, HNH and RuvC, responsible for cleaving the target strand (TS) and nontarget strand (NTS), respectively. The seminal structural characterization of the nuclease demonstrated that the HNH domain precisely cleaves between the positions 18 and 17 (18|17) of the protospacer, while the RuvC cuts between the same bases and at additional downstream positions (Jinek et al., 2012). Nonetheless, Cas9 was thought to make nearly-blunt DSBs, and the staggered cleavage remains poorly characterized thus far.

Using molecular dynamics simulation, a study reported that cleavage of the NTS between 17|16 of the target sequence was more energetically favored than 18|17, generating 1 nucleotide 5' ssDNA overhangs (Zuo & Liu, 2016). It is worth pointing that the molecular dynamics simulation experiment utilized the initial coordinates of the Cas9-sgRNA complexed with a DNA target sequence from a crystal structure trapped in a pre-cleavage state. Therefore, the role of target DNA-protein interactions in staggered cleavage was not assessed. Using budding yeast, a

subsequent study revealed that Cas9 generated 1nt 5' overhangs in a gRNA-dependent manner (Lemos et al., 2018). Notably, the authors demonstrated that the 5' overhangs are filled in, and the product of DNA repair are *templated insertions,* where the 5' overhang is used as a template by *Pol4* for the repair reaction. The association between staggered cleavage and precise templated insertions was supported by additional studies in human cells (Shi et al., 2019; Shou et al., 2018), and a general model of occasional staggered cleavage mediating insertions was proposed (Gisler et al., 2019; Molla & Yang, 2020). Furthermore, it has been proposed that the relative frequency of blunt and staggered cleavages (herein, *scission profile*) is target-dependent (Molla & Yang, 2020), although direct evidence is not available.

The studies on the mutational profiles of Cas9 targets suggests that the indel landscape is target-dependent. By assessing the DNA repair profile of 223 human genome sites targeted with Cas9, Van overbeek et al., demonstrated that the Cas9-mediated DNA repair outcome is nonrandom and reproducible across independent experiments and cell lines (van Overbeek et al., 2016). The study showed that the relative frequency of insertions and deletions was target-dependent, indicating that the sequence context is a major determinant of the indel outcome. Building on this notion, Shen et al., constructed a gRNA-target report system to investigate the repair outcome of 2,000 Cas9 gRNAs and trained inDelphi, a machine learning model for predicting the Cas9-mediated indels (Shen et al., 2018). Using inDelphi, the authors demonstrated that template-free Cas9 editing is predictable and can be capitalized for the correction of pathogenic alleles. Although precision and predictability levels varied between the target sites, the authors estimated that 5-11% of SpCas9 gRNAs in the human genome would induce a single predictable repair genotype in >50% of the repair products (Shen et al., 2018). Additional large-scale studies supported that machine learning models trained on Cas9-mediated indel repair outcome can accurately predict the repair landscape of target sequences (Allen et al., 2019; Chakrabarti et al., 2019; W. Chen et al., 2019; Leenay et al., 2019; Taheri-Ghahfarokhi et al., 2018).

Collectively, this body of work indicate that Cas9 edits are predictable and strongly determined by the target sequence context. The staggered cleavage mediates precise and predictable 1-3bp insertions, and microhomology around the Cas9 cleavage site mediates deletions. However, the frequencies and determinants of staggered cleavage remain poorly characterize for SpCas9 and engineered variants. Harnessing the scission profile might afford an opportunity to generate precise and predictable insertions.

1.6. CRISPR-Cas9 in the clinic

CRISPR-based gene editing technology has made tremendous progress since its initial discovery, fueling the development of novel gene editing therapies. Owing to the versatility and ease of use, CRISPR/Cas gene editing has been used in several clinical trials for blood disorders, cancers, immune system diseases and other genetic diseases (Zhang et al., 2023).

The first CRISPR-Cas9 in-human clinical trial aimed to test the safety and feasibility of the technology for cancer immunotherapy. In the clinical trial, two patients with advanced myeloma and one with metastatic sarcoma were enrolled to receive autologous CRISPR-engineered T cells (Stadtmauer et al., 2020). The T cells of patients were isolated and electroporated with Cas9 loaded with sgRNAs targeting two genes encoding the endogenous T cell receptor (TCR) chains - TCRα (TRAC) and TCRβ (TRBC). The PDCD1 receptor was also knocked out to improve the immune antitumor activity, followed by lentiviral transduction of the transgenic TCR. After validation of successful gene editing, modified T cells were infused back to the original donors. Interestingly, chromosomal structural variation in the form of large deletions and translocations were observed as a result of CRISPR-Cas9 gene editing, but the frequency of rearrangements decreased overtime, likely due to decreased cellular fitness from the unintended DNA repair products. No patients experienced cytokine release syndrome or other serious advert effects attributed to cell infusion. Furthermore, both patients with myeloma showed a reduction in the target antigens, and the clinical trial demonstrated that the promising CRISPR-Cas9 based therapy can be safe and feasible (Stadtmauer et al., 2020).

Following the report of the first *ex vivo* T-cell engineering using CRISPR, a landmark clinical trial demonstrated the incredible potential of CRISPR-Cas9 in curing Transthyretin amyloidosis (ATTR) in the form of an *in vivo* therapy. The disease is characterized by the misfolding of transthyretin (TTR), an enzyme normally produced in the liver. The misfolding protein aggregates into amyloid fibrils and accumulates in various tissues leading to irreversible tissue damage (Brailovsky et al., 2023). In the study, patients with hereditary ATTR were infused with lipid nanoparticles carrying the mRNA sequence for SpCas9 and a sgRNA targeting the exon 2 of TTR (NTLA-2001). Patients showed a dose-dependent response, and those that received the highest dose of the therapy had a durable mean reduction of 87% of baseline serum TTR concentration after a single round of therapy with minimal adverse effects (Gillmore Julian D. et al., 2021).

Transfusion-dependent β-thalassemia (TDT) and sickle cell disease (SCD) are life-threatening blood disorders characterized by mutations in the hemoglobin β subunit gene (HBB), hampering erythropoiesis (Frangoul et al., 2021). Patients experience moderate to severe pain likely cause by a vaso-occlusive crisis, due to obstruction of blood circulation causes by sickled red blood cells (Borhade & Kondamudi, 2024). Another groundbreaking clinical trial aimed to knock out the

BCL11A transcription factor that represses γ-globin expression in erythroid cells with the goal of upregulate the production of fetal hemoglobin. To achieve this, the CD34+ hematopoietic stem cells (HSCs) were electroporated with CRISPR-Cas9 targeting BCL11A. Patients underwent myeloablation and autologous modified CD34+ cells were infused (Frangoul et al., 2021). After over one year of the infusion, the patients maintained high levels of fetal hemoglobin and vaso-occlusive episodes were eliminated (Frangoul et al., 2021). Finally, in late 2023, the CRISPR-based therapy to treat sickle-cell disease and β-thalassemia were approved by the regulatory agencies of the United Kingdom and the United States of America (U.S. Food and Drug Administration, 2023; Wong, 2023). Taken together, CRISPR-based genome editors have sparked a revolution in the treatment of genetic diseases, emphasizing the critical need for extensive study of editor activity and safety.

Despite the growing success of CRISPR-based therapy, template-free gene editing has not yet been controlled at the level for high-precision in clinic applications. Important aspects of Cas9 nuclease and variants activity remain unexplored and under characterized owing to the lack of methods. The factors influencing the mutational landscape upon Cas9 cleavage have not been fully assessed. Despite being proposed as a regulator of the indel outcome, the direct association between scission profile and Cas9-induced DSBs remains untested at a considerable scale. Determinants of Cas9 scission decision remain elusive, and might afford an opportunity to increase gene editing precision. Finally, despite influencing the off-target landscape by changing Cas9 target sequences, it is not known whether human genetic variation can affect gene editing precision.

1.7.Aims

i. Development of a high-throughput method for mapping Cas-induced DSBs

We first aimed to develop a scalable and high throughput method for the profiling of CRISPR nuclease landscape. Ideally, the method should be easy to perform, have few enzymatic reactions and can be easily implemented in any laboratory with access to a DNA sequencer. We developed BreakTag and HiPlex BreakTag to facilitate the investigation of gRNA and CRISPR nuclease cleavage activity. Moreover, in collaboration with the IMB Bioinformatics Core Facility, we developed an analysis pipeline, BreakInspectoR, to help users analyze BreakTag data and characterize the activity of novel Cas nucleases.

ii. Characterization of Cas9 scission profile

The frequencies of the different cleavage patterns generated by SpCas9 is currently unknown owing to the lack of tools that allow systematic investigation of DSB ends. We designed BreakTag to map blunt and staggered DNA breaks generated by SpCas9 and adapted BreakInspectoR to report the scission profile of gRNAs and Cas9 variants. We developed XGScission, an XGBoost random forest classifier, in collaboration with the IMB Bioinformatics Core Facility to investigate important variables to predict the scission profile of a given target site. Moreover, we demonstrated that XGScission can be applied to characterize sequence determinants of staggered cleavage of different Cas9 variants.

iii. Test at scale the association between scission profile and indel landscape

The indel landscape of CRISPR gene editing is target dependent, and the favored repair outcome can be predicted based on the target sequence alone. We aimed to test at scale the role of scission profile in the indel outcome of target sequences by investigating the DNA repair outcomes in a scission-dependent manner.

iv. Investigate the role of common human genetic variation in gene editing precision

Common human genetic variation has been shown to affect the off-target landscape of CRISPR-based gene editing by introducing new PAMs. We sought out to investigate the role of single-nucleotide polymorphisms on scission profile and the indel landscape at polymorphic loci. We used a cloning strategy to assess the indel repair landscape at hundreds of loci using a reporter-based assay, along with an endogenous strategy by targeting polymorphic loci in immortalized cells from donors of different genetic backgrounds.

v. Characterization of scission profile of engineered Cas9 variants

Given the unique ability of BreakTag to retrace the DSB end structure of a given Cas9 target site, we sought out to characterize the scission profile of high fidelity variants developed by other laboratories. We collaborated with the IMB Protein Production Core Facility to produce these recombinant variants in house and performed BreakTag experiments to meet this goal.

vi. Assessment of the translational potential of scission-aware gRNA design

We aimed to test how the new-found knowledge on the role of scission profile of Cas9 and variants can be applied to increase gene editing precision. We set out to apply our machine learning model trained on the nucleases to assess the potential of a scission-aware gRNA designing for the correction of pathogenic single-nucleotide variants for different diseases.

Chapter 2. Results

2.1.BreakTag design and implementation

In order to investigate the off-target landscape of CRISPR gene editing and to characterize novel nucleases activity, we sought to develop a cost- and time-efficient method for mapping DNA double-strand breaks (DSBs) genome-wide. Our ideal method should be easy to perform and have minimal steps to reduce sample preparation prices and reagent requirements, streamlining sample processing.

We developed BreakTag (Patent WO2024033378A1), a minimized and highly scalable four-step protocol that maps free DNA double-strand breaks (DSB) genome-wide. This protocol has been optimized for site-specific breaks, such as those by genome editors. The procedure begins with a blunting step, in which 5' overhangs are filled in, and 3' overhangs are resected, followed by A-tailing, where a single adenine is added to the 3' end of the DSB prior to labeling. Processed ends are then labeled with a customized BreakTag linker via a ligation reaction. This linker contains a PCR handle, a sequencing primer binding site (mosaic end, ME), a unique molecular identifier (UMI) for removing PCR duplicates, and a sample barcode. The sample barcode, embedded in the linker, provides an extra layer of barcoding and increases throughput, allowing samples to be pooled and further processed in the same tube if necessary. After ligation with the BreakTag linker, gDNA is tagmented with a single-handle Tn5 containing a second PCR handle, which randomly cuts the DNA and inserts an adapter into the 5' end of the fragment (Picelli et al., 2014). Ligation of DSB ends with the BreakTag linker, followed by tagmentation with a single-handle Tn5, generates two populations of fragments: one called "*homotagged*", in which both ends of the fragment contain the same sequence added during tagmentation, and a second "*heterotagged*" population, in which fragments contain the BreakTag linker at the proximal end and the tagmentation linker at the distal end (Figure 2.1A). The latter are amenable to exponential amplification, as they contain two distinct PCR handles for primers that introduce functional p5 and p7 sequences. Homotagged fragments <1kb are not exponentially amplified, as they contain homologous sequences on each side of the fragment, forming intramolecular hairpins that avoid amplification. Moreover, these fragments do not cluster during sequencing with Illumina sequencers because they lack a p5 sequence. Fragments larger than 1kb can still amplify, but they are size-selected, as they are larger than BreakTag fragments (Figure 2.1B). Ready-to-sequence libraries are achieved in less than 6 hours, depending on the number of samples being handled,

with minimal hands-on time. The method is performed in multi-well plates, using a multichannel pipette, and automation with a liquid-handling platform is feasible.

Users can perform BreakTag in two modes: cell-based (*ex vivo*) and cell-free (*in vitro*). In the cell-based mode, CRISPR is delivered to living cells in the form of a plasmid or ribonucleoprotein (RNP), where the nuclease can cleave the target DNA sequence when the chromatin environment is intact, and DNA repair is present. Thus, BreakTag provides a snapshot of the unrepaired DSB landscape at the time of cell harvest. To demonstrate this, we delivered a photocleavable sgRNA to Cas9-expressing HEK293 cells, which allowed us to stop Cas9 cleavage and synchronize DSB repair upon 30 seconds of 365nm light delivery (Zou et al., 2021) (Figure 2.1C,D). Cell-based off-target discovery provides users with a reduced list of nominated putative sites, thereby reducing the number of loci for the validation step. However, since only a snapshot of the unrepaired DSB is assessed with DSB-labeling based methods, this mode is prone to false negatives.

In a cell-free format, genomic DNA (gDNA) is isolated, protease-treated and incubated with the RNP for digestion *in vitro* (Figure 2.1E). The biggest advantage of the cell-free format is that it can be performed in any organism as long as DNA can be extracted. Furthermore, since there is no ongoing repair and processing of DSBs, this mode allows single-nucleotide resolution of nuclease cleavage. To demonstrate this, we digested gDNA of U2OS cells with Cas9 loaded with the sgRNA targeting the FAS locus ("FAS site 1"). Cleaved sites appear as reads that rise above background and are enriched over a non-target control (Figure 2.1F). BreakTag reads are directional, with each side of the DSB mapping to opposite strands at the cleavage site (Figure 2.1F).

To demonstrate the scalability of BreakTag, we mapped the off-target landscape of 46 clinically-relevant loci (Lazzarotto et al., 2020) in single-mode targeted with Cas9, and investigated the off-targeted landscape of these gRNAs. We observed a wide range of off-targets nominated, with CXCR4 site 2 being the most specific gRNA in this dataset with 10 nominated off-target sites, and PDCD1 site 12 being the most promiscuous with 9326 sites (Figure 2.1G). This indicates that off-target activity is gRNA specific, and determinants might be explored to avoid promiscuity.

Another advantage of BreakTag over previous methods such as CHANGE-seq (Lazzarotto et al., 2020), SITE-seq (Cameron et al., 2017) and DIGENOME-seq (D. Kim et al., 2015) is the "end-repair" step that processes single-strand DNA (ssDNA) overhangs at the DSB site. Cas12a (former Cpf1a) is a Class 2 Type V compact CRISPR nuclease that can generate 5' DNA overhangs at the cleavage site (Swartz, 2019; Zetsche et al., 2015). To our knowledge, BreakTag is the only high-throughput cell-free method for both Cas12a and Cas9 off-target nomination thus far. To demonstrate this, we targeted the DNMT1 locus in the gDNA of U2OS with Cas12a and assessed the off-target landscape.

We observed a strong on-target cleavage at the expected site and at one off-target for the tested sgRNA (Figure 2.1F). In sum, BreakTag nominates the cleavage landscape of blunt and staggered CRISPR nucleases.

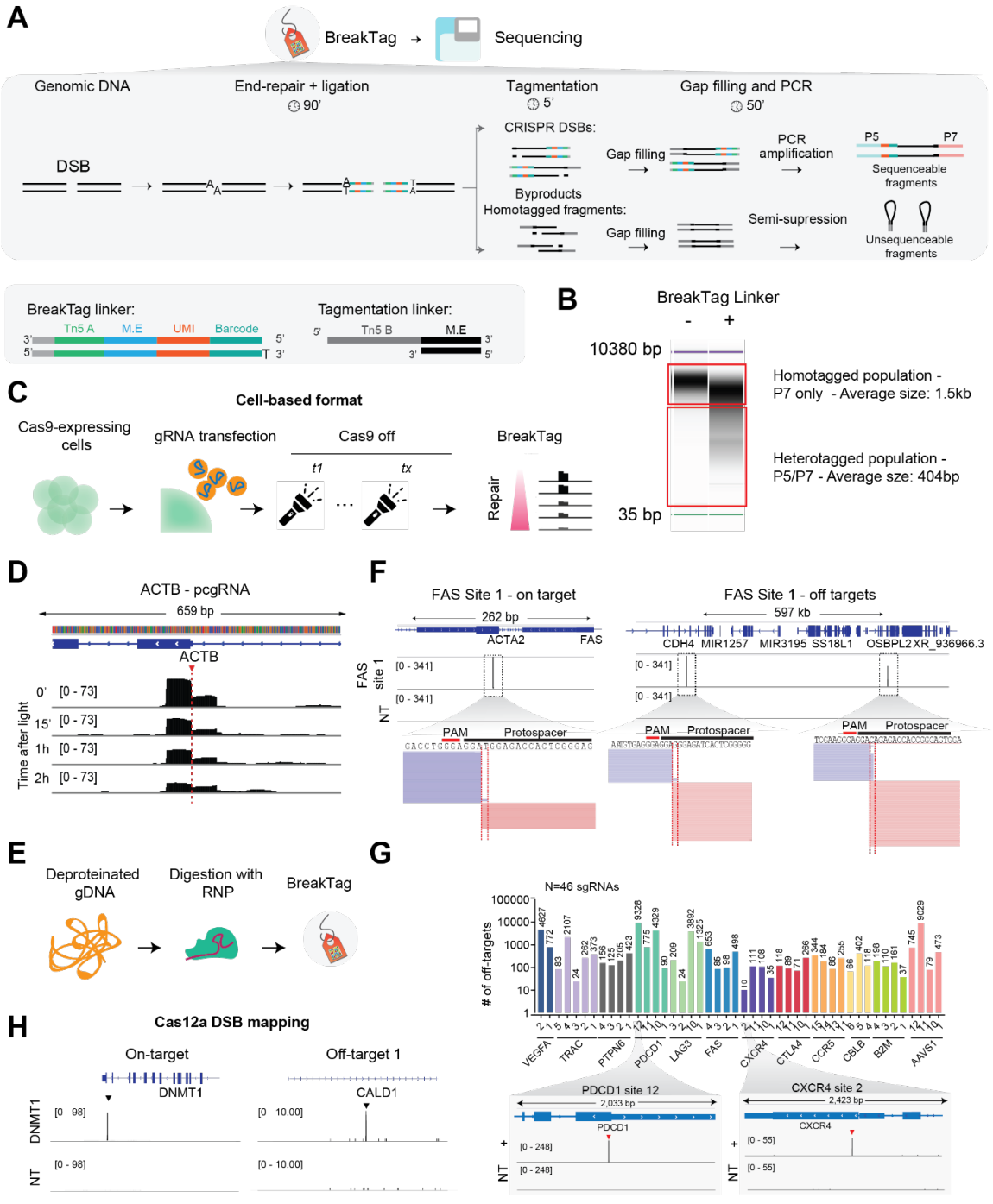


Figure 2.1. BreakTag profiles CRISPR on- and off-target DSBs. (A) Scheme depicting the biochemical workflow for BreakTag. (B) Bioanalyzer (DNA) trace showing BreakTag enrichment strategy. When BreakTag linker is omitted, small homotagged fragments form intramolecular hairpins and cannot amplify. Larger fragments can still amplify, but are not sequenced because they lack the p5 illumina sequence and are too large for clustering. When the BreakTag linker is added, smaller heterotagged fragments can amplify and are amenable for sequencing. (C) Scheme depicting BreakTag in cell-based format for repair kinetics investigation. Cas9-expressing cells are nucleofected with photocleavable sgRNAs. Cas9 nuclease activity

can be turned off by delivering light to cells. Cells are harvested over time and a snapshot of the unrepaired DSB landscape is obtained with BreakTag. (E) Cell-free BreakTag strategy. gDNA of any organism is isolated and *in vitro* digested with the CRISPR RNP. DSB-ends are then mapped with BreakTag. (F) Representative IGV snapshot showing processed BreakTag data of the on-target DSB of the "FAS site 1" gRNA (left) and two off-target sites (right). Zoomed-in views of the cut site (red dotted lines) and raw mapped reads (blue/pink rectangles) are shown below. NT: non-targeting control. gDNA from U2Os cells was used. (G) Number of off targets mapped by BreakTag in gDNA of U2OS cells digested with Cas9 and 46 different clinically relevant gRNAs (Lazzarotto et al., 2020). Representative IGV snapshots of the on-target region of "PDCD1 site 12" and "CXCR4 site 2" are shown below. (H) IGV snapshots depicting the on-target and one off-target of Cas12a targeting the DNMT1 locus in U2OS cells.

2.2.BreakTag maps Base Editor off-target activity

Current Base Editors (BEs) include a fusion of a base deaminase with nCas9 (nickase Cas9). Cytosine base editors (CBEs) catalyze C>T conversion, and adenine base editors (ABEs) catalyze A>G conversions (Anzalone et al., 2020; Gaudelli et al., 2017; Komor et al., 2016) (See 1.3.5). The nickase increases the editing efficiency by ensuring that the non-target strand copies the base edit upon repair of the nick. Although BEs are do not rely on the formation of DSBs, deamination of unintended regions have been reported and off-target has posed as a challenge in BE gene editing (Liang et al., 2019).

In order to adapt BreakTag for CBE off-target assessment, we leveraged a strategy where deaminated bases are removed using base excision repair (BER) proteins *in vitro* (D. Kim et al., 2017). USER is a mixture of E.coli uracil DNA glycosylase (UDG) and endonuclease VIII, where UDG removes deaminated cytosines (Uracil) from the DNA strand, generating an abasic site that is later nicked by endo VIII. The nick on the modified strand in combination with the nick generated by nCas9 activity on the opposite strand forms an artificial staggered DSB that can be mapped with BreakTag (Figure 2.2A), allowing investigation of BE off-target activity. To test this, we incubated the gDNA of U2OS cells with the cytosine base editor BE3 (Komor et al., 2016) loaded with a sgRNA targeting the VEGFA locus ("VEGFA site 2"). We observed a strong enrichment of artificial DSBs in the presence of USER at the on-target site and previously described off-targets (Figure 2.2B). Although not explored here, the same rationale can be applied for ABE off-target nomination by using deoxyinosine 3' endonuclease, or Endonuclease V (Liang et al., 2019). Therefore, BreakTag can be adapted for the investigation of base editor off-target activity.

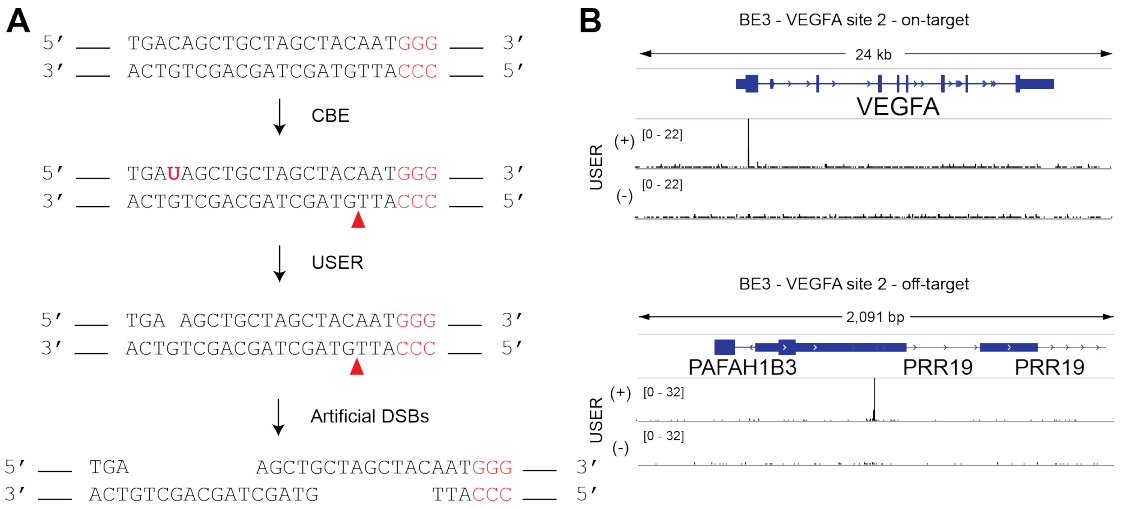


Figure 2.2. BreakTag maps off-target landscape of base editors. (A) Schematics of how artificial DSBs are generated at CBE target sites. The cytosine on the NTS is deaminated and converted to Uracil. The TS is cleaved by the HNH domain between positions 17 and 18 of the protospacer. USER is then applied to generate an abasic site followed by a nick at the modified site. The resulting artificial DSB contains 5' ssDNA overhangs that can be mapped with BreakTag. (B) IGV snapshot depicting an example of a base-edited site. BE3 (a CBE) was loaded with a sgRNA targeting the VEGFA locus ("VEGFA site 2") and incubated with gDNA of U2OS cells. The sample treated with USER shows an enrichment of the artificial DSBs compared to a sample where uracil excision was omitted.

2.3.BreakInspectoR development and implementation

To facilitate off-target discovery and the characterization of novel CRISPR nucleases, we developed BreakInspectoR, a bioinformatics pipeline for analyzing BreakTag data (https://github.com/roukoslab/breaktag). The pipeline processes raw demultiplexed FASTQ files and generates a BED file with coordinates containing DSBs. First, the pipeline filters for reads containing a sample barcode and a UMI. The clean reads are aligned to a reference genome of choice, the number of DSB per locus is counted using the UMI and a BED file is generated. Fragments that map at the same location in the genome and contain the same UMI sequence are considered PCR duplicates and are filtered out.

We developed an R package as part of BreakInspectoR for statistical analysis and plotting BreakTag data. The package uses the BED files containing the DSB coordinates as an input and identify loci with DSB read pile-up. We reasoned that using a Cas9 signature would reduce noise and increase sensitivity, so reads are filtered for those that are found in proximity of a PAM sequence (5'-NGG-3' for Cas9) and share homology to the target sequence. Finally, the pipeline tests the enrichment on the experimental sample over a non-target control to reduce false positives. Users can then filter for high-confidence off-targets using the empirical FDR score (Figure 2.3).

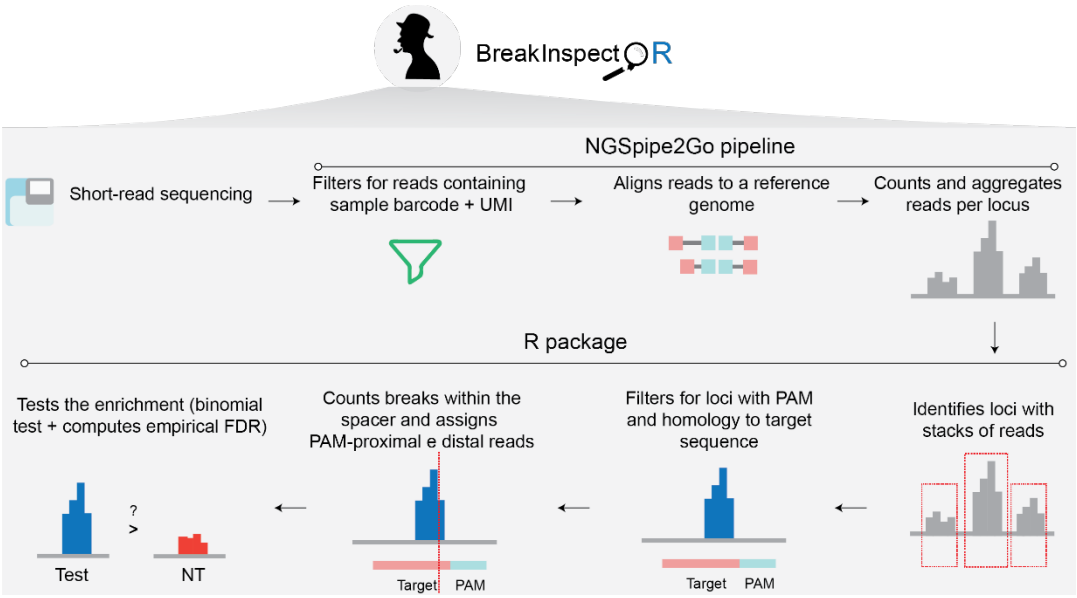


Figure 2.3. Scheme depicting BreakInspectoR workflow for analysis of BreakTag data. Initial analysis is performed using the BreakTag pipeline, where reads are filtered, duplicate reads are removed and aligned to a reference genome. BreakinspectoR identifies regions of increased signal with Cas9 signatures.

2.4.Benchmarking BreakTag against similar methods

A myriad of tools have been devised for mapping Cas9 off-target landscape, each with its own set of pros and cons (See 1.4). We tested the robustness of BreakTag by investigating the reproducibility and benchmarking the tools against previous off-target nominating methods: CHANGE-seq (Lazzarotto et al., 2020), SITE-seq (Cameron et al., 2017), Digenome-seq (D. Kim et al., 2015) and GUIDE-seq (Tsai et al., 2015) and TTISS-seq (Schmid-Burgk et al., 2020).

BreakTag showed an excellent correlation (R>=0.92) between two independent experiments when we investigated the reads per off-target (Figure 2.4A), demonstrating a robust reproducibility. Next, we sought out to compare the off-target landscape nominated by BreakTag with CIRCLE-seq (Tsai et al., 2017) and DIGENOME-seq (D. Kim et al., 2016) for which public data is available for FANCF, VEGFA_site_1 and VEGFA_site_2 targets. To confirm the editing of off-targets, we generated an amplicon deep sequencing dataset of the off-target loci nominated from the tested tools (i.e, *bona fide* off-targets). Our final list of sites comprises 1468 unique sites for VEGFA_site_ 2, 144 unique sites for VEGFA_site_1 and 12 unique sites for FANCF. Because the VEGFA_site_2 gRNA is extremely repetitive and promiscuous, validating all off-targets would be cost prohibitive. We therefore generated a TTISS-seq (Schmid-Burgk et al., 2020) dataset for VEGFA_site_2 to refine the call set and increase the chances of finding *bona fide* sites that were missed by CIRCLE-seq and Digenome-seq. TTISS-seq is an in cellulo method that builds on GUIDE-

seq (Tsai et al., 2015), and thus provides a refined list of sites available for cleavage in living cells (See 1.4 for the differences in off-target nominating strategies).

We next designed targeted amplicon sequencing panels consisting of unique and shared off-targets between the tested methods. For FANCF we selected all off-targets for the amplicon panel design. For VEGFA_site_1 and VEGFA_site_2 we arbitrarily selected sites based on the number of mismatches (1-6) and BreakTag read count (high, mid and low) and split them into 3 classes. Class I comprised targets with high BreakTag signal and 1-3MMs, Class II mid signal and 2-4MMs and finally, Class III of low signal and 1-6MMs. All TTISS-seq sites were selected for amplicon sequencing. All off-target sites nominated by TTISS-seq were selected for the amplicon panel.

All sites that contained indels for FANCF (n=2) and VEGFA site 1 (n=7) were nominated by all 3 methods, suggesting that using orthogonal *in vitro* methods is an interesting strategy for refining a superlist of off-targets for the validation step (Figure 2.4B). For the highly promiscuous VEGFA site 2, we detected 164 sites with indels, and 12 were nominated by BreakTag and TTISS-seq, but not the other methods (Figure 2.4B). The new sites identified by BreakTag can be explained by the additional end-repair step in our method that is able to probe staggered breaks (Figure 2.4B), which is absent in methodologies such as CIRCLE-seq, CHANGE-seq, SITE-seq and Digenome-seq. Nonetheless, we observed an excellent correlation between the number of sites nominated by BreakTag and CHANGE-seq (R=0.88) (Figure 2.4C)

In order to compare the performance of BreakTag with an established in *cellulo* assay, we analyzed the off-targets nominated by BreakTag and GUIDE-seq over 27 matching gRNAs (GUIDE-seq data from (Lazzarotto et al., 2020)). BreakTag nominated a larger list of off-targets (n=4394) compared to GUIDE-seq (n=240), as expected due to the nature of the assays (See 1.4.3 for the differences in off-target nominating strategies). A complete overlap between sites nominated by BreakTag and GUIDE-seq was found for 19/27 tested gRNAs (Figure 2.4D). Approximately 85% of all targets nominated by GUIDE-seq were also nominated by BreakTag across all tested gRNAs Figure 2.4E).Of note, we observed an excellent correlation between the number of off-targets nominated per gRNA for the tested methods (R=0.72, Figure 2.4D). It is worth noting that while GUIDE-seq was performed in human primary T cells from healthy donors with presumably a neutral genetic background 6, our BreakTag dataset was generated in osteosarcoma U2OS cells that contain a complex karyotype, including whole and partial chromosome losses (Janssen et al., 2011) that could account for the differences in sites nominated by the two methods. Furthermore, manual inspection of GUIDE-seq exclusive sites revealed that 6 out of the 37 are in regions identified in ENCODE which may have anomalous, unstructured, or high signal in NGS experiments (Amemiya et al., 2019).

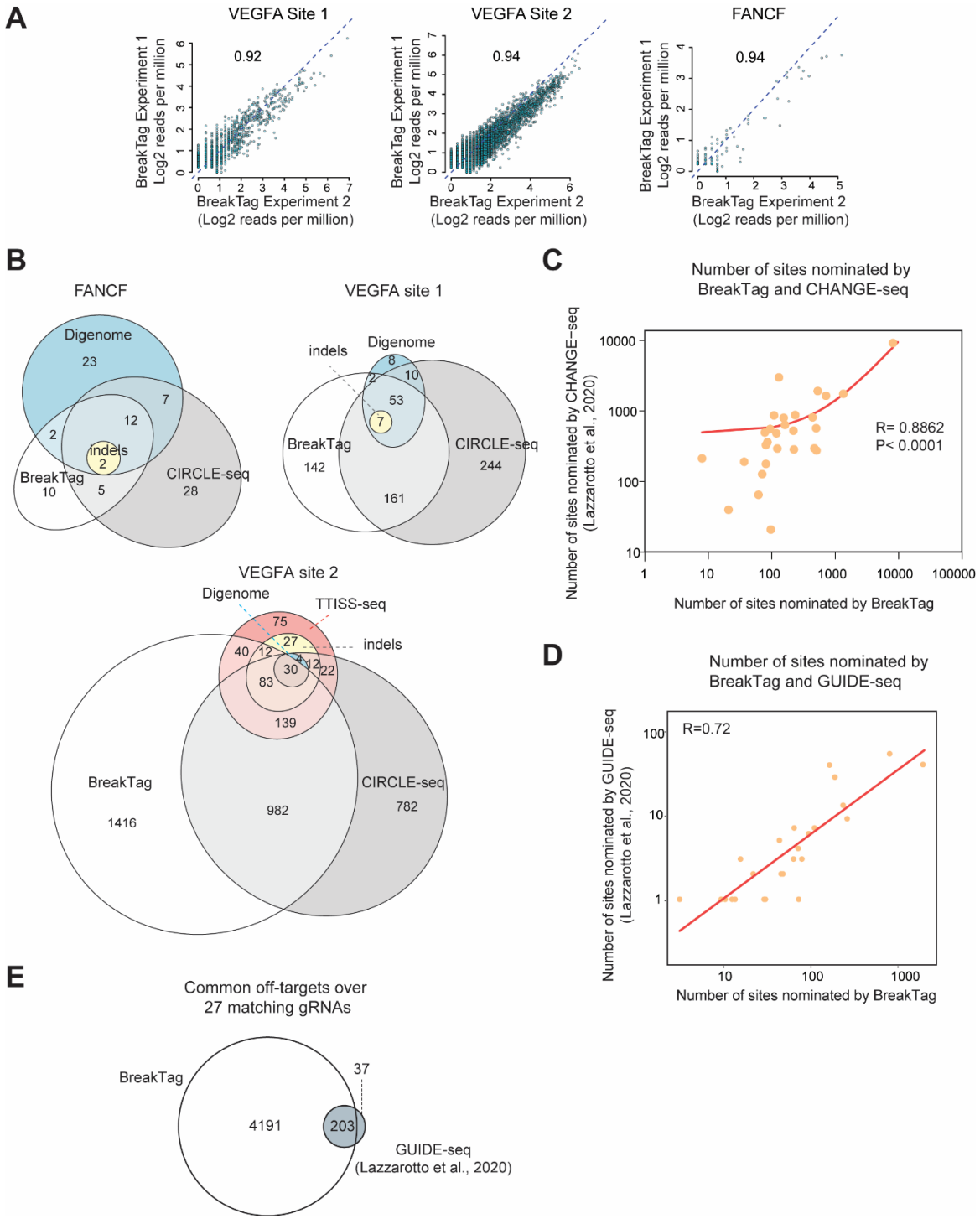


Figure 2.4. Benchmarking BreakTag against previous methods. (A) Correlation between two independent BreakTag runs for three sgRNAs commonly used in the benchmarking of off-target-nominating tools. (B) Venn diagrams showing the overlap between sites nominated by BreakTag, DIGENOME-seq (D. Kim et al., 2016), and CIRCLE-seq (Tsai et al., 2017). Off-targets were selected for validation using targeted deep sequencing. TTISS-seq (Schmid-Burgk et al., 2020) was used to generate a refined list of in cellulo VEGFA site 2 off-targets due to its high promiscuity. (C) Correlation between number of off-targets nominated by CHANGE-seq (Lazzarotto et al., 2020) and BreakTag over 44 gRNAs arbitrarily selected from the CHANGE-seq dataset. (D) Correlation between the number of off-targets nominated by BreakTag and

GUIDE-seq (Lazzarotto et al., 2020). (E) Common target sites identified by GUIDE-seq (Lazzarotto et al., 2020) and BreakTag over matching 27 gRNAs. For GUIDE-seq only targets supported by at least 8 reads, up to 6 mismatches between crRNA:DNA and an NGG PAM were considered; for BreakTag targets supported by at least 8 reads, up to 6 mismatches and a FDR<1% were considered.

2.5.HiPlex BreakTag multiplexes cell-free CRISPR off target assessment

We set out to increase the BreakTag throughput by multiplexing gDNA *in vitro* cleavage. We developed a companion approach named *HiPlex BreakTag* for the massive parallel investigation of CRISPR nuclease cleavage landscape.

The HiPlex mode combines pooling several target sequences in the same tube and high-throughput in house sgRNA synthesis, massively increasing the number of target sequences investigated per experiment (Figure 2.5A). First, targets are bioinformatically split into pools containing ~150 sequences each. A T7 promoter sequence is added at the 5' end of the target sequence, and a sequence containing homology to the SpCas9 sgRNA scaffold sequence is added to the 3' end. Because T7 *in vitro* transcription (IVT) is more efficient when the template molecule starts with a GG sequence, an extra G is added at the 5' end of the target, and the nucleotide found at position 1 of the protospacer is replaced by a G if this not a guanine (Figure 2.5B). The sequences are then ordered as oligonucleotide pools, and the IVT template is assembled via a PCR reaction using 3 primers (Figure 2.5B,C). The resulting template is used for IVT, where each reaction synthesizes a pool containing 150 sgRNAs each (Figure 2.5D).

We tested the feasibility of HiPlex BreakTag by generating a library of 1491 sequences targeting human genes (Chakrabarti et al., 2019) with Cas9 in gDNA of HepG2 cells. This procedure identified 92,375 cleaved sites (1418 of the 1491 on-targets were cut), validating the efficacy of the approach (Figure 2.5E,F).

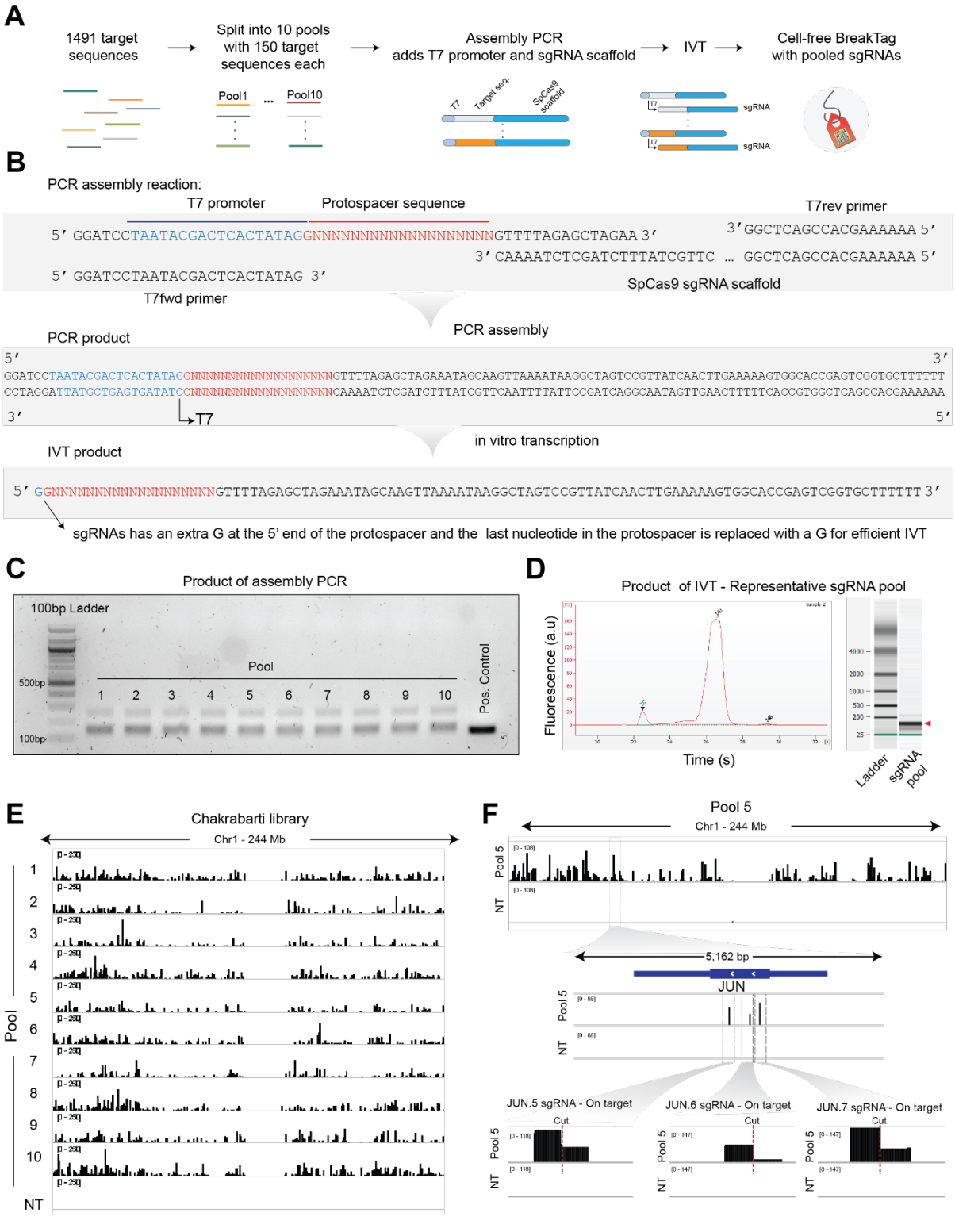


Figure 2.5. HiPlex BreakTag library construction strategy. (A) HiPlex BreakTag strategy. Cas9 target sequences are bioinformatically split into 10 pools, each containing ~150 sequences. A T7 promoter sequence was added to the 5' end of each sgRNA protospacer, and a Cas9 sgRNA scaffold sequence at the 3' end by a PCR assembly reaction, which generates a dsDNA template for T7 in vitro transcription (IVT). T7-transcribed sgRNAs were used for cell-free BreakTag. (B) PCR assembly strategy for IVT template

construction. (C) Agarose gel showing PCR product for HiPlex DNA oligo pools. (D) RNA nano BioAnalayzer of a representative sgRNA pool (Pool 1 HiPlex 1 library). Sharp peak represents pool containing 150 sgRNAs. (E) IGV snapshot of chromosome 1 of HepG2 cells digested with Pools 1–10 from the HiPlex1 library. Each bar represents a cleaved site. NT: non-targeting control. (F) IGV snapshot of chromosome 1, depicting cleaved sites for Pool 5 of the HiPlex1 dataset. Zoomed-in views of on-target DSBs of sgRNAs targeting the JUN gene are shown below.

2.6.Illuminating Cas9 nuclease activity

2.6.1.Determinants of sgRNA fidelity

Off-target activity poses as a threat for the success of gene editing efforts. Our robust HiPlex1 dataset prompted us to investigate determinants of Cas9 off-target activity. Previous studies reported that Cas9 is more permissive to crRNA:DNA mismatches (MMs) at PAM-distal regions of the target sequence (Ivanov et al., 2020; Jr et al., 2021; Lazzarotto et al., 2020). To complement previous efforts, we calculated the mismatch rate for each position along the protospacer sequence for the sgRNAs used in HiPlex1. Our analysis revealed that Cas9 is permissive to MMs if they are located at the 5' end of the protospacer, but fidelity increasingly progresses towards the PAM (Figure 2.6A). The seed portion (positions 11-20) was nearly completely absent from MMs. These data agrees with previous reports showing that MMs in the seed portion ablate R-loop formation, impeding Cas9 cleavage (Ivanov et al., 2020). Coupled with this notion, we identified an inverse correlation between the number of MMs and cleavage efficiency, indicating that mismatched target sequences are cleaved at a lower frequency (Figure 2.6B). We next asked if Cas9 PAM usage shows any degree of flexibility. We identified that ~85% of Cas9 DSBs were located next to the canonical NGG motif, but non-canonical PAMs NAG (9.29%) and NGA (1.75%) were also used, albeit with lower efficiency (Figure 2.6C).

To investigate the role of target nucleotide complexity in Cas9 off-target activity, we calculated the sequence diversity of each target sequence as the Shannon index (Lazzarotto et al., 2020). Interestingly, we observed that low-entropy target sites (i.e, repetitive loci) displayed a larger number of off-target, indicating that avoiding low complexity regions is a feasible strategy for off-target effect mitigation (Figure 2.6D).

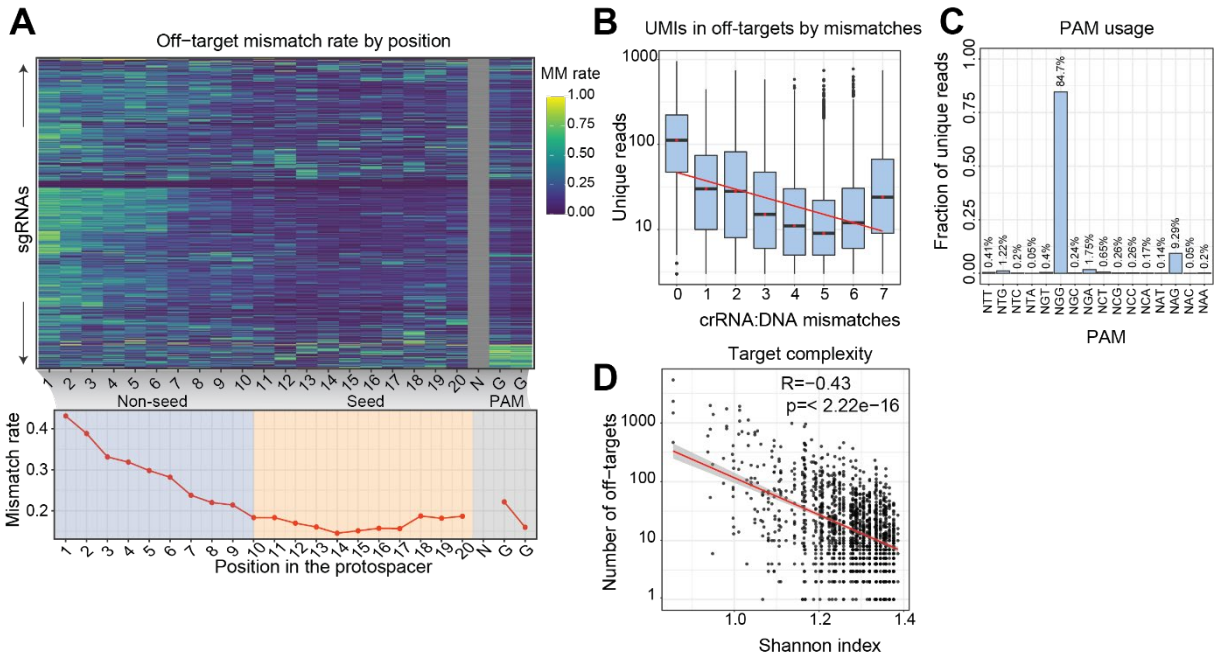


Figure 2.6. Determinants of Cas9 off-target activity. (A) Top: heatmap depicting crRNA:DNA mismatch accumulation along the protospacer of 92,375 off-target sites identified by BreakTag on 1,418 sgRNAs in the HiPlex 1 dataset. Bottom: a plot of the average mismatch rate along the protospacer. (B) Number of unique reads after deduplication using unique molecular identifiers for identified target sites containing 0–7 crRNA:DNA mismatches. (C) Percentage of unique reads for identified target sites containing noncanonical PAM sequences. (D) Correlation between the number of measured off-target cutting events and sequence complexity of the target site measured according to the Shannon index.

2.6.2.Specificity of engineered Cas9 variants

Several studies have reported the engineering of novel Cas9 variants with the goal of decreasing MM allowance and mitigate off-target effect. Different strategies have been applied for the engineering of novel Cas9 versions, with few successful attempts reducing the off-target cleavage without affecting on-target activity (see 1.4.5). Of note, a trade-off between activity and specificity has been described for Cas9 variants, which might hamper the editing efficiency in cells (Kulcsár et al., 2022).

We applied cell-free BreakTag to investigate the off-target landscape and activity of 7 different Cas9 variants previously described: HiFiCas9 (Vakulskas et al., 2018), xCas9 (Hu et al., 2018), SniperCas9 (J. K. Lee et al., 2018), HypaCas9 (J. S. Chen et al., 2017), EvoCas9 (Casini et al., 2018), and LZ3Cas9 (Schmid-Burgk et al., 2020) (Figure 2.7A). We produced the recombinant engineered variants in house (Figure 2.7B) and performed BreakTag using the Pool9 of the HiPlex 1 library in gDNA of HepG2 cells targeting 150 different endogenous loci. A total of 6303 off-targets were nominated for spCas9, whereas the variants displayed a wide range of off-targets (Figure 2.7C).

Next, we calculated the number of off-targets for each tested variants as a metric of specificity. EvoCas9 showed the highest specificity compared to SpCas9, followed by xCas9, LZ3Cas9, HiFiCas9, HypaCas9 and SniperCas9 (Figure 2.7D). To investigate if the on-target activity is retained in the tested variants, we scored each Cas9 based on their "Activity" and "Specificity". The Activity score indicates the number of reads at on-targets in comparison to SpCas9 as a metric of cleavage activity (Figure 2.7E). While SniperCas9 and HypaCas9 showed no reduction in on-target activity, we observed ~50% less on-target activity for xCas9 and EvoCas9, indicating that the mutations introduced in these variants hamper the overall cleavage activity of the nuclease (Figure 2.7C, D). These data indicate that a trade-off between specificity and activity is observed for some Cas9 variants, and BreakTag can be employed for the characterization of nuclease activity.

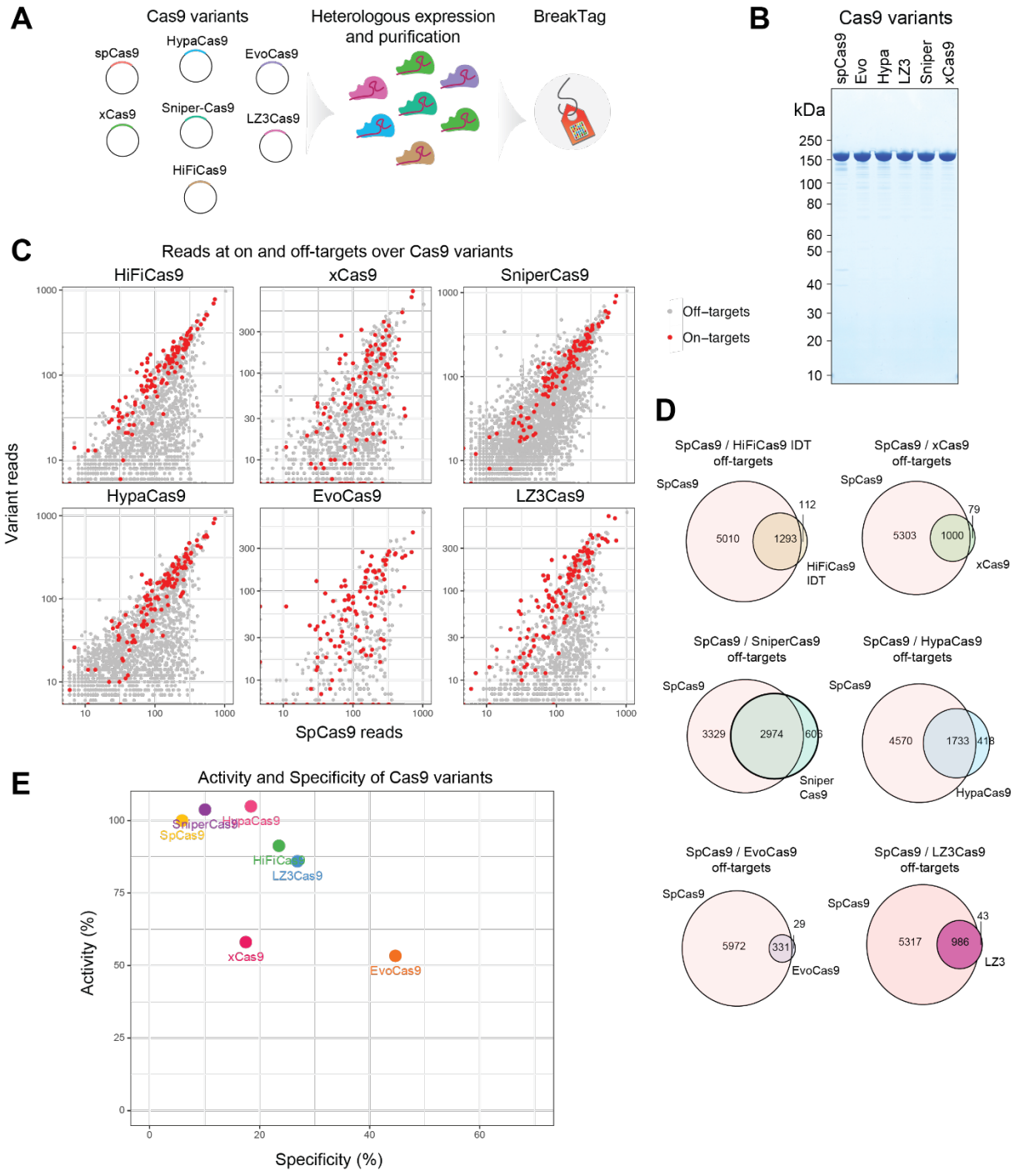


Figure 2.7. Characterization of Cas9 engineered variants. (A) Schematic of the production of engineered Cas9 variants. (B) Coomassie Blue staining of recombinant Cas9 variants used here. (C) Number of reads at on and off-targets between spCas9 and the tested variants. Red dots mark on-target sites and gray points depict off-targets. (D) Venn diagrams showing common cleaved sites mapped with BreakTag between spCas9 and the tested Cas9 variant. (E) Activity and specificity score of Cas9 variants. Activity and specificity were calculated as in (Schmid-Burgk et al., 2020). Specificity was calculated by subtracting 100 by the percentage of reads at off-targets over all sites. Activity is in relation of spCas9 was calculated by dividing the number of reads at on-targets for the variant by the reads at on-targets for spCas9.

2.7.Characterizing Cas9 DNA double-strand break scission

2.7.1.BreakTag retraces the original DNA double-strand break structure of DNA cutters

A unique advantage of BreakTag over competing *in vitro* methods such as CHANGE-seq (Lazzarotto et al., 2020), CIRCLE-seq (Tsai et al., 2017), SITE-SEQ (Cameron et al., 2017) and DIGENOME-seq (D. Kim et al., 2015), is the addition of a DSB end blunting step during sample preparation. This step uses a DNA polymerase with 3'-5' exonuclease activity, where 5' single-strand DNA overhangs (ssDNA) are filled-in and 3' ssDNA overhangs are resected, shifting the DSB read alignment. This blunting step allows the same protocol to be use for nucleases that cleave the DNA in a staggered manner, such as Cas12a (see Figure 2.1H). Coupled with that, BreakTag reads are directional, with both sides of the DSB mapping to opposite strands, providing a marker of the cut site.

We hypothesized that taking advantage of the directionality of the DSB reads in association with the processing of ssDNA overhangs at the cut site would allow us to retrace the original DSB end structure. To test this, we *in vitro* digested gDNA of HEK293 cells with a panel of restriction enzymes that cut the DNA in a blunt or staggered manner and performed BreakTag, As expected, DSB reads of a blunt cut generated with EcoRV abutted, mapped to opposite strands and started at the same location (Figure 2.8A). When enzymes that leave 1-3 nucleotide long 3' overhangs were used, a clear gap between the DSB was observed (Figure 2.8B), indicating a resection of the 3' overhangs at the cut site. Conversely, DSB reads from 5' overhangs overlapped due to the fill-in reaction (Figure 2.8C). We conclude that the DSB read overlap or gap can be used as a footprint for retracing the DSB-end structure at a given DNA cut site.

We reasoned that applying the same rationale would enable an investigation of the scission profile of Cas9-induced DSBs. The RuvC domain of Cas9 can cleave the nontarget strand at noncanonical positions, generating single-strand DNA (ssDNA) 5' overhangs (see 1.5.3). In the scenario of a blunt DSB, both the RuvC and HNH domains cleave the DNA strands between the third and fourth nucleotide ahead of the PAM sequence (positions 18 and 17, respectively), generating abutting DSB reads aligned at the expected cut site for blunt cuts (Figure 2.8D). If the RuvC domain cleaves the nontarget strand upstream of the HNH domain, 5' ssDNA overhangs are generated, and upon DSB end repair during BreakTag sample preparation, the PAM-proximal and PAM-distal reads overlap and no longer abut and should overlap, as observed for restriction enzymes that leave 5' ssDNA overhangs (Figure 2.8E). Furthermore, we adapted BreakInspectoR to parse the DSB reads into PAM-proximal and PAM-distal for Cas9 scission profile investigation (see Methods).

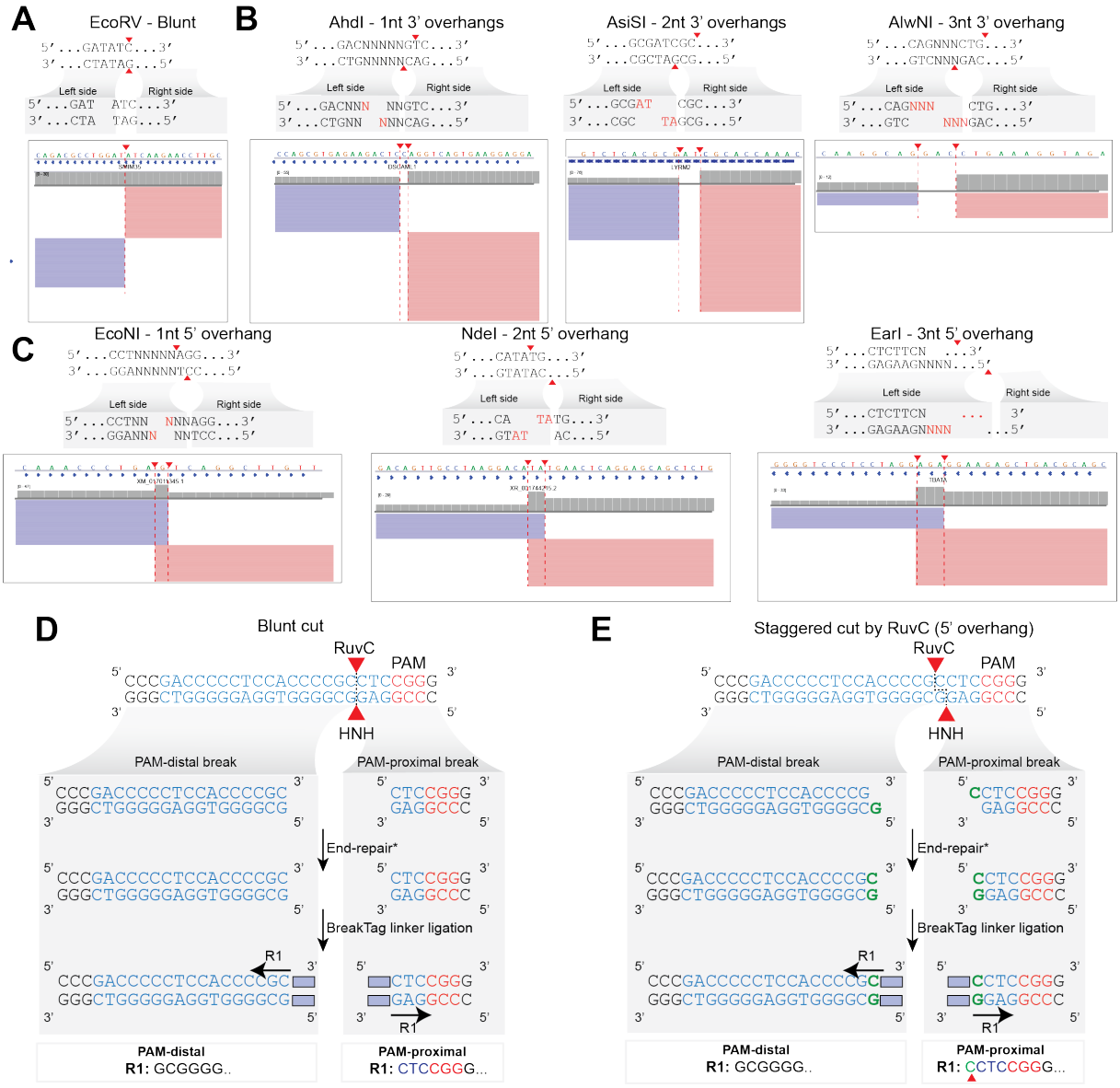


Figure 2.8. BreakTag allows profiling of Cas9 scission. gDNA of HEK293 cells was in vitro digested with a panel of restriction enzymes that generate blunt DSBs (A), 1–3 nt long 3' ssDNA overhangs (B), or 1–3 nt long 5' ssDNA overhangs at the cut site (C), and BreakTag was performed. IGV snapshots show raw mapped reads for a representative target site for each enzyme. Arrowheads indicate the start of DSB reads. Scheme depicting a blunt (D) and a staggered DSB (E) with a 1 nt 5' overhang. PAM-proximal side of the break starts 1 nt upstream (16|17) of the expected site for a blunt cut.

2.7.2.Characterizing Cas9 scission profile

It was initially believed that Cas9 leaves blunt DSBs, but later evidence supported the formation of staggered cuts by the nuclease (Gisler et al., 2019; Lemos et al., 2018; Molla & Yang, 2020; Müthel et al., 2023; Shi et al., 2019; Shou et al., 2018; Zuo & Liu, 2016). However, the available data relies on a few target sites and no study has systematically investigated the different types of DNA

DSB ends generated by Cas9. We assessed, for the first time, the frequencies of non-blunt Cas9-mediated DNA breaks using BreakTag using the footprints reported in 2.7.1 (Figure 2.9A).

We subset the HiPlex1 dataset with sites containing NGG PAM, and at least 16 reads on the PAM-proximal side of the DSB, yielding a total of 38,141 uniquely-cleaved endogenous sites for high-throughput scission profile investigation. Of note, DSB reads in HiPlex1 were significantly enriched over a non-targeting control, and different types DSB ends were observed (Figure 2.9B), with different gRNAs clustering based on the fraction of blunt reads per DSB (Figure 2.9C).

Profiling the structure of Cas9-induced DSBs revealed that Cas9 preferentially generates blunt DSBs (61.57%), but a significant portion contain 5' ssDNA overhangs (~35%) (Figure 2.9). 1 nucleotide 5’ overhangs represented the majority of non-blunt DSBs, but 2 and 3 nucleotide-long 5’ overhangs were also observed (Figure 2.9D). Interestingly, the presence of mismatches between the crRNA and gDNA influenced the Cas9 scission profile. In the absence of mismatches, 78.8% of the Cas9 DSBs were blunt, whereas approximately 19% of Cas9 DSBs were staggered (Figure 2.9D, right). At off targets, the number of blunt breaks decreased (to 55.8%), whereas the percentage of staggered breaks increased (to ~38%) (Figure 2.9D, lower right). Our data indicate that Cas9 has a preference for generating blunt DNA breaks, but staggered cuts are prevalent and slightly higher at off-targets.

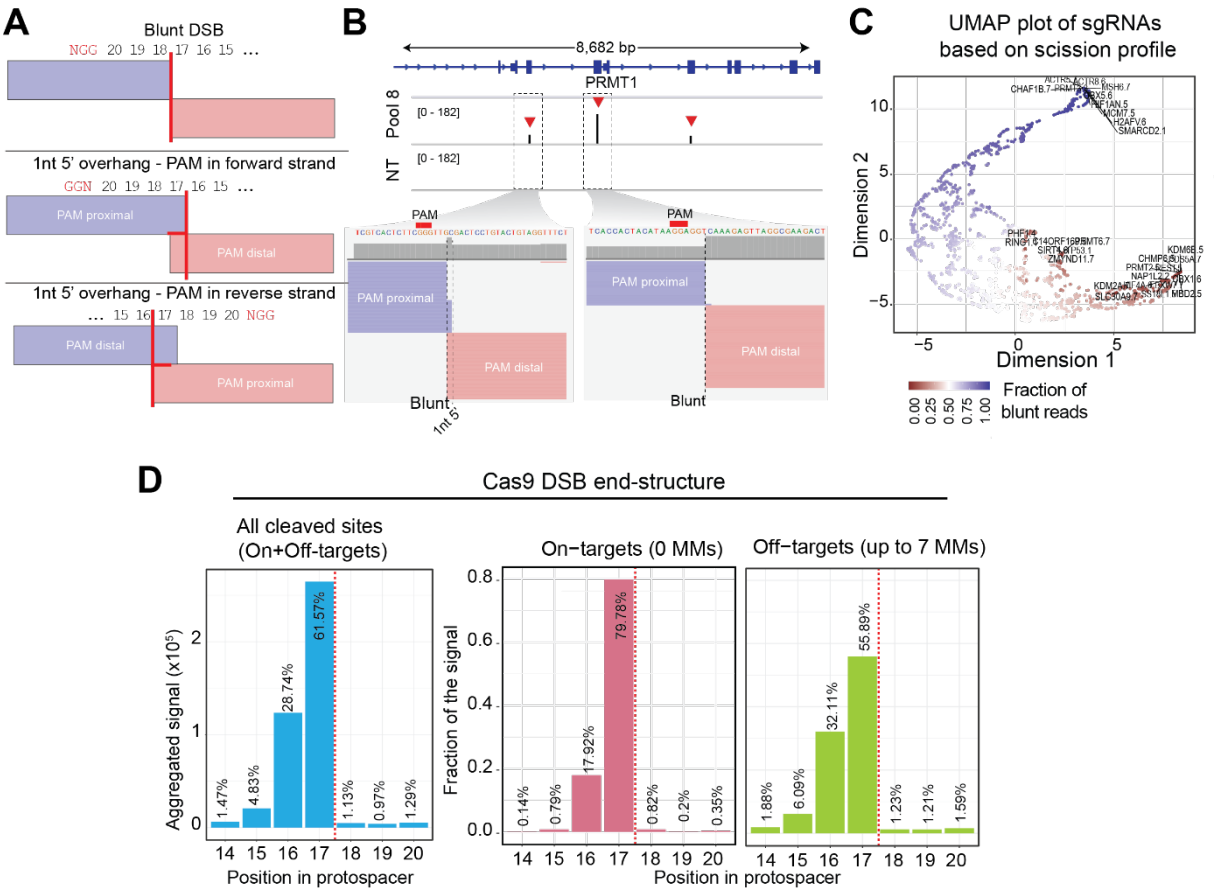


Figure 2.9. Cas9 has a flexible scission profile. (A) Schematic representation of the different read alignments for 5' overhangs in BreakTag data. Blunt DSB reads start at the same location, but 5' overhangs are shifted due to the end-repair reaction during BreakTag sample preparation. (B) Representative IGV snapshot depicting three on-target DSBs for the PRMT1 gene identified by BreakTag on gDNA (HepG2 cells) digested with HiPlex RNPs loaded with pools 1–10 of the HiPlex dataset 1. Examples of sites cut in mostly staggered (left) or blunt (right) configurations are shown below and zoomed-in. (C) UMAP representation on two dimensions of relatedness between sgRNAs based on average scission profile. Dimensions 1 and 2 are representations in a reduced dimensional space (arbitrary units) of the fraction of signal each sgRNA has in positions 14–20 of the protospacer. The color scale represents the fraction of signal at the expected cut site, ranging from 100% (blue, all blunt, no signal outside the expected cut site) to 0% (red, all staggered, all signal outside the expected cut site). The top 25 most blunt and staggered sgRNAs are highlighted in the plot. sgRNA self-organize in this representation based on their scission profile and preferred overhang length. (D) Left: aggregated signal of different DSB end structures for on/off targets in the HiPlex1 dataset. The fraction of blunt or staggered DSBs for on-targets (pink) and off-targets with up to 7 mismatches (MM; green) are shown center and right, respectively. Position 17: blunt DSBs; 16–14: 5' overhangs. Dotted line indicates the expected cut site for a blunt DSB.

2.7.3.Scission profile is not random, but a target-specific effect

Given that the sgRNAs clustered based on the proportion of blunt reads (Figure 2.9C), we next wondered if different target sequences show the same proportion of blunt/staggered reads. For this, we analyzed the position in the protospacer in which the PAM-proximal read starts, as it should indicate the original DSB end structure (Figure 2.9A). Assuming a non-flexible HNH

cleavage (Molla & Yang, 2020), the PAM-distal read should always map at position 17 regardless of the RuvC cleavage site. We observed distinct patterns in terms of PAM-proximal signal for the tested gRNAs, with most of the signal starting at position 17 (blunt DSB), but others showing increased staggeredness with 5' overhangs (Figure 2.10A). Most of the gRNAs tested displayed a mixture of blunt and staggered DSBs at different ratios for on- and off-targets (Figure 2.10B), but these ratios were target-sequence specific. For example, a sgRNA targeting the TAPBP locus showed nearly complete blunt cuts (Figure 2.10C), whereas a gRNA targeting the SUZ12 locus showed a mix between blunt and staggered reads, with a preference for 1nt 5' overhangs (Figure 2.10D).

Given that the BreakTag cell-free mode removes the chromatin environment and lacks processing of DNA DSB ends via the repair machinery, we conclude that the Cas9 scission profile is not a random phenomenon, and the target sequence composition is a major determinant of the DSB end formation.

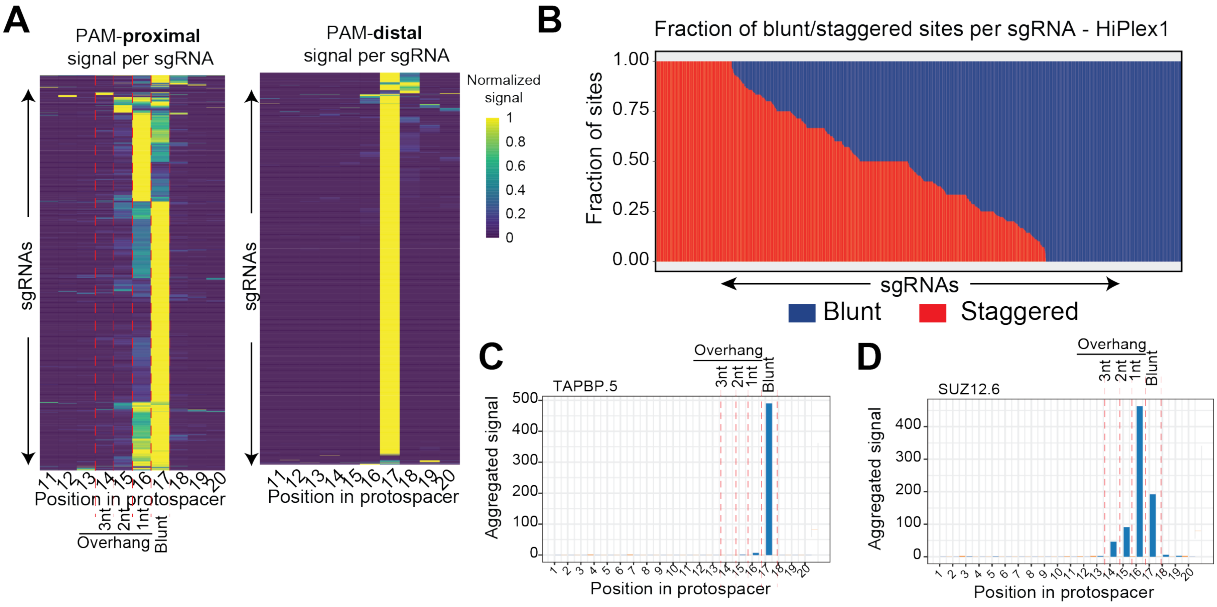


Figure 2.10. Scission profile is a target-specific effect. (A) Accumulation of reads mapped onto the PAM-proximal strand (scaled) or PAM-distal strand along the protospacer over 1,418 sgRNAs of the HiPlex1 dataset for all identified targets. For the PAM-proximal side of the cut, most sgRNAs accumulate signal at position 17 of the protospacer (corresponding to blunt DSBs), versus approximately 34% at positions 14–16 of the protospacer corresponding to staggered DSBs with 1–3 nt overhangs, respectively. (B) Fraction of blunt (blue) or staggered (red) DSBs for each sgRNA. Each column represents the fraction of blunt or staggered reads for on and off targets of a given sgRNA. (C,D) Representative examples of target sites at which Cas9 cuts preferentially in blunt or staggered configuration. Aggregated BreakTag signal along the protospacer for "TAPB.5" sgRNA on and off targets (n=3), which preferentially accumulates on position 17 of the protospacer (C). Aggregated BreakTag signal along the protospacer for "SUZ12.6" sgRNA on and off targets (n=56), which accumulates mostly on position 16 of the protospacer (D).

2.7.4.The scission profile of engineered Cas9 variants

We next sought out to investigate if mutations introduced along the Cas9 protein could affect scission profile. To test this, we leveraged the seven high fidelity Cas9 variants shown in 2.6.2 that contain mutations at different domains, including the nuclease domains and linker regions (Figure 2.11A), and performed cell-free HiPlex BreakTag with a pool of 150 sgRNAs. As a metric for comparing the frequency of staggered/blunt breaks, we calculated the "*blunt rate*", defined as the abundance of blunt DSBs profiled at the expected site for a blunt cut (between positions 17 and 18) relative to the total DSBs profiled in a region around [−3, +3] the expected cut site for the PAM-proximal read (see Materials and Methods).

We observed that, with the exception of LZ3 (Schmid-Burgk et al., 2020), the tested variants showed a normal distribution of the blunt rate (Figure 2.11B). Of all the cuts generated by the WT SpCas9, 16% were highly staggered (i.e, blunt rate < -2, 80% of staggered breaks), compared to 52.03% for LZ3 (Figure 2.11B). LZ3 showed the lowest correlation with SpCas9 (pearson, R: 0.49) compared to the other tested variants (Figure 2.11C), and LZ3 showed the poorest correlation pattern with all tested variants (Figure 2.11D). Of note, LZ3 was able to generate staggered cuts at sites previously cut blunt by SpCas9 (Figure 2.11E). These data indicate that LZ3 has a different scission profile compared to the other variants tested.

In sum, we demonstrate that the engineering of Cas9 variants affords as an opportunity to discover highly staggered variants for CRISPR-based therapeutics, and places BreakTag as a tool for characterizing the scission profile of novel genome editors.

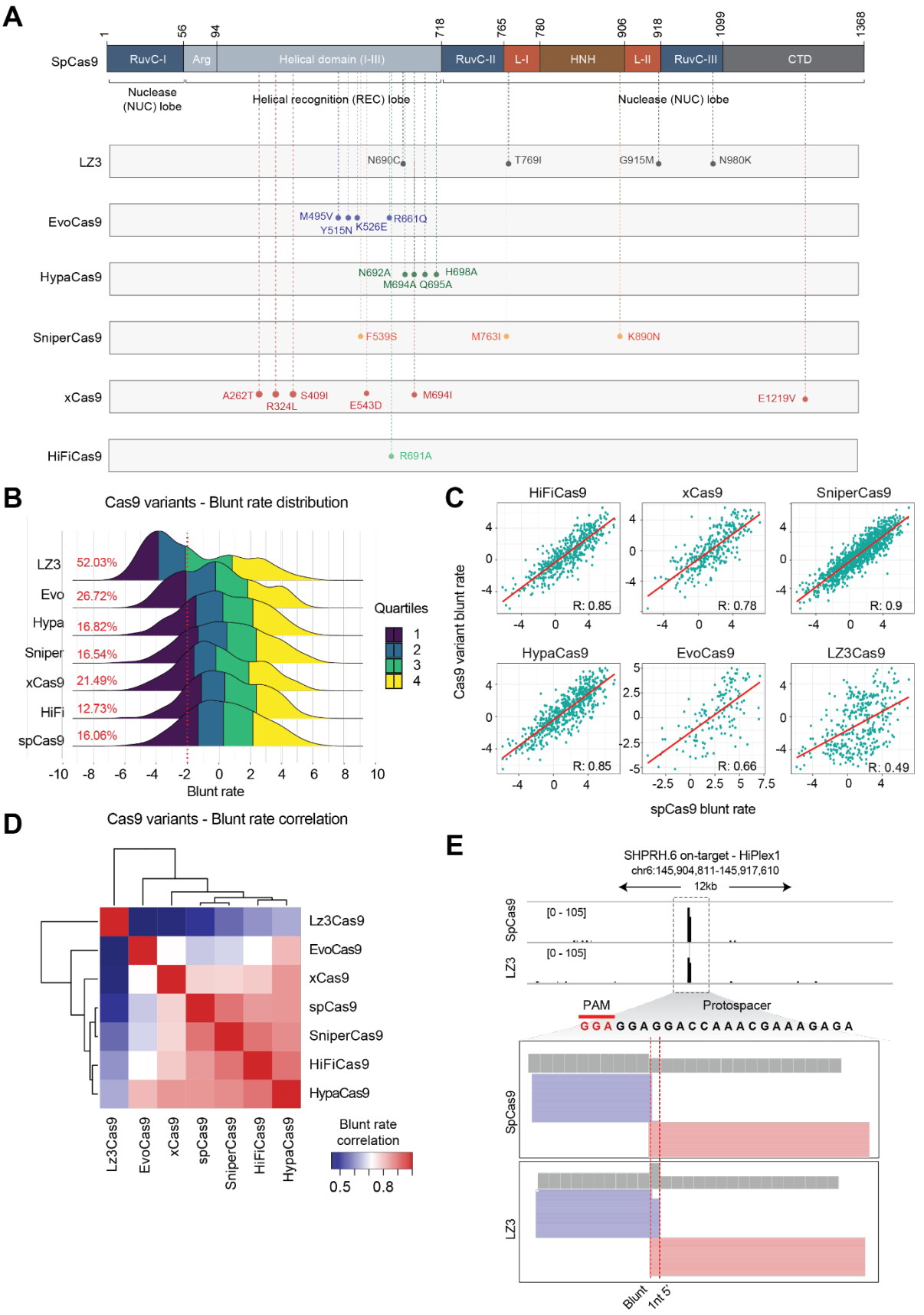


Figure 2.11. Characterization of the scission profile of six high-fidelity Cas9 variants using pool 9 of the HiPlex1 library in gDNA of HepG2 cells. (A) Schematic representation of the mutations found in the

tested Cas9 variants. Original spCas9 domain map as in Jiang et al., 2016. (B) Distribution of blunt rate for tested Cas9 variants for on and off targets. Colors show quartiles. Dashed line marks log2 rate of −2 (80% staggered DSBs). The percentage of sites with more than 80% staggered DSBs are shown. (C) Blunt rate pearson correlation between spCas9 and the different high-fidelity variant. Each point is a cleaved site (on or off target). (D) Hierarchical clustering of the blunt rate correlation between tested variants. (E) IGV snapshot showing an example of differential scission profile for the on-target sequence of SHPRH.6 sgRNA (HiPlex1 library).

2.7.5.XGScission identifies determinants of Cas9 scission profile

We identified a target-specific scission profile, indicating that the sequence composition might be an important factor underlying Cas9 incision. To identify important sequence features in SpCas9 cleavage we developed *XGScission*, an XGBoost regression model using the 2D one-hot-encoded representation of the target and guide sequence as predictors, together with the number of mismatches in the non-seed (positions 1-10) and seed (positions 11-20) parts of the protospacer (Figure 2.12A). Because the number and scission profiles of the identified targets differ greatly among the tested sgRNAs, we subset the HiPlex1 dataset as training instances to balance the nucleotide composition and scission profile. First, we limited the number of cleaved sites per sgRNA to 100 to avoid overrepresentation of highly promiscuous sgRNAs. Second, we balanced the representation of blunt (blunt reads >80%) and staggered (blunt reads >20%) sites in order to compensate for the under-representation of highly staggered targets. This resulted in a final set of 18,759 *instances* in the training set.

Our model achieved high performance, as measured by the correlation between the predicted and observed blunt rates in the cross-validated sets (R=0.74) (Figure 2.12B). This further confirms that scission profile is not random, and can be predicted with high confidence based on the sequence context of the target sequence. The high predictive power of our model allowed us to investigate key positions within the protospacer that determines whether Cas9 cleaves the target DNA in a staggered or blunt manner. We observed that positions 16–20 (5 nt upstream of the PAM) were important for predicting the scission profile, with guanines at positions 17 and 18 having the highest importance (Figure 2.12C,D). We next sought to identify sequence compositions associated with a blunt or staggered cut by interrogating the importance of each base along the protospacer of the NTS. Strikingly, we identified that a G at position 17 was predictive for a blunt DSB, whereas G at position 18 was associated with staggered DSBs (Figure 2.12E). To investigate the effects of 17G and 18G on Cas9 scission with our dataset, we grouped the cleaved sites into "blunt" (0%–33% of PAM proximal reads mapping outside of position 17: staggered reads), "middle" (33%–66% staggered reads) and "staggered" (66%–100% staggered reads). Cas9 was in general more likely to blunt cut at on-target sequences than at off targets

where mismatches are present (ANOVA: $p<2^{-16}$) (Figure 2.12F). In accordance with the model predictions, Cas9 was more likely to cleave in a blunt configuration at sites with a G at position 17 compared to sites with A, C or T, at both on and off targets (Pearson's chi-squared test: $p<2^{-16}$) (Figure 2.12F). In contrast, if a G occupied position 18, Cas9 was more likely to cleave in a staggered configuration than if A, C, or T occupied that position (Pearson's chi-squared test: $p<2^{-16}$) (Figure 2.12F).

We conclude that the base composition surrounding the DSB is a strong determinant of the Cas9 scission profile.

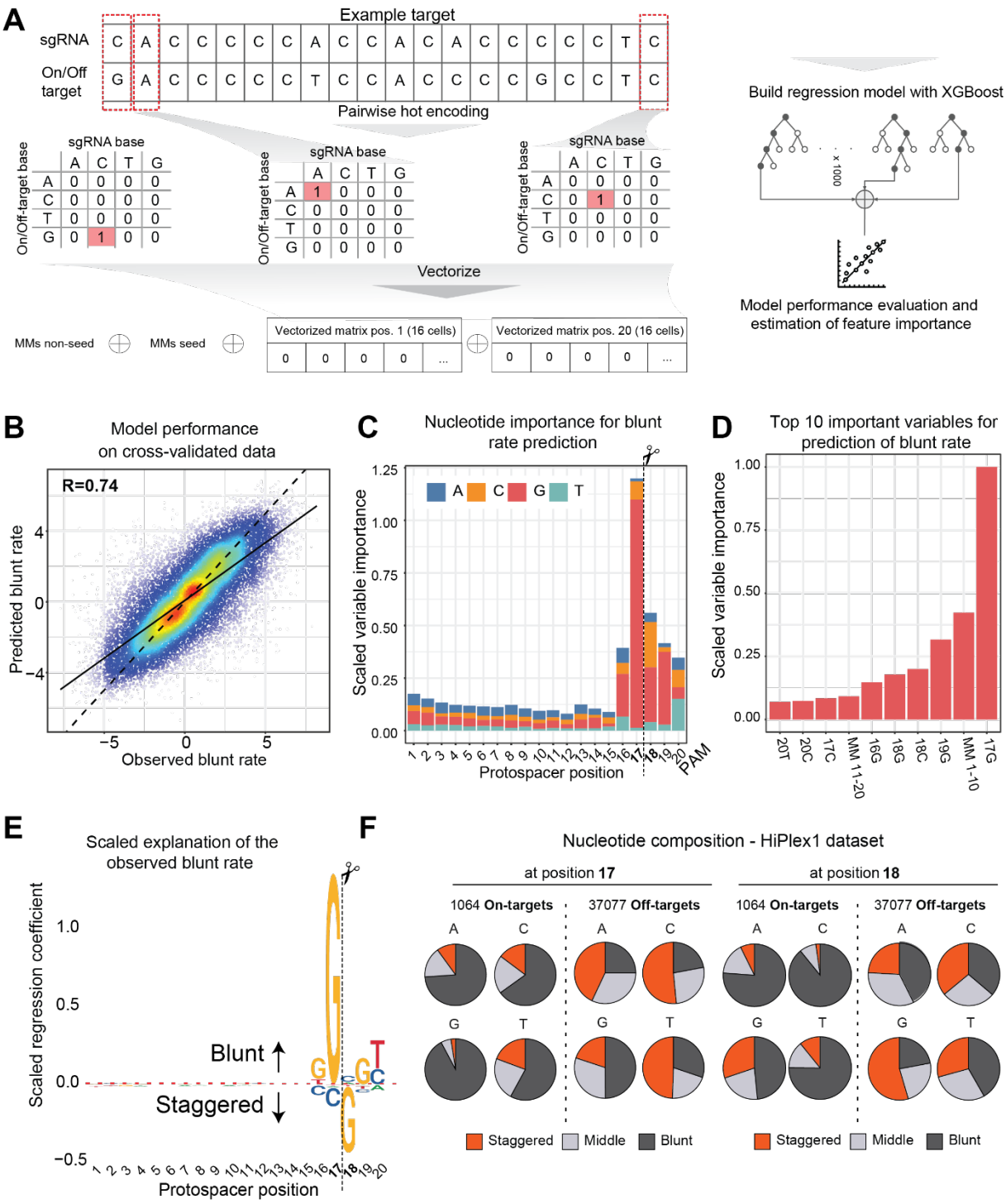


Figure 2.12. A machine learning model predicts Cas9 scission profile. (A) Overview of the machine learning strategy for predicting Cas9 scission profile. A model for regression was built using the XGBoost method and trained with a two-dimensional one-hot encoding of the correspondence between the protospacer and sgRNA sequences, in addition to the numbers of mismatches in the seed (positions 11–20) and non-seed (positions 1–10) parts of the protospacer. The training set consisted in a subset of 18,759 on and off targets with a minimum coverage of 16 reads in the PAM-proximal strand. This subset was balanced to avoid overrepresentation of highly promiscuous sgRNAs with many targets, limiting the number of targets selected per sgRNA to 100. Targets within each sgRNA were sampled randomly to try to equal the probability of selecting targets with staggered, blunt, or mixed configurations. (B) Model performance evaluation using cross-validated data. Ten rounds of cross-validation were performed. This panel shows the correspondence between expected (predicted) and observed log2 ratio of reads indicating a blunt or a

staggered cut. (C) Importance of the nucleotide composition and position in the protospacer, as estimated by the XGBoost method. Values on the y axis are scaled to the most important nucleotide+position. The dashed vertical line indicates the cut site for a blunt DSB. (D) Top ten most important variables for the prediction of blunt rate. MM 1–10: mismatches in the non-seed part of the protospacer (positions 1–10); MM 11–20: mismatches in the seed part of the protospacer (positions 11–20). (E) Observed blunt rate explained by the sequence composition of the protospacer. Coefficients of a linear regression model fit to the nucleotide composition independently on each position of the protospacer are shown as letters scaled according to the importance of that nucleotide and position, as estimated by the XGBoost model. The dashed vertical line indicates a cut site for a blunt DSB. (F) The effect of each base at positions 17 (left) and 18 (right) in the scission profile for on and off targets in the HiPlex1 library for sites with at least 16 reads in the PAM-proximal strand.

2.8.The role of DSB end structure in gene editing precision

2.8.1.Staggered DSBs are linked with precise insertions

Previous evidence supported an association between Cas9 scission and repair outcome (See 1.5.3), but the lack of scalable methods to assess scission profiles has precluded a systematic investigation.

We sought to generate a matched dataset with the scission profile and indel outcome of several target sequences. For this, we deployed our machine learning model to 2,791 genomic gRNA targets, for which the repair outcome was previously characterized (Allen et al., 2019), to predict the blunt rate for each gRNA sequence. We then selected the predicted top 700 most blunt and top 700 most staggered sites for HiPlex BreakTag (hereafter referred to as the "HiPlex2" library) to correlate their Cas9 scission profile with their repair outcome (Figure 2.13A). The predicted blunt rate of this dataset was highly correlated with the actual experimental scission profile obtained by BreakTag, confirming the robustness of our model (Figure 2.13B).

We next interrogated the most common indel outcome for the tested sgRNAs as a function of the scission profile. We observed that blunt cuts were equally represented across indel sizes (Figure 2.13C). By contrast, a striking enrichment of staggered sites was found at genomic loci that are repaired as single-nucleotide insertions (+1 indels) (Figure 2.13C). Of note, over 90% of sites with a +1 indel as the most common repair outcome were staggered DSBs, demonstrating a clear association between scission profile and DNA repair (Figure 2.13D). Staggered breaks generated more precise indels (i.e., at a higher frequency) compared to blunt cuts for −1 and +1 indels (Figure 2.13E). Collectively, these data indicate a strong association between the staggered scission and repair precision.

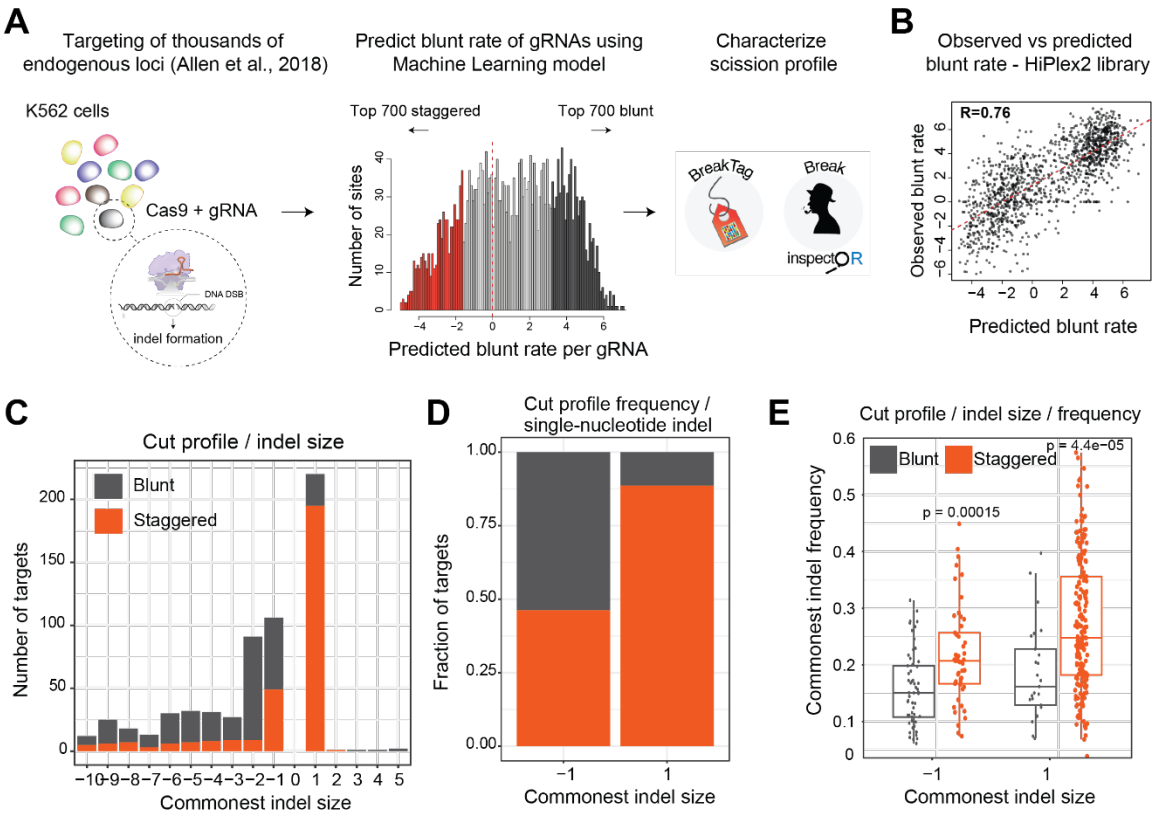


Figure 2.13. The relationship between scission profile and indel outcome. (A) Distribution of the predicted blunt rate for 2,791 gRNAs characterized previously (Allen et al., 2019). We ranked the sites according to their predicted blunt rate and selected the top 700 (most blunt) and bottom 700 (most staggered) for HiPlex BreakTag to generate matched datasets of Cas9 scission profile. (B) Correlation between predicted blunt rate by our model and observed blunt rate using BreakTag in HEK293 gDNA for top 700 staggered and top 700 blunt gRNAs identified. (C) HiPlex BreakTag (HiPlex2 dataset) was performed to assess the scission of 610 sites in a matched dataset with known repair outcomes (+1 to +5 nt insertions, −1 to −10 nt deletions). Using our model, we predicted genome sites with staggered or blunt DSBs (blunt if predicted log2 ratio of blunt vs. staggered BreakTag signal>0; otherwise staggered) from HiPlex2 library to use in this experiment. (D) Cut profile frequency in single-nucleotide indels shows a significant enrichment of +1 insertions in staggered sites (Fisher's test: odds ratio=8.99, p-value=8.345e−16). Colors represent the fraction of blunt (gray) or staggered (orange) sites showing 1 nt deletions (−1) and/or 1 nt insertions (+1). (E) Frequency of 1 nt deletions or insertions in relation to scission profile (t-test: p-value=0.00015 for −1 deletions, p-value=4.4e−5 for +1 insertions).

2.8.2.Staggered DSBs are linked with predictable indel formation

Precise insertions are desirable repair outcomes in the context of correcting pathogenic alleles and gene knockout. Massive interrogation of Cas9-mediated indel formation revealed that single nucleotide insertions are usually a duplication of the base at position 17 of the protospacer (see 1.5.3), suggesting that staggered DSBs might favor the formation of these templated insertions.

However, the lack of high-throughput scission-aware tools precluded systematic investigation of such association.

To understand the effect of scission profile on the efficiency of templated insertions, we assessed the number of loci for which the most frequent repair was a templated insertion as a factor of base composition at positions 17 and 18 of the protospacer in the HiPlex2 dataset. If the ssDNA overhang at the cut site is used as a template for repair, we would expect that the most common insertion would be a copy of the overhang sequence. Because most overhangs generated by Cas9 are 1 nt long (Figure 2.9B), we anticipated that position 17 would be duplicated in most cases (Figure 2.14A). We observed that the most common nucleotide inserted at staggered sites was a duplication of the base at position 17, indicating template insertions are a common repair outcome of staggered DSBs (Figure 2.14B). Target sites with G at position 17 showed a low number of templated insertions, as expected for blunt cuts (Figure 2.14C). By contrast, target sites with G in position 18 were more likely to use the nucleotide at position 17 as the template for the single-nucleotide insertions (Figure 2.14C). Of note, 17G loci generated a dismal indel frequency (Figure 2.14D), whereas 18G loci showed a higher precision of 1nt insertions (Figure 2.14E).

This finding highlights that staggered cleavage by Cas9 mediates the formation of precise and predictable insertions as the DNA repair outcome.

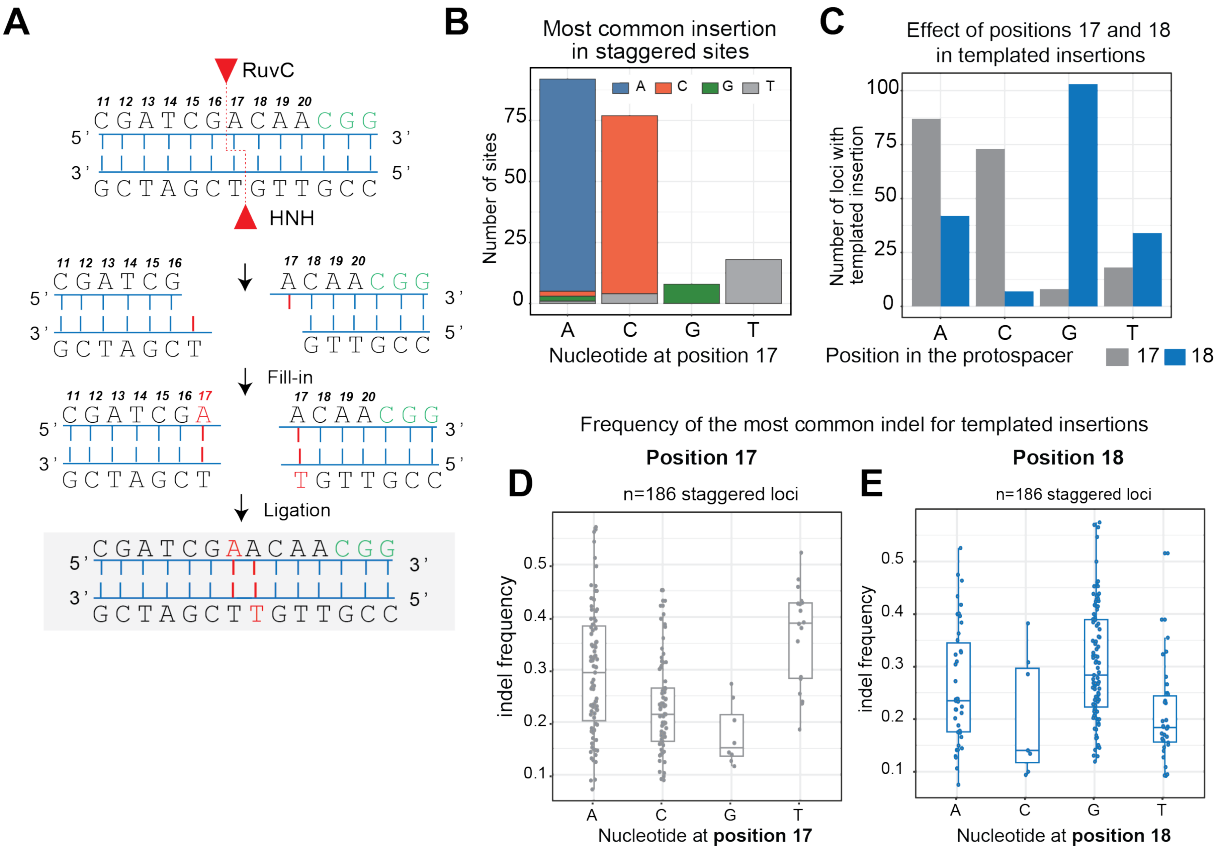


Figure 2.14. Staggered breaks mediate predictable insertions. (A) Scheme depicting how 1 nt 5' overhangs can promote templated repair, leading to 1 nt insertions. (B) Most common insertion at staggered sites according to the type of nucleotide at position 17. (C) Number of loci with templated insertion according to the base composition at positions 17 (gray) or 18 (blue). (D) Frequency of the most common indel for template insertions as a function of nucleotide at position 17 for all staggered-cleaved loci with a +1 indel as the repair outcome (n=186) from the HiPlex2 library. (E) Frequency of the most common indel for template insertions as a function of nucleotide at position 18 for 186 staggered loci with templated insertions from the HiPlex2 library.

To further demonstrate the mechanistic link between Cas9 scission profiles with the cellular indel outcome, we have employed a parallel repair outcome strategy using a scission-based design of target sequences. We employed our model trained on SpCas9 HiPlex BreakTag data on target sequences endowed by NGG and predicted their blunt rate (Figure 2.12A). Based on the predicted scission profile, we grouped the target sequences into "Blunt" or "Staggered" groups, and leveraged a gRNA-target pair strategy in order to assess the repair outcome of the predicted blunt rates in a parallel fashion. Briefly, the target sequences and the gRNA matching the protospacer were cloned into a lentiviral vector and delivered to K562 and HeLa Cas9-expressing cells (cloning strategy adapted from Allen et al., 2019, Figure 2.15A). Cells were kept under antibiotic selection, and after DSB repair, gDNA was extracted from cells and the repair outcomes were assessed using amplicon sequencing (Figure 2.15B).

We observed that for both K562 and HeLa cells, blunt cuts preferably generated deletions, whereas 1nt insertions were enriched for target sequences predicted to be cut in staggered manner (Figure 2.15C,D), with a significantly higher insertion rate found in both cell types in the staggered pool (Figure 2.15E). In line with these findings, templated insertions (i.e, duplication of the base found at position 17), represented ~76% of all insertions in the staggered pool for both cell types, as expected for DSBs containing 5' ssDNA overhangs (Figure 2.15F).

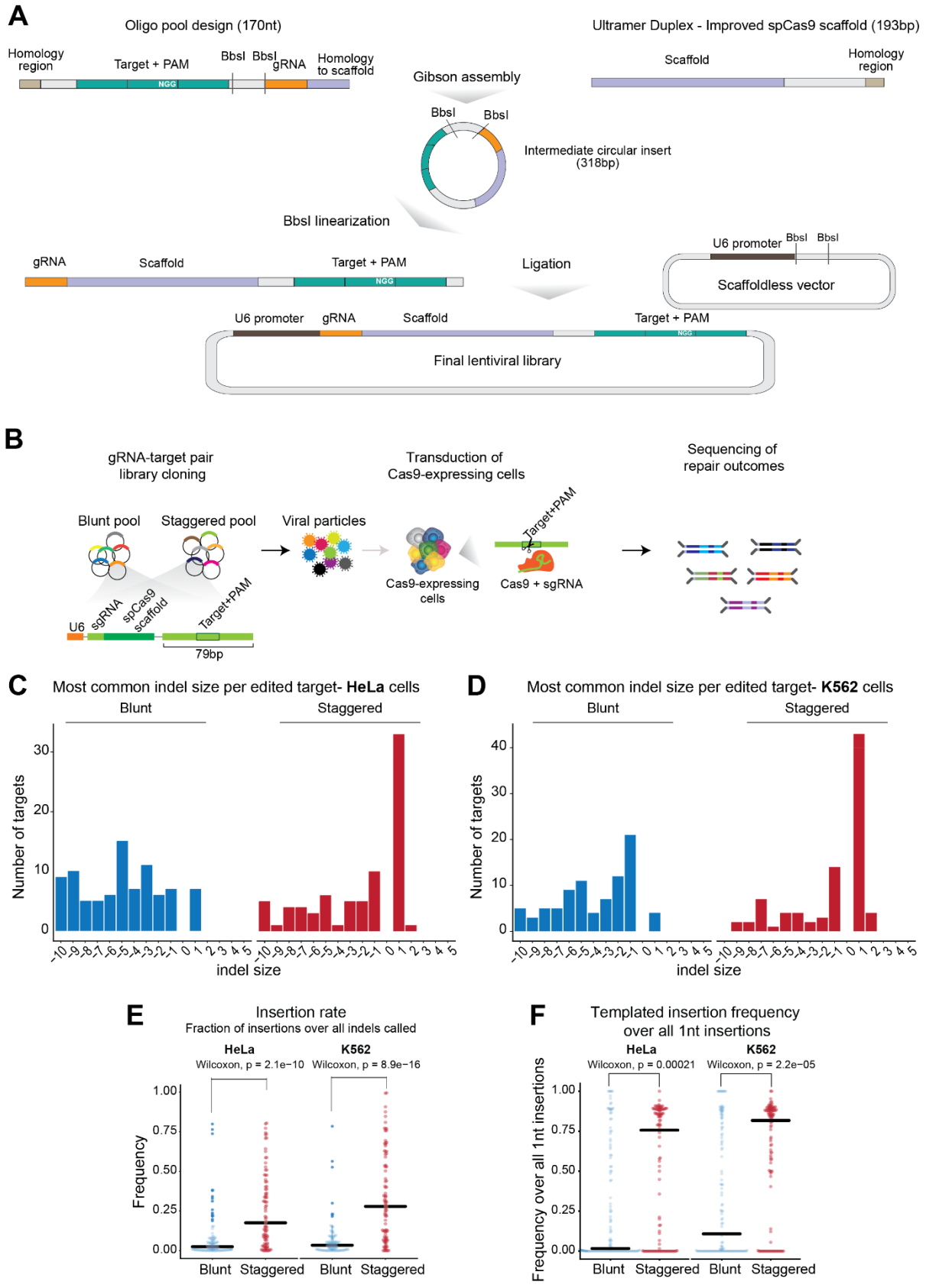


Figure 2.15. Parallel assessment of indel outcomes of target sequences predicted to be cut preferably in a blunt or staggered manner. (A) Schematics of the strategy used to clone gRNA-target pairs into a lentiviral vector. Briefly, we designed the 79nt portion of the target sequence+PAM and ordered

it in a Pool format. We performed a Gibson assembly reaction with an Ultramer Duplex containing a portion of the improved SpCas9 scaffold to assemble the insert containing the gRNA and target sequence. The intermediate circular insert was linearized and ligated into a scaffoldless pKLV2-U6(BbsI)-PKGpuro2ABFP-W (addgene #67974). (B) Schematics of gRNA-target pair experimental design. (C) Most common indel size found per edited target in HeLa-Cas9 or (D) K562-Cas9 cells. A total of 231 gRNA-target pairs (111 staggered and 120 blunt) were used for this analysis after filtering for sites with at least 100 mutated reads. (E) Insertion rate of target sequences predicted to be cleaved preferably in a blunt or staggered manner. Insertion rate was calculated as the fraction of insertion over all indels called. (F) Frequency of templated insertions over all +1 indels. Insertions were considered as templated when the inserted base is the same nucleotide found in position 17 of the protospacer.

2.8.3.Characterization of scission profile determinants of a Cas9 variant that generates highly staggered breaks

Engineering of high-fidelity Cas9 variants is a continuous effort in increasing gene editing safety, but the effect of point mutations in the different protein domains on cleavage patterns has not been formally assessed. We identified that LZ3 Cas9 has a lower bunt rate correlation with SpCas9 and other variants (Figure 2.11D) in a subset of 150 gRNAs tested. We hypothesized that this lower correlation is a consequence of different sequence determinants of staggered cleavage. To test this, we generated a HiPlex dataset with LZ3 combined with the HiPlex1 library, to investigate the frequencies and determinants of blunt and staggered cuts.

We observed that ~48% of LZ3 DSB reads accumulated at position 17, reminiscent of blunt DSBs, whereas ~47% of breaks displayed 5' overhangs (Figure 2.16A). The majority of non-blunt breaks were 1nt 5' overhangs (38.24%), but 2nt (8.44%) and 3nt (2.97%) overhangs were also observed (Figure 2.16A). Interestingly, as opposed to SpCas9, LZ3 on-target sites tended to be cleaved in a more staggered manner than off-targets, with ~48% of on-target sites displaying 1nt 5' overhangs (Figure 2.16A).

Similar to SpCas9, the proportion of blunt to staggered breaks was gRNA-dependent, indicating that sequence determinants govern LZ3 cleavage patterns (Figure 2.16B). To identify those determinants, we trained XGScission using the 2D one-hot-encoded representation of the correspondence between the 20 nucleotides of the protospacer and guide sequences as predictors, together with the crRNA:DNA mismatches for BreakTag data on LZ3. We observed that positions 20, 19 and 17 were important for predicting the blunt rate (Figure 2.16C, D), with a 19G being the top predictor of a blunt rate. Interestingly, similar to SpCas9, a 17G sequence was predictive of a blunt cut, but a 19G was highly predictive of a staggered DSB (Figure 2.16E). Finally, we investigated the insertion frequency of 19G loci in LZ3 and SpCas9 using publicly available data (Schmid-Burgk et al., 2020). We observed that 19G loci had a significantly higher insertion

rate in LZ3 compared to when other bases occupied that position, but not SpCas9 (Figure 2.16F). Altogether, these data indicate that a rational engineering of Cas9 variants might be a feasible strategy for introducing high frequency insertion at target sequences where SpCas9 cleaves in a blunt manner.

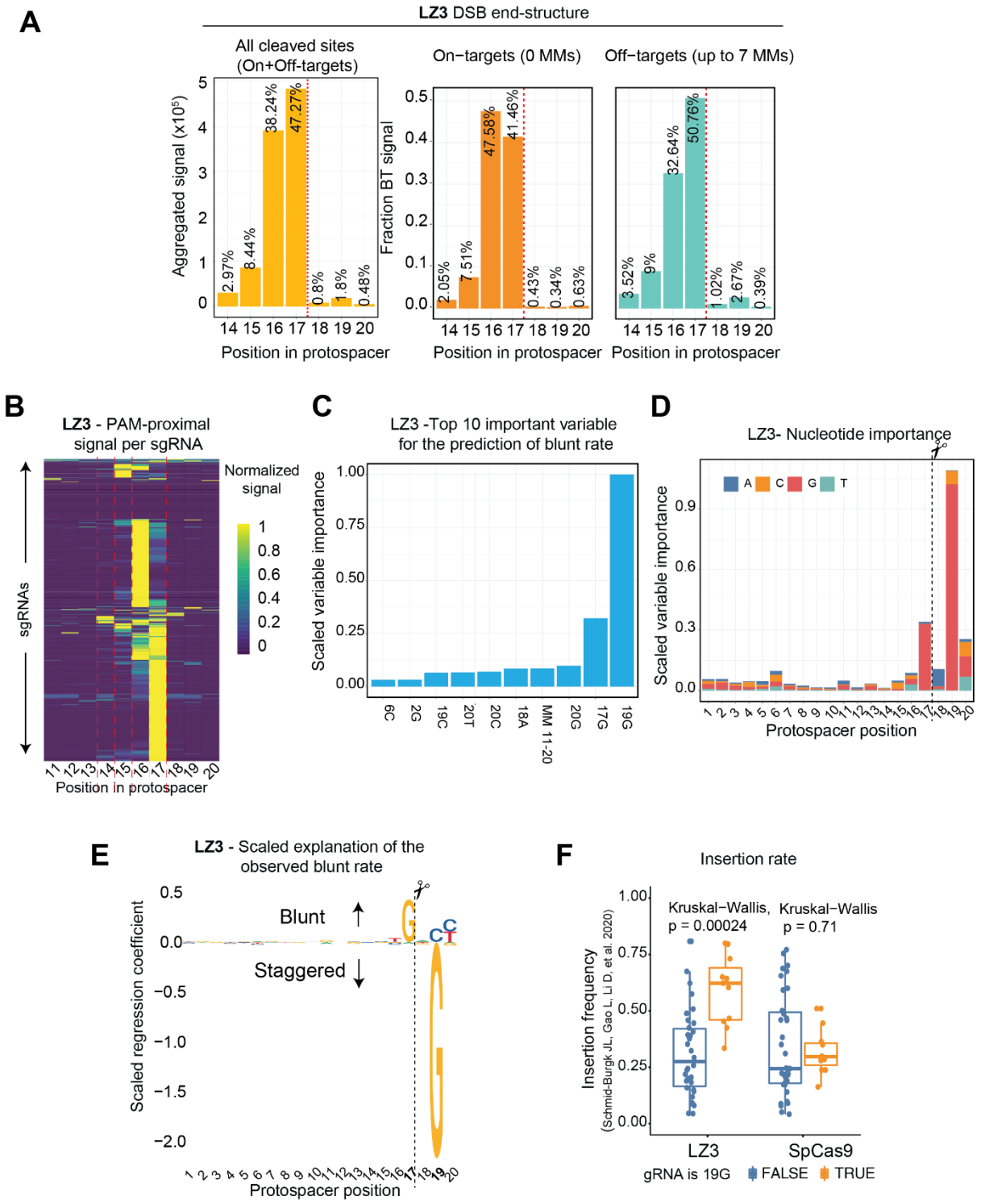


Figure 2.16. Characterization of LZ3Cas9 scission profile using BreakTag. (A) Left: aggregated signal of different DSB end structures for on/off-targets in the HiPlex1 library generated with the LZ3 nuclease. The fraction of blunt or staggered DSBs for on-targets (pink) and off-targets with up to 7 mismatches (MM; green) are shown center and right, respectively. Position 17: blunt DSBs; 16–14: 5' overhangs. Dotted line indicates the expected cut site for a blunt DSB. b, Accumulation of reads mapped onto the PAM-proximal strand (scaled) along the protospacer over 4,543 sgRNAs of the HiPlex1 library generated with the LZ3 nuclease for all identified targets with an "NGG" PAM. (C) Top ten most important variables for the prediction of LZ3 blunt rate. MM 11–20: mismatches in the seed part of the protospacer (positions 11–20). (D) Importance of the nucleotide composition and position in the protospacer, as estimated by the XGBoost method. Values on the y axis are scaled to the most important nucleotide+position. The dashed vertical line indicates the cut site for a blunt DSB. (E) Observed LZ3 blunt rate explained by the sequence composition of the protospacer. Coefficients of a linear regression model fit to the nucleotide composition independently on each position of the protospacer are shown as letters scaled according to the importance of that nucleotide and position, as estimated by the XGBoost model. The dashed vertical line indicates a cut site for a blunt DSB. (F) Boxplot comparing insertion frequency of target sites where a G occupied position 19 of the protospacer, expected to create staggered breaks in LZ3 only, to those target sites that do not.

2.9.Exploiting Cas9 scission profile for personalized gene editing

2.9.1.Single-nucleotide polymorphisms shift Cas9 scission profile

Given the strong dependency of Cas9 scission profile on the sequence context, we investigated if common human genetic variation has any effect on cleavage patterns. We focused our analysis on Single-nucleotide polymorphisms (SNPs), as they account for >99% of the genetic variation between two human genomes (1000 Genomes Project Consortium et al., 2015) and can change Cas9 target nucleotide sequences. We asked if SNPs at positions 17 or 18 of a putative Cas9 target sequence can influence scission profile of SpCas9.

To test this, we leveraged the benchmarked genomes of seven individuals where sites of variation between 2 family trios (Ashkenazim Jewish and Chinese backgrounds) and a single individual (Mormon background) were fully characterized by the Genome-in-a-Bottle (GIAB) consortium (Zook et al., 2016, 2019). We predicted the blunt rate of all putative SpCas9 target sites (i.e next to an NGG) containing a SNP at position 17 (n= 394,330) or 18 (n=395,368) of the protospacer among GIAB individuals using our machine learning model trained with SpCas9. Next, we predicted the effect of each base substitution in the Cas9 scission profile by calculating the difference between the predicted blunt rate for reference and alternative alleles. Based on our analysis, we selected 300 sites with an SNP at positions 17 or 18 and the highest predicted difference in blunt rate between the reference (REF) and alternative (ALT) allele, with the goal of identifying SNP-driven changes in the Cas9 scission profile (Fig. 4a). Finally, we generated a HiPlex BreakTag dataset of 300 sites with SNPs targeting the reference or mutant allele (hereafter referred to as the "HiPlex3" library) (Figure 2.17A).

Our empirical data revealed that at position 17, [A/C/T] -> G SNPs increased the blunt rate, indicating that the ALT allele had a higher proportion of blunt cuts for the same locus (Figure 2.17B-D). Conversely, a G -> [A/C/T] SNP at position 17 decreased the blunt rate, indicating that the ALT allele had a higher proportion of staggered cleavage (Figure 2.17B-D). Analysis of position 18 revealed a strikingly opposite pattern, with [A/T/C]>G substitutions significantly decreasing the blunt rate and strongly associated with staggered DSBs, whereas G>[A/T/C] changes were significantly associated with blunt breaks (Figure 2.17E-G). Altogether, these results indicate that SNPs can change Cas9 scission profile in a position and SNP identity-dependent manner,

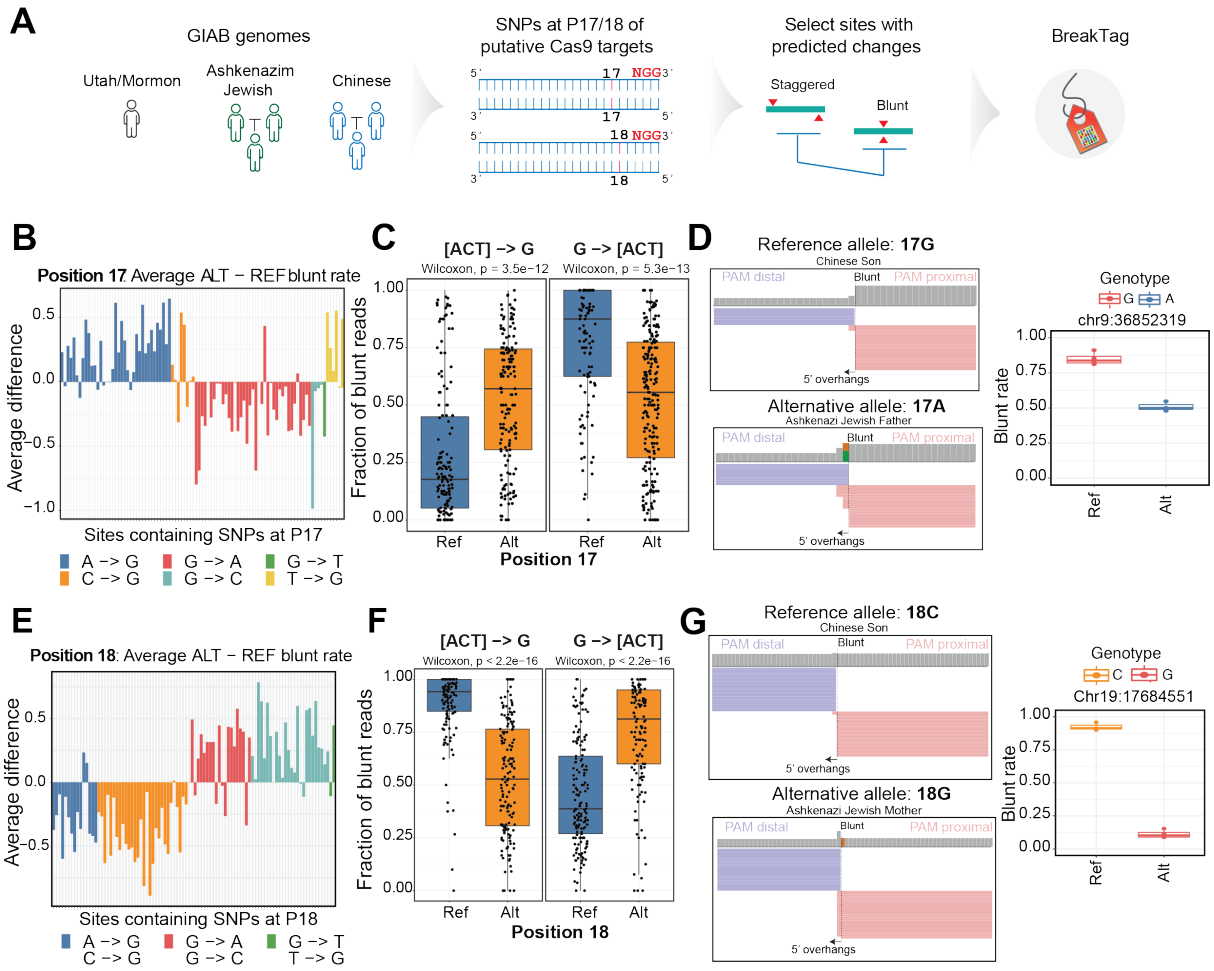


Figure 2.17. Common human genetic variation alters scission profile in a SNP and position-dependent manner. (A) Experimental design for investigating the role of human genetic variation in Cas9 scission profile. SNP databases curated from individuals of the Genome in a Bottle (GIAB) consortium (Zook et al., 2016, 2019) were used to identify Cas9 target sites containing an SNP at position 17 or 18 of a potential protospacer. Using our ML model, the blunt rate was predicted for the reference and alternative allele for each site, and sites with the highest predicted changes for position 17 or 18 were targeted using HiPlex BreakTag (HiPlex library 3) with a total of 600 sgRNAs. (B) Difference between the average blunt rate of alternative (ALT) and reference (REF) alleles containing an SNP at position 17 of the protospacer. The blunt rate was averaged between individuals with the same genotype. (C) Fraction of blunt reads out of the total number of reads in the PAM-proximal strand for the target sites containing an SNP in position 17, comparing the reference (blue) and alternative (orange) alleles; left: SNPs mutating into a G; right: SNPs mutating from a G. (D) Left: a representative IGV snapshot showing BreakTag reads of individuals harboring

the reference allele (17G, top), and an SNP (17A, bottom). Right: the blunt rates for the reference and alternative genotypes for that locus. (E) Difference between the average blunt rate in alternative (ALT) and reference (REF) alleles containing an SNP at position 18. The blunt rate was averaged between individuals with the same genotype. (F) Fraction of blunt reads over the total number of reads in the PAM-proximal strand for the target sites containing an SNP in position 18, comparing the reference (blue) and alternative (orange) alleles; left: SNPs mutating into a G; right: SNPs mutating from a G. (G) Left: a representative IGV snapshot showing BreakTag reads for individuals harboring the reference allele (18C, top), and an SNP (18G, bottom). Right: the blunt rates for the reference and alternative genotypes for that locus.

2.9.2.Single-nucleotide polymorphisms influence Cas9 scission profile and editing outcome

Given the marked effect of common human genetic variation on scission profile, we hypothesized that SNP-driven changes in Cas9 cutting have the potential to change the editing outcome in an allele-specific manner. To test that, we leveraged our gRNA-target approach to assess the DNA repair outcome at SpCas9 target sequences with a SNP found at position 17 or 18 of the protospacer with changes in scission profile as determined by BreakTag (Figure 2.18A).

As expected by the strong association between the nucleotide identity at position 17 and 18 with the SpCas9 scission profile, we observed changes in the DNA repair outcome in a SNP-depending manner. The SNPs that increased the proportion of staggered cuts at position 17 (G -> [A/C/T]) and position 18 ([A/C/T] -> G) also increased the insertion rate, as measured by the proportion insertion indels compared to deletions (Figure 2.18B, C). Conversely, the SNPs associated with a decreased proportion of staggered cuts at position 17 ([A/C/T] -> G) and position 18 (G -> [A/C/T]) also decreased the number of insertions compared to deletions (Figure 2.18B, C).

To confirm these findings, we targeted endogenous loci containing a SNP at position 17 or 18 of the protospacer with known scission profiles in lymphoblastoid cell lines from B lymphocytes derived from GIAB donors, and performed targeted ultra-deep sequencing ($\sim 10^6\times$) to assess the indel landscape of the polymorphic loci (Figure 2.18D). As an example, a G>A substitution at position 17, that is associated with a higher proportion of staggered cuts led to an increased frequency of +1 indels from 12% to 72% (fisher test, $p<2^{-16}$) (Figure 2.18E) whereas a C>G substitution at position 18, which also favors staggered Cas9 cuts (Figure 2.18F), greatly increased the frequency of +1 indels from 25% to 75% (fisher test, $p<2^{-16}$) (Figure 2.18F, G).

Taken together, our data demonstrate that genetic variation directly impacts the Cas9 scission profile along with the editing outcome, highlighting the importance of implementing variant-aware analyses of the Cas9 scission profile for more predictable and precise genome editing.

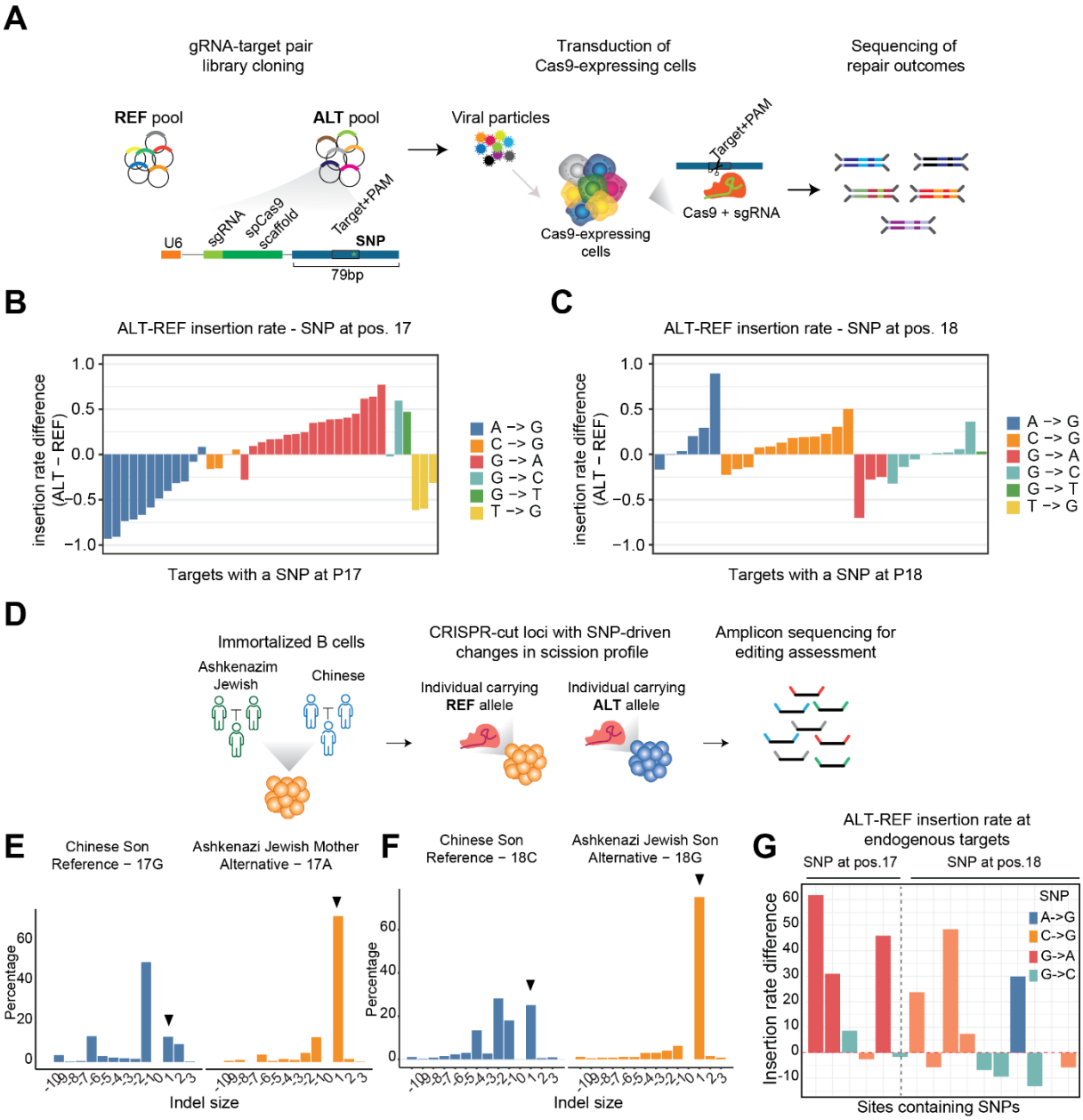


Figure 2.18. SNP-dependent changes in indel outcome profiles. (A) Schematics of gRNA-target pair experimental design for the ALT and REF pools. (B) Difference in the insertion rate of target sites containing the indicated SNPs at position 17. Targets with at least 100 mutated reads were used for the analysis. (C) Difference in the insertion rate of target sites containing the indicated SNPs at position 18. Targets with at least 100 mutated reads were used for the analysis. (D) Schematics of the experimental design for targeting the REF and ALT allele-containing GIAB donor B cells. (E) Indel size distribution of the targeted locus containing an SNP at position 17 as shown in panel G. Indels of sizes between −10 and +3 were used for this analysis. Arrow heads indicate +1 indels. (F) Indel size distribution of a locus containing an SNP at position 18 as shown in panel J. Indels of sizes between −10 and +3 were used for this analysis. Arrow heads indicate +1 indels. (G) Difference in the insertion rate of target sites containing the indicated SNPs at position 17 or 18. Positive values indicate an increase in the insertion rate in the ALT allele, and negative values indicate a decrease in the insertion rate in ALT allele compared to REF.

2.9.3.Rational editing of pathogenic deletions for gene correction

To estimate how the newly acquired insights into the scission profiles of Cas9 and variants can be leveraged for the correction of pathogenic deletions, we employed our models trained on HiPlex BreakTag data from SpCas9 or LZ3Cas9 to predict the scission profile of 1nt pathogenic deletions cataloged in the ClinVar database. Our goal was to estimate the potential of using SpCas9 or LZ3 variant-induced 1nt templated insertions to correct pathogenic deletions by restoring the frame and the original amino acid sequence, rescuing protein. In addition to SpCas9, we chose the LZ3Cas9 variant as it exhibits distinct scission profile sequence determinants that lead to higher insertion rates compared to SpCas9 at 19G loci (Figure 2.16E).

From the 31,010 pathogenic single-nucleotide deletions found in exons cataloged in ClinVar, 8,705 were endowed by a NGG PAM, and can be targeted by SpCas9 and LZ3. A total of 4,999 NGG-endowed alleles were predicted to be restored if a templated insertion takes place, rescuing the healthy protein sequence (Figure 2.19A, B). We then predicted the blunt rate of gRNAs targeting the candidate deletions for reframing and protein rescue using our model trained on SpCas9 and LZ3. We observed that 2,276 alleles were predicted to be cut preferably staggered by SpCas9, and 2,582 by LZ3. From the staggered alleles, 938 and 1,212 were predicted to be cleaved in a highly staggered manner by SpCas9 and LZ3, respectively, suggesting that templated insertions would be highly favored (Figure 2.19B). From the highly staggered alleles, we observed that 321 were shared between both nucleases, but the majority was variant-exclusive (607 for Cas9, 865 LZ3, in total 1,793 unique target sites), indicating different sequence determinants expand the number of target sites that could be cleaved in a highly staggered manner for favoring templated insertions (Figure 2.19C). We confirmed that pre-selection of target sites in which Cas9 induces staggered breaks compared to blunt, increases the frequency of templated +1 insertions that could be used to rescue 39 pathogenic single nucleotide deletions cataloged in ClinVar (Figure 2.19D) using the cellular assay used before (Figure 2.14B). As anticipated, the insertion rate and the frequency of templated insertions over all +1 indels was significantly enriched in the subset of target candidates predicted to be cut highly staggered compared to highly blunt ($p=8.6^{-08}$) (Figure 2.19E-G) demonstrating, as proof of principle, that preselection of target sites in which Cas9 cuts staggered can be used to correct clinically relevant pathogenic deletions. Among those corrected deletions, a single nucleotide deletion (ClinVar rs2077957264) in exon 1 creates a premature translational stop signal (p.Leu24*) in the TRMU gene, which has been reported to be associated with Acute Infantile Liver Failure 11, and a gRNA targeting the deletion was predicted to be cut in a highly staggered manner (Figure 2.19H). Upon targeting this deletion, we observed that most indels were insertions (Figure 2.19H), with the vast majority being templated insertions. The

inserted base would recover the frame and the original amino acid sequence, disrupting the stop codon and recovering the original protein sequence (Figure 2.19I).

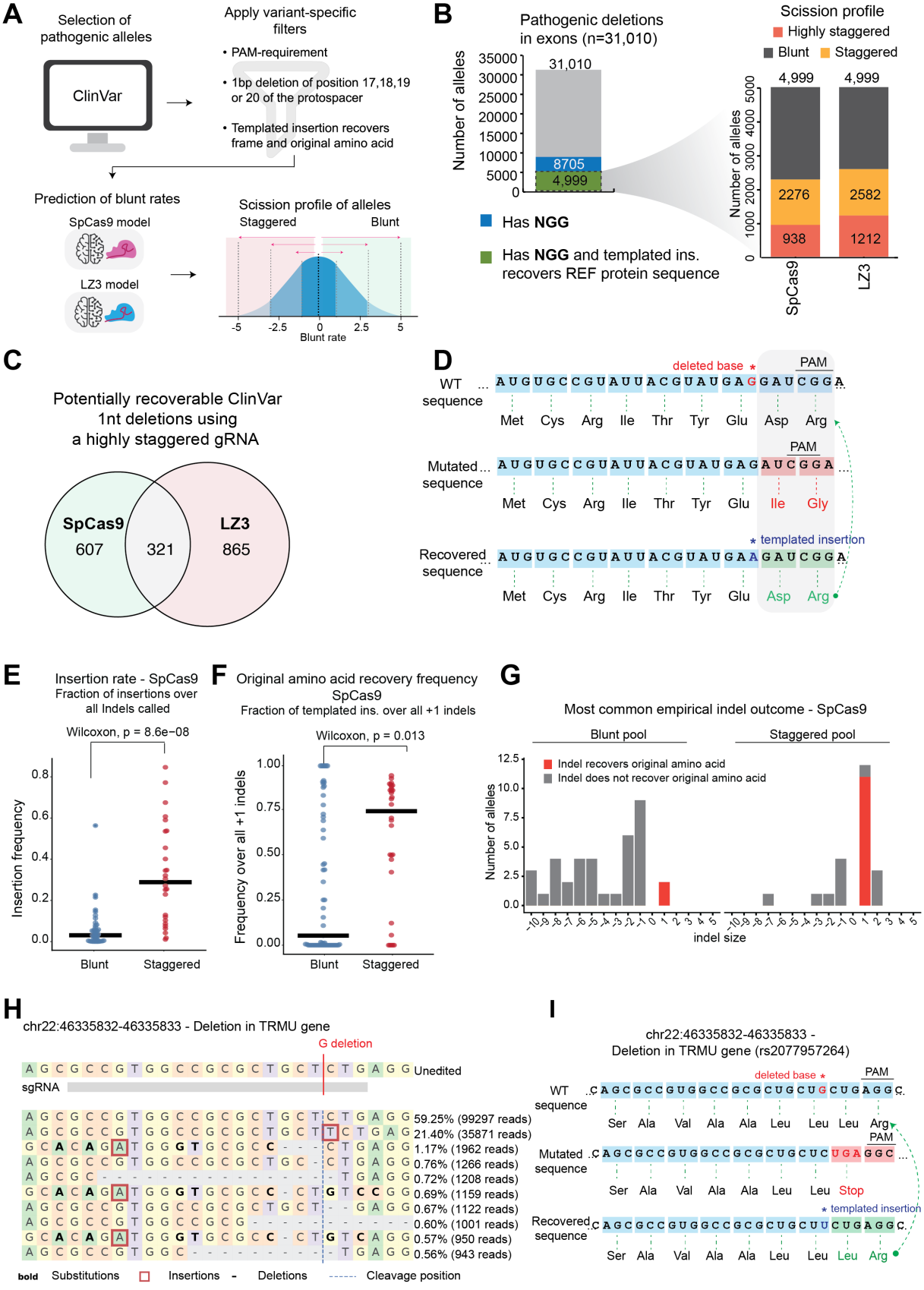


Figure 2.19. Cas9 variants expand the pool of pathogenic alleles amenable for correction. (A) Schematics depicting the workflow for the prediction of scission-aware targeting of pathogenic deletions. (B) Barplot (left) show the number of pathogenic deletions in exons that contain an NGG (blue) or that

contain an NGG and a templated insertion recovers the reference protein sequence and frame (green). Horizontal barplots (right) show the predicted scission profile of gRNAs targeting pathogenic deletions with LZ3 or SpCas9. Blunt indicates gRNAs with blunt rate >0, staggered < 0, and highly staggered ≤-2. (C) Venn diagrams depicting the overlap between pathogenic alleles that are predicted to be cleaved in a highly staggered manner by LZ3 or SpCas9. (D) Example of 1nt deletion generating a frameshift mutation, and a templated insertion rescuing the frame and original amino acid sequence. (E) Insertion rate of pathogenic 1nt deletions predicted to be cleaved in a blunt or staggered manner. (F) Rate of original protein sequence recovery, as measured by the frequency of templated insertions (i.e, duplication of the base found at position 17 of the protospacer) over all +1 indels. (G) Most common indel outcome for alleles in the blunt or staggered pool. (H) Table depicting the top 10 repair outcome alleles after targeting the 1nt deletion in the TRMU gene with SpCas9. (I) Example of a pathogenic allele in the staggered pool. The 1nt deletion generates a stop codon in the TRMU gene.

Taken together, we developed a framework for characterization of Cas9 nucleases' activity, fidelity and scission profile. We characterized we characterized the Cas9 endonuclease scission profile and identified that the sequence of CRISPR–Cas9 target sites, human genetic variation and alternative Cas9 variants are three principal influencers of Cas9 cleavage pattern and, therefore, of gene editing outcomes. We found that staggered cleavage by Cas9 is non-random, and highly influenced by the sequence composition of the seed portion of the target site. Furthermore, we demonstrate that a scission-aware gRNA design can be leveraged for increased gene editing precision of compensating insertions at alleles carrying pathogenic deletions.

Chapter 3. Discussion

3.1.BreakTag expands the toolkit of off-target discovery

The safety of CRISPR-based therapy is paramount for the success of gene editing efforts, and off-target assessment has become standard practice in designing CRISPR-Cas gene editing strategies. Here we designed and implemented BreakTag, a high-throughput and easily scalable method for the nomination of off-targets and CRISPR-Cas nuclease characterization.

Several methods have been developed to assess gRNA/nuclease fidelity, differing mostly at the DSB enrichment step. In SITE-seq, high molecular weight DNA is *in vitro* digested with the ribonucleoprotein (RNP), and a biotinylated adapter is ligated to the Cas9-generated DSB. A subsequent pull-down with magnetic streptavidin beads allows selective enrichment of DSBs for library preparation (Cameron et al., 2017). CIRCLE-seq (Lazzarotto et al., 2018; Tsai et al., 2017) and CHANGE-seq (Lazzarotto et al., 2020) rely on the circularization of genomic DNA, followed by selective linearization of circles containing a target sequence by cleavage using RNPs. The linear DNA molecules are enriched for library preparation, yielding a sensitive representation of off-targets. However, the circularization step is relatively inefficient, CIRCLE-seq requires large amounts of large material (~25 µg of DNA per reaction (Lazzarotto et al., 2018)) and may preclude off-target discovery on rare samples. CHANGE-seq bypasses this issue by deploying a tagmentation-based circularization, requiring significantly less DNA (1 µg per reaction (Lazzarotto et al., 2020)). In these methods, the DSB enrichment step is a stand-alone reaction in the workflow, requiring extra nucleic acid purification steps that increase sample handling time. BreakTag enriches DSBs using a semi-suppression approach of non-cut molecules where DNA DSB ends are selectively enriched during the PCR step (see 2.1). This PCR-embedded enrichment eliminates the requirement of separate steps for enrichment of DSBs, yielding a fast protocol with fewer enzymatic reactions and DNA clean-up steps, reducing the starting material necessary for successful library preparation. Furthermore, BreakTag can be easily automated with a liquid handling platform, greatly expediting off-target discovery and nuclease characterization (Longo et al., 2024).

We further demonstrate that BreakTag can be used to investigate DSB repair dynamics. In a proof-of-principle experiment, we transfected photocleavable gRNAs (pcgRNAs (Zou et al., 2021)) to Cas9-expressing cells. Following a short UV pulse delivery, the gRNA is cleaved and Cas9 can no longer cut and therefore DNA repair is synchronized. By performing BreakTag at different

timepoints, we demonstrate that the DSB break signal is reduced over time (Figure 2.1D). The cell-based format might be used to map off-targets in a physiological context, but further experimentation is necessary for benchmarking against *in cellulo* methods. On the other hand, the cell-free format allows single-nucleotide resolution of DNA DSB ends since there is no ongoing DNA repair in deproteinated genomic DNA. We demonstrated this by rapidly mapping the off-target of 46 clinically-relevant sgRNAs in cell-free mode with a wide range of off-target activity. Furthermore, unlike earlier *in vitro* methods that lack an "end-repair" step, cell-free BreakTag can be applied for off-target investigation of Cas12a, a staggered-cleaving nuclease, without changing the workflow.

We developed BreakInspectoR for BreakTag data analysis. The pipeline pre-processes the sequencing reads and maps them to a reference genome. Sites that shared homology to a target sequence (up to 7 mismatches) and found next to a PAM sequence are selected, and the number of unique breaks is counted. An enrichment test is then performed between the test sample versus the non-target control (where no Cas9-induced DSBs are expected). The pipeline provides users with meaningful outputs for sgRNA activity investigation and might be applied to other tools such as sBLISS (Bouwman et al., 2020; Gothe et al., 2019; Yan et al., 2017).

Base editors are a fusion of Cas9 nickase with a cytosine or adenine deaminase. This class of gene editors work by directly modifying the target locus without the need of a DNA DSB, generating a nick on the target strand and a base deamination on the non-target strand (Anzalone et al., 2020; Gaudelli et al., 2017; Komor et al., 2016). Although they are regarded as a safer alternative to Cas9 because no cleavage is required, detrimental effects such as off-target (Lei et al., 2021; Slesarenko et al., 2022) and genotoxicity (Fiumara et al., 2023) have been reported. BreakTag can only map free DSB ends and, in principle, base editor off-target mapping is precluded. We adapted the BreakTag workflow to induce artificial DSBs at sites of base editing *in vitro* by performing a base excision repair reaction as previously described (D. Kim et al., 2021; Liang et al., 2019), and demonstrated that BE off-target nomination is feasible with a simple protocol adaptation. However, further experimentation will be necessary to benchmark off-target nomination against similar methods tailored to BEs.

The relatively poor overlap between off-targets for the same gRNA nominated by different in vitro methods is well recognized in the field, and it can be accounted for by differences in cell type, ploidy of the locus, sequencing depth and DSB-enrichment strategies (Atkins et al., 2021; Cromer et al., 2023). We benchmarked BreakTag against *in vitro* and *in cellulo* tools using publicly available data and experiments performed by us across 3 different gRNAs. We observed that the majority of *bona fide* off-targets (i.e, those that had an indel) were mapped by two or more *in in*

vitro tools, suggesting that using different methods of off-target nomination is an interesting strategy to ensure all true off-targets are mapped. Importantly, we observed a strong correlation between the number of off-targets called by BreakTag and CHANGE-seq across 46 gRNAs (Lazzarotto et al., 2020). It is worth pointing out that the additional "end-repair" step in the BreakTag allows the enrichment of those that are cut in a highly staggered manner, influencing the pool of off-targets mapped. One disadvantage is the relatively high background compared to in cellulo methods, since DSBs generated by intrinsic cell processes (such as transcription, replication) and mechanical breaks generated during DNA extraction are also sequenced with BreakTag and might mask off-targets that fall in those regions. We circumvent this issue by performing an enrichment test against a non-targeting control of similar sequencing depth, filtering out potential sequencing artifacts and highly fragile sites. Compared to GUIDE-seq (Tsai et al., 2015), an *in cellulo* method, BreakTag nominated a larger list of off-targets across 27 gRNAs. In vitro methods rely on the digestion of proteinase-treated genomic DNA with ribonucleoproteins and by default provide a larger list of off-targets than *in cellulo* assays, since the chromatin accessibility can protect from Cas9 cleavage (Horlbeck et al., 2016; Lazzarotto et al., 2020; Yarrington et al., 2018). Nonetheless, we observed a good correlation between the number of off-targets nominated by GUIDE-seq and BreakTag across matching gRNAs (Figure 2.4D).

3.2.Increasing throughput with HiPlex BreakTag

We reasoned that BreakTag throughput can be greatly increased by combining several gRNAs in the same reaction. We generated HiPlex BreakTag by performing sgRNA production in pooled format. In our HiPlex library 1, previously reported gRNA sequences (Chakrabarti et al., 2019) were used to generate 10 pools containing approximately 150 sgRNAs each. Genomic DNA is digested with Cas9 loaded with the gRNA sequence and based on sequence homology, BreakInspectoR assigns off-targets to their parental gRNAs. This approach increases the number of on- and off-targets inflicted per sample, allowing the production of a robust dataset for further analysis as demonstrated here. We envision that the gRNA pooled format can be used to generate data amenable for training models where a large amount of cleavage data is necessary, such as XGScission for dissecting Cas nucleases' sequence determinants for scission profile. Further experimentation is required to investigate the limit number of gRNAs per pool in HiPlex mode, and we anticipate that optimization might be necessary if more gRNAs per pool are used. Moreover, the sgRNAs in HiPlex mode are generated using a T7 *in vitro* transcription reaction that might introduce biases due to transcription sequence preferences by the RNA polymerase and therefore, singleton should be the mode of choice for off-target discovery.

BreakTag might assign a given off-target to multiple parental gRNAs if it contains >6MMs based on homology alone. To help users design HiPlex pools, we developed a script that automatically sorts gRNA sequences into different pools based on hamming distance, mitigating wrong assignment of off-targets to a parental gRNA (https://github.com/roukoslab/breakinspectoR).

We leveraged out robust dataset to investigate factors that influence off-target activity. We observed that mismatches between the crRNA and the target DNA sequence show a preference to accumulate at PAM-distal portions of the target sequence, and gradually decrease towards the PAM portion, as shown by others (Jr et al., 2021; Lazzarotto et al., 2020). Mismatches in the seed portion disrupt R-loop formation and consequently ablate cleavage by Cas9 (Jiang et al., 2016) and the accumulation of mismatches decreases cleavage efficiency. We observed that although most off-target sequences were endowed by the classical NGG PAM, non-canonical PAMs such as NAG and NGA activated cleavage, albeit with lower frequency, agreeing with previous data that showed that non-NGG PAMs greatly reduce Cas9 binding (Jr et al., 2021). Finally, the observed inverse correlation between the target sequence complexity and the number off-targets mapped indicate that targeting complex regions is an interesting strategy for mitigating unintended editing,

3.3.Scission-aware profiling of Cas9-mediated DSBs

The scission profile remained an under characterized and overlooked aspect of Cas9 biology. Previous reports demonstrated that the RuvC of SpCas9 can cut at upstream positions of the HNH, generating staggered DSBs with 5' overhangs (Chauhan et al., 2023; Jinek et al., 2012; Jr et al., 2021; Lemos et al., 2018; Molla & Yang, 2020; Shi et al., 2019; Shou et al., 2018; Slaman et al., 2023; Stephenson et al., 2018). Thus far, available evidence was limited to few gRNAs, and systematic investigation of Cas9 scission profile has been precluded due to lack of tools that can map DSB ends at scale. Furthermore, we adapted BreakInspectoR to parse out the read 1 start site for the PAM-proximal and PAM-distal side of the break, providing a strand resolution of the DSB ends. This modification allowed us to investigate at scale the different types of cuts generated by Cas9, by calculating the distance between the *expected* cut site for a blunt break versus the *real* cut site after DSB end processing. Because 5' overhangs are filled in and 3' overhangs are chewed back, the difference between *real vs. expected* cut site coordinate indicates the size and identity of the staggered cut. We observed that SpCas9 has a preference for blunt cuts as previously demonstrated, but 5' overhangs of 1-3 nucleotides represented nearly 1/3 of all cleavage sites. Moreover, the scission profile was a gRNA-dependent effect. The majority of gRNAs showed a mixture of blunt and staggered SpCas9-mediated cuts, and a smaller portion showing complete

blunt or staggered cleavage. This indicated that the scission profile is not random, but rather target-specific.

Using NUCLEA-seq, a previous study systematically tested the effect of mismatches (MMs) across every position in the protospacer between the crRNA and the target DNA sequence over three gRNAs, and observed that MMs shifted the RuvC domain cleavage site, generating staggered cleavage (Jr et al., 2021). A study elucidated the mechanism of MM acceptance of Cas9 by determining the crystal structure of the nuclease coupled with mismatched substrates (Pacesa et al., 2022). The study found that the formation of noncanonical base pairs induce rearrangements on the REC2/3 domain of Cas9 that enable cleavage. We observed here that on-target sequences had a preference for blunt cleavage compared to off-targets with up to 7 MMs, presumably due to similar conformational changes in Cas9 imposed by PAM-distal MMs that might shift the RuvC nicking site. Regardless of the underpinning mechanism, our observation that off-targets display a higher frequency of staggered cleavage suggests that methods that do not contain an "end-repair" step might miss *bona fide* off-targets cleaved in a highly staggered manner.

Using XGScission, we identified a core region in the seed (positions 16-20 of the protospacer) with a marked importance for predicting the blunt rate. The same core positions were identified in other studies where machine learning models were trained on CRISPR repair products in order to predict the indel outcome based on the target sequence (Allen et al., 2019; Chakrabarti et al., 2019; W. Chen et al., 2019; Leenay et al., 2019). Moreover we identified strong sequence determinants, with a G base symmetry at the cut site for blunt and staggered cleavage: The G base at position 17 of the protospacer favored a blunt cut, whereas the G at position 18 favored a staggered cut. We anticipate that XGScission will be useful to predict SpCas9-gRNA scission profile for different applications, especially in the context of favoring insertions over deletions.

3.4.Characterization of Cas9 variants cleavage activity

Given the robustness and scalability of nuclease activity investigation with BreakTag, we anticipated that the platform could be applied to test the activity of engineered Cas9 variants, from cleavage efficiency to scission profile.

A common strategy for the mitigation of off-target activity is the engineering of Cas9 variants by introducing mutations in different domains, with the goal of reducing gRNA:DNA mismatch acceptance (see 1.4.5). However, one common drawback is the reduction of overall cleavage activity depending on the residue being targeted, hindering an enzyme that generates fewer off-target loci, but also has lower on-target cleavage efficiency (Allemailem et al., 2023; Kulcsár et al.,

2022). We demonstrated that the data-rich BreakTag output can be applied to the multi-level characterization the following Cas9 variants: HiFiCas9 (Vakulskas et al., 2018), xCas9 (Hu et al., 2018), SniperCas9 (J. K. Lee et al., 2018), HypaCas9 (J. S. Chen et al., 2017), EvoCas9 (Casini et al., 2018), and LZ3Cas9 (Schmid-Burgk et al., 2020). We observed a clear trade-off between *specificity* and *activity,* for example, the variant EvoCas9 showed ~40% increase in specificity, but the activity was ~50% that of SpCas9. These results are in line with another study where the activity and specificity of the same variants was tested using TTISS-seq, an *in cellulo* method, and a trade-off with similar levels was reported (Schmid-Burgk et al., 2020), demonstrating that the BreakTag readout is an accurate surrogate of variant activity in cells.

We identified the variant LZ3 (Schmid-Burgk et al., 2020) as a high fidelity enzyme with a remarked reduction of blunt rate correlation with SpCas9. The blunt rate distribution for LZ3 skewed towards staggered cuts, and using XGScission, we identified that the staggered determinant was a 19G, unlike SpCas9. Coupled with this finding, the LZ3 variant showed a different insertional profile compared to SpCas9, and the authors identified a preference for +1 indels at 19G (Schmid-Burgk et al., 2020). The variant contains 4 point mutations in different domains: N690C (recognition lobe), T769I (linker 1), G915M (linker 2) and N980K (RuvC domain) that might confer higher specificity and/or different sequence determinants. Another study identified that a G915F mutation changed the scission profile of Cas9 when translocation junctions were analyzed (Shou et al., 2018). Interestingly, the Gly915 residue of SpCas9 interacts with position 18 of the protospacer (Jiang et al., 2016), indicating that the interaction between the Linker2 domain and the protospacer mediates flexible cleavage. These results indicate that modulation of scission profile and repair outcome can be achieved by engineering new Cas9 variants, and we envision that BreakTag-BreakInspectoR-XGScission will be used for the rapid characterization of new engineered nucleases (Figure 3.1).

Figure 3.1. Workflow for the multiscale characterization of engineered Cas9 variants with BreakTag suite of tools.

3.5.Scission profile as a major determinant of indel outcome

The DNA repair landscape of Cas9-induced DSBs is not random, and can be predicted solely based on the spacer sequence. By analyzing the repair outcome of over 200 sites targeted with Cas9 in human cells, Van Overbeek and colleagues (van Overbeek et al., 2016) demonstrated that the indel landscape and the repair outcomes are reproducible and non-random. Using yeast as a model, Lemos and colleagues further demonstrated that SpCas9 creates 1nt 5' overhangs that are filled in before DSB end ligation, generating templated insertions (Lemos et al., 2018). These findings were reproduced in mammalian cells (Gisler et al., 2019; Molla & Yang, 2020; Shi et al., 2019; Shou et al., 2018), reinforcing the notion that cleavage patterns influence the indel landscape. Finally, machine learning models trained on massive Cas9-mediated DSB repair data revealed that the repair outcome is highly predictable by the target sequence (Allen et al., 2019; Chakrabarti et al., 2019; W. Chen et al., 2019; Leenay et al., 2019; Shen et al., 2018; Taheri-Ghahfarokhi et al., 2018). Although a clear mechanistic link between scission and repair outcome was demonstrated, this direct relationship has not been tested at scale. Thus far, the frequencies of different types of DSB ends generated by Cas9 remain unexplored.

We observed a clear preference for staggered cuts being repaired as single-nucleotide insertions, whereas blunt cleavages generated random indel sizes. Interestingly, the repair outcome of staggered cuts was predictable. Those staggered sites repaired as insertions were generally templated insertions, with the overhang sequence being filled in and upon ligation of DSB ends, a duplication of the base found at position 17 of the protospacer was generated. As a complementary effort, we devised a gRNA-target pair strategy generated by others (Allen et al., 2019). The strategy relies on the parallel cloning of the target sequence + gRNA in the same cassette, and upon transfer to a lentiviral vector, Cas9 expressed cells can be infected. The repair outcome can be assessed in parallel with the same primer pair, bypassing the need to design a primer-pair for every loci analyzed as in the endogenous strategy. In line with our previous findings, the analysis revealed a preference for templated insertions (i.e, duplication of the base found at position 17 of the protospacer) for those target sites predicted to be cleaved in a staggered manner by SpCas9.

Given the predictability of indel landscape by the target sequence, machine learning models have been devised to extract important features for insertions and deletions (Allen et al., 2019; Chakrabarti et al., 2019; W. Chen et al., 2019; Leenay et al., 2019; Shen et al., 2018; Taheri-Ghahfarokhi et al., 2018). In terms of sequence determinants, the models largely agreed that a 17G locus favors a deletion, whereas an 18G locus favors insertions (Leenay et al., 2019; Shen et al., 2018). In line with these observations, our XGScission sequence determinant analysis for flexible scission agrees with these previous data, further confirming the intrinsic link between staggered

cleavage and precise and predictable insertions. It is worth pointing out that the analysis of scission profiles is derived from DSB end patterns, constituting a preliminary stage preceding the DNA repair outcome in the cascade of events from Cas9 cleavage to indel results. The indel landscape is shaped by different DNA repair pathways influenced by the chromatin environment (Schep et al., 2021; Xue & Greene, 2021), which might account for the slight deviation in sequence determinants of insertions and deletions identified by computational predictors trained on repair outcome data compared to cleavage determinants identified by BreakTag (Figure 3.2). While we demonstrate a strong link between the scission profile and the repair outcome, suggesting that the structure of DSB ends significantly influences the size of indels, we do not anticipate that the scission profile-trained model will outperform existing *in silico* models at predicting indel outcomes. Moreover, the models trained on repair outcomes revealed that deletions are also predictable base on microhomology surrounding the cut site, repaired by microhomology end joining (MMEJ). Although we observed a preference for blunt cuts being repaired as deletions, it remains to be assessed if adding scission profile information to the models can increase the predictability of deletion outcomes.

Manipulating DDR factors offers a promising approach to increase the frequency of templated insertions at staggered DNA cuts. It has been shown that KU-60019, an ATM inhibitor, significantly increases the frequency of single-nucleotide insertions (Bermudez-Cabrera et al., 2021). Hussman and colleagues used Repair-seq to map the effect of 476 DDR genes in the distribution of indels, and identified that knockdown of Pol λ and Pol θ decreased the percentage of insertions, suggesting that these polymerases mediate this indel class (Hussmann et al., 2021). Additionally, Pol λ has been identified as the primary polymerase for filling in 5' overhangs at Cas9 staggered cuts (Mehryar et al., 2023), with similar findings in a budding yeast model for Pol4, a human Pol λ homologue (Lemos et al., 2018). Moreover, fusing Cas9 with DNA polymerases has been shown to increase templated insertion rates at staggered cuts (Nakade et al., 2022; Yang et al., 2023). Collectively, this body of work demonstrates that the frequency of templated insertions can be enhanced at staggered cuts through various methods, and our new knowledge on the determinants of scission profiles can further advance these strategies

While adapting the indel-trained machine learning model to new Cas9 variants is feasible, it can be cumbersome due to the necessity for time-consuming cloning, the generation of Cas9 variant-expressing cell lines, and the design of costly large amplicon panels for sequencing endogenous targets. In contrast, our biochemical assays provide a comprehensive characterization of Cas9 nucleases at multiple levels. This data-rich approach allows us to assess variant fidelity, activity,

and assess determinants of nuclease scissions within a single assay, offering a broad view of nuclease function and accelerating the characterization of new nuclease variants.

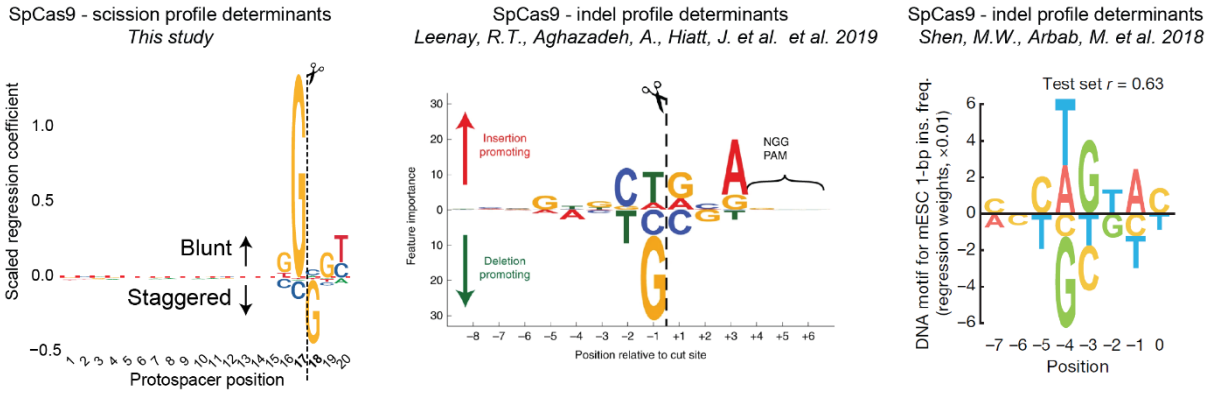


Figure 3.2. Comparison of sequence determinants for SpCas9 cleavage profile (this study) and insertions/deletions (Leenay et al., 2019; Shen et al., 2018).

3.6.Common human genetic variation affects gene editing precision

CRISPR-Cas gene editing therapy has been largely designed on a human reference genome, and therefore, fails to account for common human genetic variation. It is estimated that a typical genome differs from the reference genome at 4.1-5 million sites, and ~99.9% of this variation is explained by single nucleotide polymorphisms (SNPs) and short indels (1000 Genomes Project Consortium et al., 2015). In the context of targeted gene editing, genetic variation can be seen as a double-edge sword. On one hand, genetic variation can be explored for allele-specific targeting (Keough et al., 2019), since mismatches between the target DNA and the sgRNA in the seed portion will disrupt cleavage. On the other hand, genetic variation might hamper the success of CRISPR-based therapy by generating neo off-targets at polymorphic sites (Lessard et al., 2017). A recent study demonstrated that a gRNA targeting the BCL11A enhancer – a target for curing sickle cell disease and β-thalassemia, has a population-dependent new off-target produced by an allele common in African-ancestry populations, illustrating how genetic variation affects the off-target landscape (Cancellieri et al., 2022).

Applying a variant-aware strategy for gRNA design is desirable strategy to account for human genetic variation in gene editing. For example, after empirically assessing the off-target landscape of a given CRISPR-based therapy in cell models, participants could be included or excluded based on the presence of the variant in the participant's genome that generates neo off-targets (Saha, 2023). Alternatively, given the scalability of BreakTag, the off-target landscape of new gene editing therapies can be tested in several genomes in parallel.

Here we uncovered another important role of genetic variation in gene editing therapy by altering scission profile and the indel landscape in a SNP and position-dependent fashion. We compared the insertion rate at sites with changes in scission profile and observed that SNPs that increased the proportion of blunt cuts, decreased the insertion frequency when the DNA repair outcome was assessed. In line, SNPs that increased the proportion of staggered cuts also increased the proportion of insertions at polymorphic sites. We further confirmed these findings by targeting ALT and REF endogenous alleles in immortalized B cells of the GIAB donors carrying polymorphic alleles. SNPs that increase the frequency of single-nucleotide insertions might be explored in the context of gene knockout in those participating in gene editing therapy. Although polymorphic loci tend to not be used as a target in CRISPR-Cas therapy due to the formation of gRNA:DNA seed mismatches, previous work demonstrated that the use universal base inosine in the gRNA is able to bypass the cleavage-prohibitive MM in the seed portion (Krysler et al., 2022). These data point out that beyond affecting the off-target landscape, common human genetic variation is an important determinant of scission profile, as changes in the target sequence ultimately affect scission profile, and consequently, the DNA repair outcome.

3.7.Leveraging scission profile information for the correction of pathogenic deletions

Given the gRNA-dependent predictable indel landscape, machine learning models trained on the repair outcome of Cas9-induced DSBs allow analysis of the potential for template-free correction of pathogenic alleles. This is highly advantageous, since HDR rates are low in most cells, and the formation of indels is the most common repair outcome (Xue & Greene, 2021). Using a predictor of repair outcome, previous work demonstrated that MMEJ-induced predictable Cas9-induced deletions can correct microduplicated alleles of cells from patients with Hermansky-Pudlak syndrome (HSP1 gene) and Menkes disease (ATP7A). Targeting the pathogenic microdeletion with Cas9 resulted in the restoration of the wild type sequence, and therefore, protein function (Shen et al., 2018).

The higher precision and predictability of Cas9-induced staggered breaks affords an opportunity for leveraging the formation of single-nucleotide insertions in a controlled manner. We hypothesized that templated single-nucleotide insertions may be harnessed to counteract pathogenic single-nucleotide deletions, correcting protein frame. By exploring the degeneration of the genetic code, templated insertions may be induced without changing the original protein sequence.

We assessed the potential of a scission-aware gRNA design for template-free allele correction. We demonstrated that ~57% of single-nucleotide deletions in ClinVar found next to an available PAM can be corrected without changing the original amino acid sequence. From the correctable alleles, our model predicted that ½ is cleaved in a staggered manner by SpCas9 or LZ3. We further stablished that a scission-aware gRNA design can be leveraged for favoring 1nt insertions in a proof-of-principle experiment by demonstrating the preference for templated insertions in a pool of ClinVar alleles predicted to be cleaved in a staggered manner. Although we focused on alleles where a templated insertion maintains the codon identity and corrects the frame, some proteins can tolerate single amino acid substitution without affecting protein structure, and the pool correctable pathogenic single-nucleotide deletions might be larger than we estimate here (Cheng et al., 2023). Another study demonstrated that scission-aware gRNA design for the correction of pathogenic deletions is an achievable goal. The authors targeted the CAPN3 c.550delA allele in cells of patients with limb girdle muscular dystrophy and demonstrated that a 1nt templated insertion generated by a 1nt staggered cut by Cas9 restored protein function (Müthel et al., 2023).

We demonstrated that a scission-based gRNA design can be harnessed for predicting candidate pathogenic deletions, ~75% of pathogenic deletions in ClinVar do not have an available PAM for SpCas9 (NGG) and therefore cannot be targeted. Moreover, the intrinsic sequence determinants of Cas9 for staggered cleavage limits which loci will be cut in a highly staggered manner. In the context of pathogenic deletions, a high fidelity, sequence-agnostic staggered and PAMless Cas9 will considerably expand the pool of alleles amenable for correction. A study demonstrated that mutations in the PAM-interacting domain of SpCas9 relax PAM requirements and a new variant, SpRY, can target virtually any sequence (Walton et al., 2020). We anticipate that combining PAMless mutations with those that increase staggered cleavage will have an important role in correcting pathogenic deletions via templated insertions. We envision that BreakTag will facilitate the characterization of novel CRISPR nucleases as demonstrate here, accelerating the discovery of highly precise CRISPR nucleases.

3.8.Summary

We characterized here determinants of the Cas9 flexible scission profile the ratios of different DSB-end structures generated by the nuclease and engineered variants. We tested at scale the association between scission profile and precise and predictable insertions, demonstrating that gRNA-intrinsic scission profiles can be harnessed for the correction of pathogenic deletions. Moreover, we devised platform for the multiscale characterization of new engineered CRISPR nuclease variants.

In summary, we characterized the Cas9 endonuclease scission profile and established that the sequence of CRISPR-Cas9 target sites, human genetic variation, and alternative Cas9 variants, are three principal influencers of Cas9 cleavage pattern, and therefore, of gene editing outcomes (Figure 3.3). Our work illuminates the fundamental properties of Cas9 nuclease activity and lays the foundation for harnessing the flexible scission profile of Cas9 and engineered variants for precise, predictable and personalized genome editing.

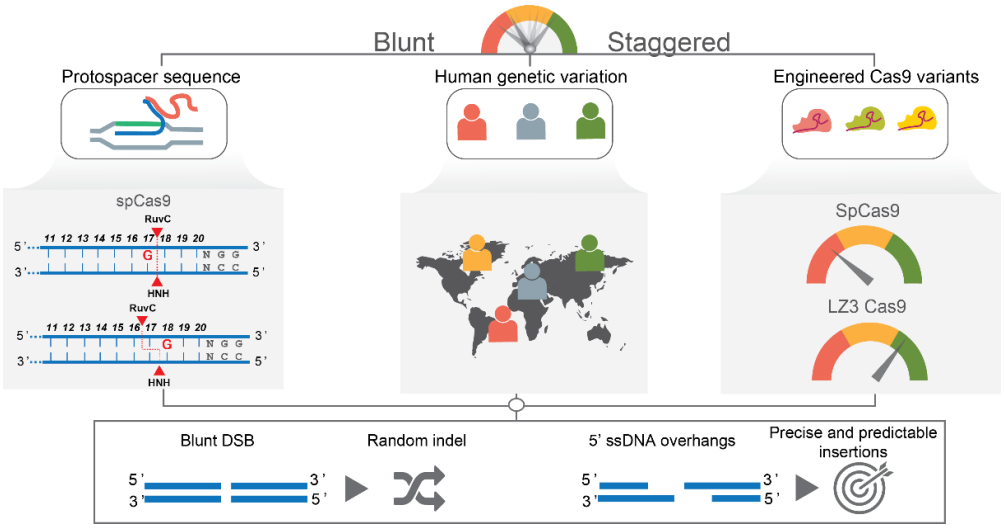


Figure 3.3. A model of the determinants of Cas9 scission profile identified using BreakTag. The protospacer sequence, human genetic variation, and engineering Cas9 variants can dictate Cas9 scission profile, which is strongly associated with precise and predictable genome editing.

Chapter 4. Material and Methods

4.1. Cell culture

Human osteosarcoma U2OS cells, human embryonic kidney cells (HEK293), and HepG2 cells were cultured in DMEM (GIBCO 41965062) supplemented with 10% FBS (PanBiotech P40-37500), 100 U/mL penicillin–streptomycin and 2 mM L-glutamine. Cell lines were maintained in a humidified incubator at 37°C supplemented with 5% CO_2.

Lymphoblastoid cell lines from B cells from the Chinese son (GM24631), Ashkenazi Jewish son (GM24385), and Ashkenazi Jewish mother (GM24143) were obtained from Coriell and maintained in RPMI (GIBCO 11875093) supplemented with 15% FBS (PanBiotech P40-37500), 1 mM sodium pyruvate (GIBCO 11360070), 100 U/mL penicillin–streptomycin (GIBCO 15140122) and 2 mM L-glutamine (GIBCO 25030081).

Table 1. Cell lines used in this study.

Cell line	**Source**
U2OS	ATCC, HTB-96
HEK293T	ATCC, CRL-3216
HepG2	Gift from Dr. Julian Koenig's Lab (IMB)
LCL from Chinese Son	Coriell, GM24631
LCL from Ashkenazi Jewish Son	Coriell, GM24385
LCL from Ashkenazi Jewish Mother	Coriell, GM24143

4.2. Genomic DNA extraction

The genomic DNA of cells was extracted using the Qiagen Blood and Tissue Kit (Qiagen 69506) following the manufacturer's instructions with the following modification: after washing the spin column with buffer AW2, gDNA was eluted in nuclease-free H_2O and stored at −20°C. Extracted gDNA was quantified using a Qubit Fluorometer and a dsDNA High Sensitivity (HS) kit (Invitrogen Q32853).

gDNA of Genome in a Bottle individuals was purchased from Coriell: Female Utah/Mormon (NA12878), Ashkenazi Jewish Son (NA24385), Ashkenazi Jewish Father (NA24149), Ashkenazi Jewish Mother (NA24143), Chinese Son (NA24631), Chinese Father (NA24694), and Chinese Mother (NA24695).

Table 2. Genomic DNA sources used in this study.

Donor	**Source**
Female Utah/Mormon	NA12878
Ashkenazi Jewish Son	NA24385
Ashkenazi Jewish Father	NA24149
Ashkenazi Jewish Mother	NA24143
Chinese Son	NA24631
Chinese Father	NA24694
and Chinese Mother	NA24695

4.3. Expression and purification of homemade Tn5

Expression and purification of hyperactive Tn5 (E54K, L372P) were performed by IMB Protein Production Core Facility as described elsewhere (Hennig et al., 2018) with the following modifications: Tn5 was expressed as an N-terminal His_6-GST fusion followed by a 3C protease cleavage site. GSH affinity purification followed PEI precipitation of nucleic acids from the

soluble lysate. The fusion protein was cleaved using a recombinant His_6-3C protease. His_6-tagged protease and GST were separated from untagged Tn5 using reverse Ni–NTA-affinity chromatography, and purification was performed as described elsewhere (Hennig et al., 2018).

4.4.Tn5 loading and BreakTag linker preparation

Tn5-B adapter was prepared by mixing 100 µM Tn5ME-B (Illumina FC-121-1031) and 100 µM Tn5MErev (Picelli et al., 2014) (Table 3) resuspended in annealing buffer (50 mM NaCl, 40 mM Tris, pH 8) at a 1:1 ratio. The oligos were annealed in a thermocycler programmed as follows:

Step	Temperature	Time
1	95°C	5 min
2	65°C	−0.1°C/s
3	65°C	5 min
4	4°C	−0.1°C/s
5	4°C	Hold

Homemade Tn5 was loaded with pre-annealed Tn5 B adapter for 1 hour at room temperature with agitation (300 rpm) in a thermoshaker.

The BreakTag linker was prepared by combining 10 µM BreakTag_fwd and 10 µM BreakTag_rev oligos (Table 3) in T4 polynucleotide kinase buffer (NEB M0201S). The oligos were annealed in a thermocycler programmed as follows:

Step	Temperature	Time
1	95°C	5 min
2	Cool to 25°C	−0.1°C/s

3	25°C	Hold

Table 3. Oligonucleotide sequences used in BreakTag.

Name	**Sequence 5'-3'**
Tn5MErev	/5Phos/CTGTCTCTTATACACATCT
Tn5ME-B	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG
BTLinker_1_rev	CGATTGAGGCCGGGTCTCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNNNNNNCTCACACGT
BTLinker_1_fwd	/5phos/CGTGTGAGNNNNNNNNCTGTCTCTTATACACATCTGACGCTGCCGACGAGACCCGGCCTCAATCGAA
N7XX	CAAGCAGAAGACGGCATACGAGAT**XXXXXXXX**GTCTCGTGGGCTCGG
S5XX	AATGATACGGCGACCACCGAGATCTACAC**XXXXXXXX**TCGTCGGCAGCGTC
/5phos/ indicates 5' phosphorylation	
Underscored sequence indicates sample barcode	
XXXXXXXXX indicates i5/i7 index sequence	

4.5. *In vitro* digestion of gDNA with Cas9 ribonucleoproteins, Cas12 and Base editors

Ribonucleoproteins (RNPs) were assembled by mixing Cas9, Cas12a or Base editor with the sgRNA at equimolar ratios in NEB 3.1 buffer (NEB B72030), followed by incubation at 37°C for 10

min. For HiPlex BreakTag, pools were mixed with the nuclease at a 2:1 ratio. An input of 500 ng of gDNA was mixed with each RNP at a final concentration of 90 nM and incubated at 37°C for 1 hour in a thermocycler with the lid set at 37°C. The reaction was terminated by adding RNase A (Thermo Scientific 10753721) and proteinase K (NEB P8107) at final concentrations of 0.8 and 0.2 μg/μL, respectively, at 37°C for 20 min, followed by incubation at 55°C for 20 min. Nuclease-digested gDNA was purified with DNA AMPure XP beads (1.2x volumes, Beckman Coulter A63881).

For Cytosine Base Editor off-target experiment, digested gDNA was incubated with the USER enzyme in 1x rCutSmart buffer for 2 hours at 37°C in a thermocycler. Followed by USER digestion, gDNA was purified using 1.2x volumes of DNA AMpure XP beads.

4.6.HiPlex sgRNA library design and production

Sequences for HiPlex1 (Chakrabarti et al., 2019) and HiPlex2 (Allen et al., 2019) pools were bioinformatically split into 10 pools. Each pool contained 150 gRNAs for HiPlex1 and 140 gRNAs for HiPlex2, modified as follows: the last nucleotide at the 5' end of the gRNA sequence (position 20) was replaced with a G for efficient T7 transcription. A T7 promoter sequence 5'-GGATCCTAATACGACTCACTATAG-3' was added at the 5' end of the protospacer, and a SpCas9 scaffold sequence 5'-GTTTTAGAGCTAGAA-3' was added at the 3' end. The sequences were ordered as DNA oPools (IDT) and reconstituted in nuclease-free H_2O at 100 μM. In-house production of sgRNAs was performed using the HighYield T7 sgRNA Synthesis Kit (SpCas9) (Jena Bioscience RNT-105). In brief, each pool (1 μM) was used for an assembly PCR reaction using three primers: T7fwd_sRNA: 5'-GGATCCTAATACGACTCACTATAG-3', T7rev_sgRNA: 5'-AAAAAAGCACCGACTCGG-3' and SpCas9_scaffold: 5'-AAAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTTAACTTGCTATT TCTAGCTCTAAAAC-3'. To increase complexity and avoid PCR bias, we performed three separate PCR reactions for each pool, which were then combined before *in vitro* transcription. The expected size of the assembled DNA template was confirmed on an agarose gel and used directly for T7 *in vitro* transcription. Three *in vitro* transcription reactions per pool were performed for increased yield, and were incubated for 90 min at 37°C. *In vitro* transcription products were purified using 2× volumes of Agencourt RNAclean XP magnetic beads (Beckman Coulter A66514) and resuspended in nuclease-free H_2O. RNA concentration was estimated using Qubit RNA Broad Range (Invitrogen Q10211).

4.7.BreakTag procedure and sequencing

DNA DSB ends of nuclease-digested gDNA were repaired and 3'-adenylated using the NEBNext Ultra II End Repair/dA-Tailing Module (NEB E7546) according to the manufacturer's instructions with the following modification: the total volume of the reaction was halved by using half the volume of the reagents. Labeling of DSB ends by ligation with the BreakTag linker was performed using the NEBNext Ultra II Ligation Module (NEB E7595) according to the manufacturer's instructions with the following modifications: the total volume of the reaction was halved by using half the volume of the reagents, and the USER enzyme digestion step was omitted. The BreakTag linker was used at a final concentration of 50 nM per sample. Labeled DNA was size-selected using 0.7× volumes of DNA AMPure XP beads (Beckman Coulter A63987) to remove nonligated linkers and then eluted in nuclease-free H_2O. Tagmentation with in-house Tn5 was performed in 10 mM Tris-HCl (pH 7.5) buffer containing 10 mM $MgCl_2$ and 25% *N*,*N*-dimethylformamide (DMF, Sigma–Aldrich 227056). Tagmentation reactions were assembled using 100–200 ng of DNA as input. Hyperactive Tn5 was used at a final concentration of 1.25 ng/µL. The tagmentation mix was then incubated at 55°C for 5 min in a preheated thermocycler followed by termination with 0.2% SDS at room temperature for 5 min. Libraries were amplified with the NEBNext Ultra II Q5 Master Mix (M0544) in a thermocycler programmed as follows:

Step	**Temperature**	**Time**	
1	72°C	5 min	Gap-filling reaction
2	98°C	30 s	
3	98°C	10 s	
4	63°C	30 s	14 loops
5	72°C	60 s	
6	72°C	5 min	
7	12°C	Hold	

Amplified and barcoded samples were size-selected to remove PCR byproducts by performing two consecutive 0.5× volume right-tail + 0.35× volume left-tail size selections using DNA AMPure XP beads (Beckman Coulter A63987). Final libraries were quantified using a Qubit dsDNA High Sensitivity Assay Kit and fragment size distribution was assessed on a BioAnalyzer High Sensitivity DNA chip. Libraries were pooled and sequenced on a NextSeq 500/550 platform with NextSeq 500/550 High Output Kit v2 chemistry for SE 1×75 bp sequencing or NovaSeq PE 2×150 bp.

4.8. BreakTag data analysis with BreakInspectoR

Initial preprocessing was done in a Linux cluster using the BreakTag NGSpipe2go pipeline (https://gitlab.rlp.net/imbforge/NGSpipe2go/-/tree/master/pipelines/breaktag). The pipeline processes raw reads as they are output by the sequencer and generates a BED file with coordinates containing DSBs. Raw reads (single- or paired-end) were first scanned and those not containing the expected 8-nt UMI followed by the 8-nt sample barcode in the 5' end of read 1 were discarded. Valid reads were aligned to the human reference genome version hg38 downloaded from UCSC with timestamp 2014-01-15 21:14, using the *mem* command in BWA (version 0.7.17-r1188) (Li, 2013) with a seed length of 19 and default scoring/penalty values for mismatches, gaps, and read clipping. Reads mapped with a minimum quality score Q=60 were retained to ensure we worked only with uniquely mapping reads. A final deduplication step was performed in which spatial consecutive reads mapping within a window of 30 nt and their UMIs differing by up to two mismatches were considered close PCR duplicates, and only one was kept. The resulting reads were aggregated per position and reported as a BED file.

Subsequent analysis was done using the BreakInspectoR package in R (https://gitlab.rlp.net/imbforge/breakinspectoR), which performs a guided search toward putative on/off targets. Starting from the previously generated BED files, BreakInspectoR identifies stacks of read ends near a PAM as candidate loci for containing a DSB, and calculates a *p*-value and false discovery rate for each site identified, considering also the signal found in a nontargeted library. The identification of sites required the function *breakinspectoR()* to search for stacks of at least three read ends at a distance of 3 nt from an "NGG" PAM, which is preceded by a protospacer sequence that differs by seven mismatches at most from the sgRNA sequence. Only breaks identified in standard chromosomes were retained.

4.9. Blunt rate estimation

For each site identified by BreakInspectoR, we analyzed the scission profile using the *scission_profile_analysis()* function. This function analyzes the signal in the PAM-proximal side and returns a table in the form of a *data.frame* attached as metadata columns of a *GRanges* object (Lawrence et al., 2013). The table extends the coordinates of the original DSB with the signal found around the position at which the enzyme is expected to cut, a *p*-value and false discovery rate that assess the significance of the signal found outside the expected cut site compared to the nontarget library, and the classification of a site according to its preference for forming blunt or staggered breaks. We performed the analysis by using the function to look in a region between [−3, +3] nucleotides up-/downstream of the expected cut site; for Cas9 this was 3 nt upstream (toward the 5' end) from the PAM. In order to avoid sites that could mislead the analysis, we focused only on sites with an "NGG" PAM, for which, in principle, expected cut sites are readily identified. Finally, from the table generated by *scission_profile_analysis()*, we could calculate the blunt rate for a site. We did this in two ways: 1) as a fraction of the signal found in the expected cut site (PAM 3 nt upstream, i.e., position 17 of the protospacer) and the total amount of signal in the region [−3, +3] around the cut site; and 2) as a log2 ratio of the signal in the expected cut site versus the signal in the region [−3, +3] around the cut site after excluding the signal in the cut site.

4.10. Machine learning model for the prediction of blunt rates

We trained a machine learning model to predict scission profiles using the XGBOOST flavor of the Gradient Boosting Machine algorithm implemented in the H2O.ai framework (*Python Interface for H2O. Module Version 3.38.0.2*, 2014/2016). The software was installed in the Bioconductor R container release version 3.15 (Huber et al., 2015) (bioconductor/bioconductor_docker:RELEASE_3_15). We tuned the hyperparameters of the algorithm to use 1,000 trees of unlimited depth, DART as the booster algorithm (Rashmi & Gilad-Bachrach, 2015), and five folds for K-fold cross-validation with automatic fold assignment of instances.

Because the number and scission profiles of the identified targets differ greatly among sgRNAs, we used only a subset of the total identified targets as training instances. In defining the subset, we selected only highly covered sites with at least 16 raw reads in the PAM-proximal side, and accounted for specific biases. We limited the number of targets selected per sgRNA to 100 in order to avoid biases toward highly promiscuous sgRNA sequences that we identified to cut promiscuously, and additionally sampled staggered targets with a probability K^{-1}, where K is the ratio between the number of staggered (blunt reads <20%) and blunt (blunt reads >80%) targets for a specific sgRNA, in order to pick more from the pool of staggered targets and compensate for

their under-representation in the total set of identified targets. This resulted in a final set of 18,759 *instances* in the training set.

The *response* variable to be predicted was the log2 ratio between the number of raw reads mapped in the PAM-proximal side exactly at position 17 of the protospacer (the expected cut site) and the sum of raw reads mapped in the PAM-proximal side found in positions 14–16 and 18–20 of the protospacer. A pseudocount was added to both the denominator and numerator of this fraction in order to avoid a division by 0.

We tried to reflect in the *predictor* variables both the targeted protospacer sequence and the actual target sequence used to guide the Cas9 enzyme, along with the mismatches between the two. To that end, we performed one-hot encoding by constructing a 4×4 matrix for each of the 20 positions of the protospacer, each row representing one of the possible nucleotides (A, C, G, T) to occupy that position in the targeted protospacer, and in each column the same for the sgRNA sequence. The matrix was filled with *0* with the exception of the cell representing the nucleotide in the protospacer (row) and the sgRNA (column) for that position, which would contain *1*. Each matrix was converted into a vector of length 16 by concatenating the column vectors, and finally the 20 vectors were concatenated into one large vector of length 320 with the final representation of the one-hot encoding. In addition, we included an additional predictor variable representing the number of mismatches between the targeted protospacer and the sgRNA sequence in the first 10 positions of the protospacer, and a second variable representing the mismatches in the last 10 positions of the protospacer. In total, we used 322 variables to represent each training instance.

We also produced a reduced model using only the last 10 nucleotides of the protospacer as predictor variables under the same conditions (algorithm, hyperparameters, training instances, response variable, and encoding of the predictor variables).

Sequence motifs related to the scission profile were produced with the ggseqlogo package in R (Wagih, 2017/2023). The height of the nucleotide was calculated as the coefficient of the linear model relating the centered log2 blunt rates to the presence of a specific nucleotide at a certain position of the protospacer, independent of the other positions, and scaled to the nucleotide importance at that position, as calculated by the XGBOOST algorithm when training the model.

4.11.Selection of sites containing SNPs in Genome in a Bottle individuals for HiPlex BreakTag

We downloaded the VCF file containing the single-nucleotide variants (SNVs) called in the Genome in a Bottle (GIAB) (Zook et al., 2016). We filtered the files to retain single-nucleotide polymorphisms (SNPs) only, and retrieved the 20 bp of sequence context around those sites. We retained two subsets of 394,585 and 395,392 putative CRISPR-Cas9 target sites that contain an "NGG" PAM preceded by a protospacer containing at positions 17 or 18 (respectively) an SNP found in at least one of the GIAB samples. We then used the reduced machine learning model, which uses only the last 10 positions of the protospacer, to predict the expected blunt rate of those putative target sites for the reference allele sequence targeted with an sgRNA matching the reference sequence, and also for the mutated allele targeted with an sgRNA containing the mutation. The top 150 sites with the lowest blunt rates (75 in sense and 75 in antisense strands) and the top 150 sites with highest blunt rates (75 in sense and 75 in antisense strands) were selected for HiPlex BreakTag sgRNA pool generation. For greater statistical power, we selected sites for which the alternative allele is found in three or four donors.

4.12. Analysis of SNP-driven changes in scission profile

We used the *scission_profile_analysis()* function in BreakInspectoR to obtain the scission profile of the 300 sites picked from the previously selected SNP-containing sites in GIAB genomes. We calculated the blunt rate as the fraction of the BreakTag signal in the expected cut site (position 17 of the protospacer) with respect to the total signal in the region [−3, +3] around the cut site, obtaining an approximation for the number of blunt breaks compared to the total number of breaks as captured by BreakTag. For the visualizations comparing the blunt rate and the genotype, we selected highly covered sites with at least 16 raw reads in the PAM-proximal side and reference and alternative genotype information in at least one sample for each genotype.

4.13. Nucleofection of lymphoblastoid cells

For the preparation of RNP complexes, sgRNAs targeting SNP-containing loci were generated in-house using the HighYield T7 sgRNA Synthesis Kit (SpCas9) (Jena Bioscience RNT-105). Two hundred picomolar sgRNA was mixed with 100 pM Alt-R S.p. Cas9-GFP V3 (IDT 10008100) and incubated at room temperature for 10 min. A total of 5 × 105 cells per reaction were resuspended in SF Cell Line 4D-Nucleofector solution (Lonza V4XC-2032) and nucleofected in a 4D-NucleoFector system using the pulse code DN-100. Nucleofected cells were transferred to a plate containing culture medium and kept in a humidified incubator at 37°C supplemented with 5% CO2 for 3 days before gDNA was extracted for indel analysis.

4.14. Amplicon sequencing and editing analysis using CRISPResso2

The gDNA of lymphoblastoid cells nucleofected with RNPs was extracted 3 days after CRISPR delivery using the Qiagen Blood and Tissue Kit (Qiagen 69506) according to the manufacturer's instructions. Approximately 100 ng of gDNA from each sample was used for locus amplification using the primers listed in Table 4. Amplicon libraries were generated as described elsewhere (Yau & Rana, 2018) with the following modifications: a first round of amplification using the NEBNext Ultra II Q5 Master Mix (M0544) was performed with 33 cycles. The amplified DNA was purified using a 1× volume of DNA AMPure XP beads (Beckman Coulter A63987) and the entire purified product was used for a second round of PCR with primers containing p5 and p7 sequences for Illumina sequencing Table 5. Amplicons were pooled and sequenced in a MiniSeq sequencer with the MiniSeq Mid Output Kit (Illumina FC-420-1004) in single-read mode and 150 cycles.

Indel analysis was performed in a local Linux cluster using CRISPresso2 (Clement et al., 2019) using the following parameters: *--amplicon_min_alignment_score 50 --quantification_window_size 10 --quantification_window_center -3 --exclude_bp_from_left 0 --exclude_bp_from_right 0 --ignore_substitutions --plot_window_size 20 --min_frequency_alleles_around_cut_to_plot 0.*

Table 4. Primers used for first round of amplification of polymorphic loci in cells of GIAB donors.

Locus	Forward sequence	Reverse sequence
p18_chr19_17684551	CTACACGACGCTCTTCCGATCTGCATTTACCATTCACAGCCTC	GACGTGTGCTCTTCCGATCTCACCAACACATAGCAGACC
p17_chr9_36852319	CTACACGACGCTCTTCCGATCTGGAAGAGGAAACCCAGTGAC	GACGTGTGCTCTTCCGATCTTCCCAATCTGCGACACAAG
p18_ref_chr5_57408285	CTACACGACGCTCTTCCGATCTCTTTCCTTCATCTGGTGGCAG	GACGTGTGCTCTTCCGATCTCATGTGCCCTCAACCCAAAT
p18_ref_chr5_67196897	CTACACGACGCTCTTCCGATCTGTCCCTTTTGCATACCACCC	GACGTGTGCTCTTCCGATCTCAGAAGCTGTTGGAGTTGGC
p18_ref_chr6_7441004	CTACACGACGCTCTTCCGATCTtctcctttcccctctcctga	GACGTGTGCTCTTCCGATCTctcacTAGCGGGGAAGGAAG

p18_ref_chr14_55194777	CTACACGACGCTCTTCCGATCTgctaacctggcattcaaacca	GACGTGTGCTCTTCCGATCTccgcaagctgacagaagaat
p18_ref_chr14_71031307	CTACACGACGCTCTTCCGATCTGCAAATGGAGAAGGGGCATT	GACGTGTGCTCTTCCGATCTGATGATTTGCCCACTCAGCC
p18_ref_chr15_60914261	CTACACGACGCTCTTCCGATCTTGGCTGTGCTAATTTGCTGT	GACGTGTGCTCTTCCGATCTCAAAGCATGGTGGTGGACTT
p18_ref_chr16_73298741	CTACACGACGCTCTTCCGATCTTGCTCTTAACTGTGAGGCCA	GACGTGTGCTCTTCCGATCTACGTCAGGCATCACTATCAGT
p18_ref_chr16_81168168	CTACACGACGCTCTTCCGATCTCCTCACTGGCATGCTGTAGA	GACGTGTGCTCTTCCGATCTCGGAGATACAGGGAGCTCG
p18_ref_chr17_79628954	CTACACGACGCTCTTCCGATCTACCGGCCCTTTGGAATAGG	GACGTGTGCTCTTCCGATCTgaatGAACAAAGGCCCTGCA
p18_ref_chr18_11986327	CTACACGACGCTCTTCCGATCTCCTTGTCCGGCCAGATCTG	GACGTGTGCTCTTCCGATCTACCTGAGAAGCGGTCCATG
p18_ref_chr19_17684550	CTACACGACGCTCTTCCGATCTACCATTCACAGCCTCAGGAT	GACGTGTGCTCTTCCGATCTCACCCAGCACCAACACATAG
p17_ref_chr1_10690138	CTACACGACGCTCTTCCGATCTAGTGGGTGTATCTGGCCAAG	GACGTGTGCTCTTCCGATCTTGTGGAGGCTTCTAGGAACC
p17_ref_chr1_177422658	CTACACGACGCTCTTCCGATCTTGGAAGTGTGTGCAGTCTGA	GACGTGTGCTCTTCCGATCTagtatttgtatcccctgtgcct
p17_ref_chr1_27118527	CTACACGACGCTCTTCCGATCTCTCTCCTAACGCCCTGACTC	GACGTGTGCTCTTCCGATCTGCCCTGAAAGTTTTGCCTCA
p17_ref_chr12_64945057	CTACACGACGCTCTTCCGATCTtcactaaatccagcccacact	GACGTGTGCTCTTCCGATCTACAATCGAGGTCCCAGGAAA

Table 5. Indexed primers used in the second round of amplification of polymorphic loci.

Primer	Sequence
P5_Scriptseq	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T
P7_i701_Scriptseq	CAAGCAGAAGACGGCATACGAGATCGGTTCAAGTGACTGGAGTTCAGACGTGTGCTCTTCCGAT*C*T
P7_i702_Scriptseq	CAAGCAGAAGACGGCATACGAGATGCTGGATTGTGACTGGAGTTCAGACGTGTGCTCTTCCGAT*C*T
P7_i703_Scriptseq	CAAGCAGAAGACGGCATACGAGATTAACTCGGGTGACTGGAGTTCAGACGTGTGCTCTTCCGAT*C*T
P7_i704_Scriptseq	CAAGCAGAAGACGGCATACGAGATTAACAGTTGTGACTGGAGTTCAGACGTGTGCTCTTCCGAT*C*T
P7_i705_Scriptseq	CAAGCAGAAGACGGCATACGAGATATACTCAAGTGACTGGAGTTCAGACGTGTGCTCTTCCGAT*C*T
P7_i706_Scriptseq	CAAGCAGAAGACGGCATACGAGATGCTGAGAAGTGACTGGAGTTCAGACGTGTGCTCTTCCGAT*C*T

4.15. Engineered Cas9 variant cloning, expression and purification

Cloning, expression and purification of Cas9 variants was performed by the IMB Protein Production Core Facility. The pET-Cas9-NLS-6xHis expression vectors for Cas9 variants were generated by using Gibson Assembly. As a PCR template for the expression vector backbone, pET wildtype Cas9-NLS-6xHis was used (Zuris et al., 2015) (Addgene Plasmid #62933). The PCR templates for the Cas9 variants were pX165-LZ3 Cas9 (Addgene Plasmid #140561), pX165-

ev°Cas9 (Addgene Plasmid #140569), pX165-xCas9 (Addgene Plasmid #140568), pX165-HypaCas9 (Addgene Plasmid #140567), and pX165-SniperCas9 (Addgene Plasmid #140560).

The pET expression vectors containing the different C-terminally NLS-His_6-tagged Cas9 variants were transformed into E. Coli BL21 (DE3) CodonPlus (Agilent), which were grown at 37°C and 140 rpm until an OD_{600} value of 0.5 was achieved. Cultures were cooled to 18°C on ice and protein expression was induced using IPTG at a final concentration of 0.5 mM, and incubated for a further 21 h at 18°C and 140 rpm. Cells were harvested by centrifugation (4,000 × *g*, 15 min), resuspended in ice-cold lysis buffer (30 mM Tris-HCl, 500 mM NaCl, 10 mM imidazole, 1 mM $MgCl_2$, 1 mM TCEP, 5% glycerol, 1× Complete protease inhibitor, 100 U/ml Benzonase, pH 8.0) and lysed by high-pressure homogenization at 28 kpsi (Constant Systems CF1 cell disruptor). Cells were cleared by centrifugation (40,000 × *g*, 30 min, 4°C) and the cleared lysate was applied to a HisTrap FF 5 ml column (Cytiva), using an automated chromatography system (Biorad NGC Quest Plus; used for all chromatography steps). The column was washed with 20 CV wash buffer (30 mM Tris-HCl, 500 mM NaCl, 10 mM imidazole, 5 % glycerol) and the Cas9 variants were eluted from the Ni–NTA column by applying a linear gradient of 10–500 mM imidazole (containing 30 mM Tris-HCl, 500 mM NaCl, 5% glycerol). The eluted proteins were diluted 1:10 in a low-salt buffer (25 mM Na–Hepes, pH 7.2, 100 mM NaCl, 5% glycerol), applied to a HiTrap Heparin 5 ml column (Cytiva) and eluted by applying a linear NaCl gradient from 100 to 1000 mM. Elution fractions containing the Cas9 variants were pooled and concentrated using Amicon Ultra 15 spin concentrators (Merck). Concentrated proteins were applied to a gel filtration column (Superdex 200 16/60 pg, Cytica, 40 mM Na–Hepes, pH 7.4, 400 mM NaCl, 10% glycerol). Peak fractions containing the Cas9 variants were pooled, concentrated to 6.4 g/l and diluted 1:2 with 86% glycerol to a final concentration of 3.2 g/l (20 µM). Aliquots of Cas9 variants were snap-frozen in liquid nitrogen and stored at −80°C. HiFiCas9 was purchased from Integrated DNA technologies (IDT # 1081060) (Vakulskas et al., 2018).

4.16. Prediction of blunt rates of gRNAs targeting pathogenic deletions

The full set of variants annotated in ClinVar as of April 2023, comprising a total of 2,122,310 variants, was downloaded from NIH's FTP server (https://ftp.ncbi.nih.gov/pub/clinvar/vcf_GRCh38/clinvar.vcf.gz). Only variants which were 1nt deletions, located in standard chromosomes, overlapping an exon annotated in TxDb.Hsapiens.UCSC.hg38.knownGene (data package made from resources at UCSC on 16:50:30 +0000 Thu, 07 Apr 2022) and annotated in ClinVar as "Pathogenic" or "Likely_pathogenic" were considered (31,010 variants). We focused on a subset of 8,705 deletions that had an NGG motif

directly adjacent to them in either strand, and up to 4nt upstream. Those sites were candidates for being cut by Cas9 in a staggered manner, which could potentially induce a templated +1 insertion as the repair outcome, correcting the frameshift in the pathogenic allele and potentially recovering the original protein sequence. We calculated that a total of 4,999 of those deletions would recover the original protein sequence with a templated +1 insertion. Next, we designed "in-silico" the gRNA sequences that would target the regions containing the deletions, and estimated the blunt rate using the previously described XGBoost models for SpCas9 and LZ3 trained with the HiPlex library. Those sites predicted to be cut in a highly staggered manner (log2 blunt rate < -2) in which a templated insertion would recover the original protein were finally reported as pathogenic variants being potentially treated with a CRISPR-Cas9 therapy.

4.17. Construction of gRNA-target pair lentiviral libraries

Using our XGBoost models for SpCas9, we predicted the blunt rate of human genome sites, and selected 150 sites predicted to be cut mostly blunt, and 150 sites predicted to be cut mostly staggered. For the “ALT” and “REF” libraries all gRNAs used in the HiPlex3 dataset were used. The cloning strategy of gRNA-target pair lentiviral libraries was adapted from Allen et al., 2019 (Allen et al., 2019). Briefly, a scaffoldless lentiviral expression vector, pKLV2-U6(BbsI)-PKGpuro2ABFP-W, was generated by removing the improved gRNA SpCas9 scaffold from pKLV2-U6gRNA5(BbsI)-PKGpuro2ABFP-W (a gift from Kosuke Yusa, Addgene plasmid #67974), allowing us to use either the improved or conventional scaffold sequences, as in the original strategy (Allen et al., 2019). The deletion was generated by amplifying two fragments encompassing the 5’ end of the AmpR cassette to U6 promoter and PGK promoter of the 3’ end of the AmpR cassette, followed by Gibson assembly. The empty vector was transformed into Stabl3 chemically competent cells, single colonies were picked and scaffold deletion was confirmed via Sanger sequencing.

For the library cloning step, we generated a 170nt oligonucleotide pool (IDT) encoding the gRNA and a portion of the allele sequence containing 79 nucleotides with the target sequence + PAM in the center for the 4 individual libraries. The oligonucleotide was amplified with primers compatible with the scaffold used, and a Gibson assembly was used to fuse the amplified pool to a 193nt Ultramer Duplex (IDT) encoding the improved version of the gRNA scaffold and a spacer sequence (Allen et al., 2019). Three separated Gibson assembly reactions were performed per pool at a 1:1 molar ratio, followed by an incubation for 1 h at 50°C, and subsequently pooled for column-based purification (Monarch PCR & DNA Clean-up kit, New England Biolabs #T1030S) and removal of linear DNA was achieved by treating the samples with Plasmid-safe ATP-dependent DNAse (Epicentre). The intermediate circular insert and scaffoldless vector were

linearized with the FastDigest BpiI (IIs class) kit (Thermo Scientific, FD1014) for 30 minutes and ligated in triplicates per pool (T4 DNA ligase, NEB). The replicates were pooled and transformed in Stabl3 chemically competent cells.

4.18. Transduction of gRNA-target lentiviral pools into Cas9-expressing cells

For lentiviral packaging of gRNA-target libraries, the gRNA-target libraries were independently co-transfected with the two packaging plasmids and the supernatants were pooled and concentrated 50-100-fold. Packaging and transduction were performed as described previously (Papapetrou & Sadelain, 2011). Briefly, we produced the viruses by co-transfection of 293T cells with each of the four library pools and two helper plasmids, psPAx2 and pMD2.g encoding the VSV-G envelope and the lentiviral gag-pol genes, respectively. We harvested the lentiviral vector-containing supernatant twice, at ~42 and 66 hours post-transfection and concentrated it by using Lenti-X-concentrator (Takara, 631232). We plated 300,000 cells in a well of 6-well plate and transduced with the vector supernatants and 4µg/ml polybrene in a total volume of 2mL. After 48 hours, the transduced cells were removed from the 6-well plates, and one fifth of the cells were tested for BFP expression by flow cytometry (BD Canto), while the rest were plated in 10cm2 tissue-culture dishes for selection with puromycin (1µg/ml). The transduced cells were kept under puromycin selection for 5 days. On the last day, cells were collected and tested for BFP expression, and genomic DNA was isolated using the Qiagen Blood and Tissue Kit (Qiagen 69506).

4.19. gRNA-target pair amplicon sequencing library preparation

The region containing the gRNA sequence and 79nt portion of the allele was amplified using the Fwd_pool and Rev_pool primers (Table 6) with the NEBNext® Ultra™ II Q5® Master Mix (New England Biosciences #M0544) with the following program: 98°C for 60s, 24 loops of 98°C for 10s and 72°C for 30s, followed by a final extension at 72°C for 2 minutes. The PCR product was purified using 0.9x volumes of DNA AMPure XP beads (Beckman Coulter A63987) and eluted in nuclease-free water. The entire product of clean-up was used for a second PCR round with indexed primers (Table 5) with the following conditions: 98°C for 60s, 13 loops of 98°C for 10s, 67°C for 10s and 72°C for 20s, followed by a final extension at 72°C for 2 minutes. The indexed libraries were pooled and the band corresponding to the amplicon size (464 base pairs) was excised from a 2% agarose gel, purified and sequenced in paired-end mode (2x150bp) in a NextSeq2000 sequencer with 40% PhiX spike-in.

Table 6. Primers used for the amplification of the gRNA-target pairs.

Primer	**Sequence (5'-3')**
Fwd_pool	CTACACGACGCTCTTCCGATCTTCTTGTGGAAAGGACGAAACA
Rev_pool	GACGTGTGCTCTTCCGATCTCTACCCGGTAGAATTGGATCCAAAC

4.20. Analysis of gRNA-target repair outcomes

A total of 80,701,952 M paired-end (2x150bp) reads were sequenced in a NextSeq2000 P1 FC. The first read in pair was used solely to estimate the abundance of each gRNA, as it reads into the gRNA portion of the construct. The second pair that reads into the target sequence, was reverse complemented with the fastx_toolkit (http://hannonlab.cshl.edu/fastx_toolkit) and stripped from the first 57 bases and kept only the immediate 79nt using Trimmotatic 67 with options SE HEADCROP:57 CROP:79, which would keep only the 79nt-long portion of the read containing the actual amplicon of the targeted sequence. Processed reads from technical replicates were merged in a single fastq file, and indels were called using CRISPResso2 (Clement et al., 2019) in pooled mode (CRISPRessoPooled), restricting the analysis to regions with at least 100 aligned reads and ignoring substitutions other than indels. gRNAs with detected activity in WT cells not expressing Cas9 that had been reported in the CRISPResso2 analysis with at least 100 edited reads, were excluded from the analysis. For the rest, we extracted from the CRISPResso2 analysis output the length of the indel, the frequency of the most common +1 insertion over all edited sequences, and the inserted nucleotide. We considered a protein sequence being recovered when the inserted nucleotide in the most common +1 indel was the very same nucleotide found immediately after the cutsite determined by CRISPResso2.

Chapter 5. Cited References

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Allemailem, K. S., Almatroodi, S. A., Almatroudi, A., Alrumaihi, F., Al Abdulmonem, W., Al-Megrin, W. A. I., Aljamaan, A. N., Rahmani, A. H., & Khan, A. A. (2023). Recent Advances in Genome-Editing Technology with CRISPR/Cas9 Variants and Stimuli-Responsive Targeting Approaches within Tumor Cells: A Future Perspective of Cancer Management. *International Journal of Molecular Sciences*, *24*(8), 7052. https://doi.org/10.3390/ijms24087052

Allen, F., Crepaldi, L., Alsinet, C., Strong, A. J., Kleshchevnikov, V., De Angeli, P., Páleníková, P., Khodak, A., Kiselev, V., Kosicki, M., Bassett, A. R., Harding, H., Galanty, Y., Muñoz-Martínez, F., Metzakopian, E., Jackson, S. P., & Parts, L. (2019). Predicting the mutations generated by repair of Cas9-induced double-strand breaks. *Nature Biotechnology*, *37*(1), 64–82. https://doi.org/10.1038/nbt.4317

Amemiya, H. M., Kundaje, A., & Boyle, A. P. (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Scientific Reports*, *9*(1), Article 1. https://doi.org/10.1038/s41598-019-45839-z

Amitai, G., & Sorek, R. (2016). CRISPR–Cas adaptation: Insights into the mechanism of action. *Nature Reviews Microbiology*, *14*(2), Article 2. https://doi.org/10.1038/nrmicro.2015.14

Anzalone, A. V., Koblan, L. W., & Liu, D. R. (2020). Genome editing with CRISPR–Cas nucleases, base editors, transposases and prime editors. *Nature Biotechnology*, *38*(7), 824–844. https://doi.org/10.1038/s41587-020-0561-9

Anzalone, A. V., Randolph, P. B., Davis, J. R., Sousa, A. A., Koblan, L. W., Levy, J. M., Chen, P. J., Wilson, C., Newby, G. A., Raguram, A., & Liu, D. R. (2019). Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature*, *576*(7785), 149–157. https://doi.org/10.1038/s41586-019-1711-4

Ashley, E. A. (2016). Towards precision medicine. *Nature Reviews Genetics*, *17*(9), Article 9. https://doi.org/10.1038/nrg.2016.86

Atkins, A., Chung, C.-H., Allen, A. G., Dampier, W., Gurrola, T. E., Sariyer, I. K., Nonnemacher, M. R., & Wigdahl, B. (2021). Off-Target Analysis in Gene Editing and Applications for Clinical Translation of CRISPR/Cas9 in HIV-1 Therapy. *Frontiers in Genome Editing*, *3*(August), 1–26. https://doi.org/10.3389/fgeed.2021.673022

Aymard, F., Aguirrebengoa, M., Guillou, E., Javierre, B. M., Bugler, B., Arnould, C., Rocher, V., Iacovoni, J. S., Biernacka, A., Skrzypczak, M., Ginalski, K., Rowicka, M., Fraser, P., & Legube, G. (2017). Genome-wide mapping of long-range contacts unveils clustering of DNA double-strand breaks at damaged active genes. *Nature Structural and Molecular Biology*, *24*(4), 353–361. https://doi.org/10.1038/nsmb.3387

Aymard, F., Bugler, B., Schmidt, C. K., Guillou, E., Caron, P., Briois, S., Iacovoni, J. S., Daburon, V., Miller, K. M., Jackson, S. P., & Legube, G. (2014). Transcriptionally active chromatin recruits homologous recombination at DNA double-strand breaks. *Nature Structural and Molecular Biology*, *21*(4), 366–374. https://doi.org/10.1038/nsmb.2796

Bae, S., Park, J., & Kim, J.-S. (2014). Cas-OFFinder: A fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics (Oxford, England)*, *30*(10), 1473–1475. https://doi.org/10.1093/bioinformatics/btu048

Bao, X. R., Pan, Y., Lee, C. M., Davis, T. H., & Bao, G. (2021). Tools for experimental and computational analyses of off-target editing by programmable nucleases. *Nature Protocols*, *16*(1), 10–26. https://doi.org/10.1038/s41596-020-00431-y

Becker, S., & Boch, J. (2021). TALE and TALEN genome editing technologies. *Gene and Genome Editing*, *2*, 100007. https://doi.org/10.1016/j.ggedit.2021.100007

Bermudez-Cabrera, H. C., Culbertson, S., Barkal, S., Holmes, B., Shen, M. W., Zhang, S., Gifford, D. K., & Sherwood, R. I. (2021). Small molecule inhibition of ATM kinase increases CRISPR-Cas9 1-bp insertion frequency. *Nature Communications*, *12*, 5111. https://doi.org/10.1038/s41467-021-25415-8

Biehs, R., Steinlage, M., Barton, O., Juhász, S., Künzel, J., Spies, J., Shibata, A., Jeggo, P. A., & Löbrich, M. (2017). DNA Double-Strand Break Resection Occurs during Non-homologous End Joining in G1 but Is Distinct from Resection during Homologous Recombination. *Molecular Cell*, *65*(4), 671-684.e5. https://doi.org/10.1016/j.molcel.2016.12.016

Biernacka, A., Zhu, Y., Skrzypczak, M., Forey, R., Pardo, B., Grzelak, M., Nde, J., Mitra, A., Kudlicki, A., Crosetto, N., Pasero, P., Rowicka, M., & Ginalski, K. (2018). I-BLESS is an ultra-sensitive method for detection of DNA double-strand breaks. *Communications Biology*, *1*(1). https://doi.org/10.1038/s42003-018-0165-9

Borhade, M. B., & Kondamudi, N. P. (2024). Sickle Cell Crisis. In *StatPearls*. StatPearls Publishing. http://www.ncbi.nlm.nih.gov/books/NBK526064/

Bouwman, B. A. M., Agostini, F., Garnerone, S., Petrosino, G., Gothe, H. J., Sayols, S., Moor, A. E., Itzkovitz, S., Bienko, M., Roukos, V., & Crosetto, N. (2020). Genome-wide detection of DNA double-strand breaks by in-suspension BLISS. *Nature Protocols*, *15*(12), 3894–3941. https://doi.org/10.1038/s41596-020-0397-2

Brailovsky, Y., Rajapreyar, I., & Alvarez, R. (2023). TTR Amyloidosis. *JACC Case Reports*, *10*, 101759. https://doi.org/10.1016/j.jaccas.2023.101759

Brambati, A., Sacco, O., Porcella, S., Heyza, J., Kareh, M., Schmidt, J. C., & Sfeir, A. (2023). RHINO directs MMEJ to repair DNA breaks in mitosis. *Science (New York, N.Y.)*, *381*(6658), 653–660. https://doi.org/10.1126/science.adh3694

Brinkman, E. K., Chen, T., de Haas, M., Holland, H. A., Akhtar, W., & van Steensel, B. (2018). Kinetics and Fidelity of the Repair of Cas9-Induced Double-Strand DNA Breaks. *Molecular Cell*, *70*(5), 801-813.e6. https://doi.org/10.1016/j.molcel.2018.04.016

Brunet, E., & Jasin, M. (2018). Induction of chromosomal translocations with CRISPR-Cas9 and other nucleases: Understanding the repair mechanisms that give rise to translocations. *Advances in Experimental Medicine and Biology*, *1044*, 15–25. https://doi.org/10.1007/978-981-13-0593-1_2

Burdett, T., & Nuseibeh, S. (2023). Changing trends in the development of AAV-based gene therapies: A meta-analysis of past and present therapies. *Gene Therapy*, *30*(3), 323–335. https://doi.org/10.1038/s41434-022-00363-0

Cameron, P., Fuller, C. K., Donohoue, P. D., Jones, B. N., Thompson, M. S., Carter, M. M., Gradia, S., Vidal, B., Garner, E., Slorach, E. M., Lau, E., Banh, L. M., Lied, A. M., Edwards, L. S., Settle, A. H., Capurso, D., Llaca, V., Deschamps, S., Cigan, M., … May, A. P. (2017). Mapping the genomic landscape of CRISPR-Cas9 cleavage. *Nature Methods*, *14*(6), 600–606. https://doi.org/10.1038/nmeth.4284

Cancellieri, S., Zeng, J., Lin, L. Y., Tognon, M., Nguyen, M. A., Lin, J., Bombieri, N., Maitland, S. A., Ciuculescu, M.-F., Katta, V., Tsai, S. Q., Armant, M., Wolfe, S. A., Giugno, R., Bauer, D. E., & Pinello, L. (2022). Human genetic diversity alters off-target outcomes of therapeutic gene editing. *Nature Genetics*, *138*(Supplement 1), 3993–3993. https://doi.org/10.1038/s41588-022-01257-y

Carroll, D. (2011). Genome Engineering With Zinc-Finger Nucleases. *Genetics*, *188*(4), 773–782. https://doi.org/10.1534/genetics.111.131433

Casini, A., Olivieri, M., Petris, G., Montagna, C., Reginato, G., Maule, G., Lorenzin, F., Prandi, D., Romanel, A., Demichelis, F., Inga, A., & Cereseto, A. (2018). A highly specific SpCas9 variant is identified by in vivo screening in yeast. *Nature Biotechnology*, *36*(3), 265–271. https://doi.org/10.1038/nbt.4066

Cebrian-Serrano, A., & Davies, B. (2017). CRISPR-Cas orthologues and variants: Optimizing the repertoire, specificity and delivery of genome engineering tools. *Mammalian Genome*, *28*(7), 247–261. https://doi.org/10.1007/s00335-017-9697-4

Chakrabarti, A. M., Henser-Brownhill, T., Monserrat, J., Poetsch, A. R., Luscombe, N. M., & Scaffidi, P. (2019). Target-Specific Precision of CRISPR-Mediated Genome Editing. *Molecular Cell*, *73*(4), 699-713.e6. https://doi.org/10.1016/j.molcel.2018.11.031

Chatterjee, N., & Walker, G. C. (2017). Mechanisms of DNA damage, repair and mutagenesis. *Environmental and Molecular Mutagenesis*, *58*(5), 235–263. https://doi.org/10.1002/em.22087

Chauhan, V. P., Sharp, P. A., & Langer, R. (2023). Altered DNA repair pathway engagement by engineered CRISPR-Cas9 nucleases. *Proceedings of the National Academy of Sciences*, *120*(11), e2300605120. https://doi.org/10.1073/pnas.2300605120

Chen, J. S., Dagdas, Y. S., Kleinstiver, B. P., Welch, M. M., Sousa, A. A., Harrington, L. B., Sternberg, S. H., Joung, J. K., Yildiz, A., & Doudna, J. A. (2017). Enhanced proofreading governs CRISPR-Cas9 targeting accuracy. *Nature*, *550*(7676), 407–410. https://doi.org/10.1038/nature24268

Chen, W., McKenna, A., Schreiber, J., Haeussler, M., Yin, Y., Agarwal, V., Noble, W. S., & Shendure, J. (2019). Massively parallel profiling and predictive modeling of the outcomes of CRISPR/Cas9-mediated double-strand break repair. *Nucleic Acids Research*, *47*(15), 7989–8003. https://doi.org/10.1093/nar/gkz487

Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L. H., Zielinski, M., Sargeant, T., Schneider, R. G., Senior, A. W., Jumper, J., Hassabis, D., Kohli, P., & Avsec, Ž. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, *381*(6664), eadg7492. https://doi.org/10.1126/science.adg7492

Christian, M., Cermak, T., Doyle, E. L., Schmidt, C., Zhang, F., Hummel, A., Bogdanove, A. J., & Voytas, D. F. (2010). Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics*, *186*(2), 757–761. https://doi.org/10.1534/genetics.110.120717

Clement, K., Rees, H., Canver, M. C., Gehrke, J. M., Farouni, R., Hsu, J. Y., Cole, M. A., Liu, D. R., Joung, J. K., Bauer, D. E., & Pinello, L. (2019). CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nature Biotechnology*, *37*(3), Article 3. https://doi.org/10.1038/s41587-019-0032-3

Cohen-Tannoudji, M., Robine, S., Choulika, A., Pinto, D., El Marjou, F., Babinet, C., Louvard, D., & Jaisser, F. (1998). I-SceI-Induced Gene Replacement at a Natural Locus in Embryonic Stem Cells. *Molecular and Cellular Biology*, *18*(3), 1444–1448.

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., Wu, X., Jiang, W., Marraffini, L. A., & Zhang, F. (2013). Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science*, *339*(6121), 819–823. https://doi.org/10.1126/science.1231143

Cromer, M. K., Majeti, K. R., Rettig, G. R., Murugan, K., Kurgan, G. L., Bode, N. M., Hampton, J. P., Vakulskas, C. A., Behlke, M. A., & Porteus, M. H. (2023). Comparative analysis of CRISPR off-target discovery tools following ex vivo editing of CD34+ hematopoietic stem and progenitor cells. *Molecular Therapy*, *31*(4), 1074–1087. https://doi.org/10.1016/j.ymthe.2023.02.011

Crosetto, N., Mitra, A., Silva, M. J., Bienko, M., Dojer, N., Wang, Q., Karaca, E., Chiarle, R., Skrzypczak, M., Ginalski, K., Pasero, P., Rowicka, M., & Dikic, I. (2013). Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nature Methods*, *10*(4), 361–365. https://doi.org/10.1038/nmeth.2408

Deltcheva, E., Chylinski, K., Sharma, C. M., Gonzales, K., Chao, Y., Pirzada, Z. A., Eckert, M. R., Vogel, J., & Charpentier, E. (2011). CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature*, *471*(7340), 602–607. https://doi.org/10.1038/nature09886

Dobbs, F. M., van Eijk, P., Fellows, M. D., Loiacono, L., Nitsch, R., & Reed, S. H. (2022). Precision digital mapping of endogenous and induced genomic DNA breaks by INDUCE-seq. *Nature Communications*, *13*(1), 3989. https://doi.org/10.1038/s41467-022-31702-9

Doetschman, T., Gregg, R. G., Maeda, N., Hooper, M. L., Melton, D. W., Thompson, S., & Smithies, O. (1987). Targetted correction of a mutant HPRT gene in mouse embryonic stem cells. *Nature*, *330*(6148), Article 6148. https://doi.org/10.1038/330576a0

Donohoue, P. D., Pacesa, M., Lau, E., Vidal, B., Irby, M. J., Nyer, D. B., Rotstein, T., Banh, L., Toh, M. S., Gibson, J., Kohrs, B., Baek, K., Owen, A. L. G., Slorach, E. M., van Overbeek, M., Fuller, C. K., May, A. P., Jinek, M., & Cameron, P. (2021). Conformational control of Cas9 by CRISPR hybrid RNA-DNA guides mitigates off-target activity in T cells. *Molecular Cell*, *81*(17), 3637-3649.e5. https://doi.org/10.1016/j.molcel.2021.07.035

Dziubańska-Kusibab, P. J., Berger, H., Battistini, F., Bouwman, B. A. M., Iftekhar, A., Katainen, R., Cajuso, T., Crosetto, N., Orozco, M., Aaltonen, L. A., & Meyer, T. F. (2020). Colibactin DNA-damage signature indicates mutational impact in colorectal cancer. *Nature Medicine*, *26*(7), 1063–1069. https://doi.org/10.1038/s41591-020-0908-2

Ewaisha, R., & Anderson, K. S. (2023). Immunogenicity of CRISPR therapeutics—Critical considerations for clinical translation. *Frontiers in Bioengineering and Biotechnology*, *11*. https://doi.org/10.3389/fbioe.2023.1138596

Fiumara, M., Ferrari, S., Omer-Javed, A., Beretta, S., Albano, L., Canarutto, D., Varesi, A., Gaddoni, C., Brombin, C., Cugnata, F., Zonari, E., Naldini, M. M., Barcella, M., Gentner, B., Merelli, I., & Naldini, L. (2023). Genotoxic effects of base and prime editing in human hematopoietic stem cells. *Nature Biotechnology*, 1–15. https://doi.org/10.1038/s41587-023-01915-4

Frangoul, H., Altshuler, D., Cappellini, M. D., Chen, Y.-S., Domm, J., Eustace, B. K., Foell, J., de la Fuente, J., Grupp, S., Handgretinger, R., Ho, T. W., Kattamis, A., Kernytsky, A., Lekstrom-Himes, J., Li, A. M., Locatelli, F., Mapara, M. Y., de Montalembert, M., Rondelli, D., … Corbacioglu, S. (2021). CRISPR-Cas9 Gene Editing for Sickle Cell Disease and β-Thalassemia. *The New England Journal of Medicine*, *384*(3), 252–260. https://doi.org/10.1056/NEJMoa2031054

Gasiunas, G., Barrangou, R., Horvath, P., & Siksnys, V. (2012). Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(39), 2579–2586. https://doi.org/10.1073/pnas.1208507109

Gaudelli, N. M., Komor, A. C., Rees, H. A., Packer, M. S., Badran, A. H., Bryson, D. I., & Liu, D. R. (2017). Programmable base editing of T to G C in genomic DNA without DNA cleavage. *Nature*, *551*(7681), 464–471. https://doi.org/10.1038/nature24644

Gersbach, C. A., Gaj, T., & Barbas, C. F. I. (2014). Synthetic Zinc Finger Proteins: The Advent of Targeted Gene Regulation and Genome Modification Technologies. *Accounts of Chemical Research*, *47*(8), 2309–2318. https://doi.org/10.1021/ar500039w

Gillmore Julian D., Gane Ed, Taubel Jorg, Kao Justin, Fontana Marianna, Maitland Michael L., Seitzer Jessica, O'Connell Daniel, Walsh Kathryn R., Wood Kristy, Phillips Jonathan, Xu Yuanxin, Amaral Adam, Boyd Adam P., Cehelsky Jeffrey E., McKee Mark D., Schiermeier Andrew, Harari Olivier, Murphy Andrew, … Lebwohl David. (2021). CRISPR-Cas9 In Vivo Gene Editing for Transthyretin Amyloidosis. *New England Journal of Medicine*, *385*(6), 493–502. https://doi.org/10.1056/NEJMoa2107454

Gisler, S., Gonçalves, J. P., Akhtar, W., de Jong, J., Pindyurin, A. V., Wessels, L. F. A., & van Lohuizen, M. (2019). Multiplexed Cas9 targeting reveals genomic location effects and gRNA-based staggered breaks influencing mutation efficiency. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-09551-w

Gothe, H. J., Bouwman, B. A. M., Gusmao, E. G., Piccinno, R., Petrosino, G., Sayols, S., Drechsel, O., Minneker, V., Josipovic, N., Mizi, A., Nielsen, C. F., Wagner, E. M., Takeda, S., Sasanuma, H., Hudson, D. F., Kindler, T., Baranello, L., Papantonis, A., Crosetto, N., & Roukos, V. (2019). Spatial Chromosome Folding and Active Transcription Drive DNA Fragility and Formation of Oncogenic MLL Translocations. *Molecular Cell*, *75*(2), 267-283.e12. https://doi.org/10.1016/j.molcel.2019.05.015

Guo, C., Ma, X., Gao, F., & Guo, Y. (2023). Off-target effects in CRISPR/Cas9 gene editing. *Frontiers in Bioengineering and Biotechnology*, *11*, 1143157. https://doi.org/10.3389/fbioe.2023.1143157

Hennig, B. P., Velten, L., Racke, I., Tu, C. S., Thoms, M., Rybin, V., Besir, H., Remans, K., & Steinmetz, L. M. (2018). Large-scale low-cost NGS library preparation using a robust Tn5 purification and tagmentation protocol. *G3: Genes, Genomes, Genetics*, *8*(1), 79–89. https://doi.org/10.1534/g3.117.300257

Heyer, W.-D., Ehmsen, K. T., & Liu, J. (2010). Regulation of homologous recombination in eukaryotes. *Annual Review of Genetics*, *44*, 113–139. https://doi.org/10.1146/annurev-genet-051710-150955

Hoeijmakers, J. H. J. (2009). DNA damage, aging, and cancer. *The New England Journal of Medicine*, *361*(15), 1475–1485. https://doi.org/10.1056/NEJMra0804615

Horlbeck, M. A., Witkowsky, L. B., Guglielmi, B., Replogle, J. M., Gilbert, L. A., Villalta, J. E., Torigoe, S. E., Tjian, R., & Weissman, J. S. (2016). Nucleosomes impede Cas9 access to DNA in vivo and in vitro. *eLife*, *5*, e12677. https://doi.org/10.7554/eLife.12677

Hu, J. H., Miller, S. M., Geurts, M. H., Tang, W., Chen, L., Sun, N., Zeina, C. M., Gao, X., Rees, H. A., Lin, Z., & Liu, D. R. (2018). Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. *Nature*, *556*(7699), 57–63. https://doi.org/10.1038/nature26155

Huang, M. E., Qin, Y., Shang, Y., Hao, Q., Zhan, C., Lian, C., Luo, S., Liu, L. D., Zhang, S., Zhang, Y., Wo, Y., Li, N., Wu, S., Gui, T., Wang, B., Luo, Y., Cai, Y., Liu, X., Xu, Z., … Meng, F.-L. (2024). C-to-G editing generates double-strand breaks causing deletion, transversion and translocation. *Nature Cell Biology*, *26*(2), 294–304. https://doi.org/10.1038/s41556-023-01342-2

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Oleś, A. K., … Morgan, M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, *12*(2), 115–121. https://doi.org/10.1038/nmeth.3252

Hussmann, J. A., Ling, J., Ravisankar, P., Yan, J., Cirincione, A., Xu, A., Simpson, D., Yang, D., Bothmer, A., Cotta-Ramusino, C., Weissman, J. S., & Adamson, B. (2021). Mapping the genetic

landscape of DNA double-strand break repair. *Cell*, *184*(22), 5653-5669.e25. https://doi.org/10.1016/j.cell.2021.10.002

Ishino, Y., Shinagawa, H., Makino, K., Amemura, M., & Nakata, A. (1987). Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. *Journal of Bacteriology*, *169*(12), 5429–5433. https://doi.org/10.1128/jb.169.12.5429-5433.1987

Ivanov, I. E., Wright, A. V., Cofsky, J. C., Palacio Aris, K. D., Doudna, J. A., & Bryant, Z. (2020). Cas9 interrogates DNA in discrete steps modulated by mismatches and supercoiling. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(11), 5853–5860. https://doi.org/10.1073/pnas.1913445117

Jaenisch, R., & Mintz, B. (1974). Simian Virus 40 DNA Sequences in DNA of Healthy Adult Mice Derived from Preimplantation Blastocysts Injected with Viral DNA. *Proceedings of the National Academy of Sciences of the United States of America*, *71*(4), 1250–1254.

Janssen, A., van der Burg, M., Szuhai, K., Kops, G. J. P. L., & Medema, R. H. (2011). Chromosome Segregation Errors as a Cause of DNA Damage and Structural Chromosome Aberrations. *Science*, *333*(6051), 1895–1898. https://doi.org/10.1126/science.1210214

Jiang, F., & Doudna, J. A. (2017). *CRISPR – Cas9 Structures and Mechanisms*. 505–531.

Jiang, F., Taylor, D. W., Chen, J. S., Kornfeld, J. E., Zhou, K., Thompson, A. J., Nogales, E., & Doudna, J. A. (2016). Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science*, *351*(6275), 867–871. https://doi.org/10.1126/science.aad8282

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science*, *337*(6096), 816–821. https://doi.org/10.1126/science.1225829

Jinek, M., Jiang, F., Taylor, D. W., Sternberg, S. H., Kaya, E., Ma, E., Anders, C., Hauer, M., Zhou, K., Lin, S., Kaplan, M., Iavarone, A. T., Charpentier, E., Nogales, E., & Doudna, J. A. (2014). Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science (New York, N.Y.)*, *343*(6176), 1247997. https://doi.org/10.1126/science.1247997

Jr, S. K. J., Hawkins, J. A., Johnson, N. V., Jung, C., Hu, K., Rybarski, J. R., Chen, J. S., Doudna, J. A., Press, W. H., & Finkelstein, I. J. (2021). Massively parallel kinetic profiling of natural and engineered CRISPR nucleases. *Nature Biotechnology*, *39*(January). https://doi.org/10.1038/s41587-020-0646-5

Keough, K. C., Lyalina, S., Olvera, M. P., Whalen, S., Conklin, B. R., & Pollard, K. S. (2019). AlleleAnalyzer: A tool for personalized and allele-specific sgRNA design. *Genome Biology*, *20*, 167. https://doi.org/10.1186/s13059-019-1783-3

Kim, D., Bae, S., Park, J., Kim, E., Kim, S., Yu, H. R., Hwang, J., Kim, J. I., & Kim, J. S. (2015). Digenome-seq: Genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nature Methods*, *12*(3), 237–243. https://doi.org/10.1038/nmeth.3284

Kim, D., Kang, B. C., & Kim, J. S. (2021). Identifying genome-wide off-target sites of CRISPR RNA–guided nucleases and deaminases with Digenome-seq. *Nature Protocols*, *16*(2), 1170–1192. https://doi.org/10.1038/s41596-020-00453-6

Kim, D., Kim, S., Kim, S., Park, J., & Kim, J. S. (2016). Genome-wide target specificities of CRISPR-Cas9 nucleases revealed by multiplex Digenome-seq. *Genome Research*, *26*(3), 406–415. https://doi.org/10.1101/gr.199588.115

Kim, D., Lim, K., Kim, S. T., Yoon, S. H., Kim, K., Ryu, S. M., & Kim, J. S. (2017). Genome-wide target specificities of CRISPR RNA-guided programmable deaminases. *Nature Biotechnology*, *35*(5), 475–480. https://doi.org/10.1038/nbt.3852

Kim, Y. G., Cha, J., & Chandrasegaran, S. (1996). Hybrid restriction enzymes: Zinc finger fusions to Fok I cleavage domain. *Proceedings of the National Academy of Sciences*, *93*(3), 1156–1160. https://doi.org/10.1073/pnas.93.3.1156

Koeppel, J., Weller, J., Peets, E. M., Pallaseni, A., Kuzmin, I., Raudvere, U., Peterson, H., Liberante, F. G., & Parts, L. (2023). Prediction of prime editing insertion efficiencies using sequence features and DNA repair determinants. *Nature Biotechnology*, *41*(10), 1446–1456. https://doi.org/10.1038/s41587-023-01678-y

Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A., & Liu, D. R. (2016). Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature*, *533*(7603), 420–424. https://doi.org/10.1038/nature17946

Krysler, A. R., Cromwell, C. R., Tu, T., Jovel, J., & Hubbard, B. P. (2022). Guide RNAs containing universal bases enable Cas9/Cas12a recognition of polymorphic sequences. *Nature Communications*, *13*(1), 1–13. https://doi.org/10.1038/s41467-022-29202-x

Kulcsár, P. I., Tálas, A., Ligeti, Z., Krausz, S. L., & Welker, E. (2022). SuperFi-Cas9 exhibits remarkable fidelity but severely reduced activity yet works effectively with ABE8e. *Nature Communications*, *13*(1), 6858. https://doi.org/10.1038/s41467-022-34527-8

Lander, E. S. (2016). The Heroes of CRISPR. *Cell*, *164*(1), 18–28. https://doi.org/10.1016/j.cell.2015.12.041

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., & Carey, V. J. (2013). Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology*, *9*(8), e1003118. https://doi.org/10.1371/journal.pcbi.1003118

Lazzarotto, C. R., Malinin, N. L., Li, Y., Zhang, R., Yang, Y., Lee, G. H., Cowley, E., He, Y., Lan, X., Jividen, K., Katta, V., Kolmakova, N. G., Petersen, C. T., Qi, Q., Strelcov, E., Maragh, S., Krenciute, G., Ma, J., Cheng, Y., & Tsai, S. Q. (2020). CHANGE-seq reveals genetic and epigenetic effects on CRISPR–Cas9 genome-wide activity. *Nature Biotechnology*, *38*(11), 1317–1327. https://doi.org/10.1038/s41587-020-0555-7

Lazzarotto, C. R., Nguyen, N. T., Tang, X., Malagon-Lopez, J., Guo, J. A., Aryee, M. J., Joung, J. K., & Tsai, S. Q. (2018). Defining CRISPR–Cas9 genome-wide nuclease activities with CIRCLE-seq. *Nature Protocols*, *13*(11), Article 11. https://doi.org/10.1038/s41596-018-0055-0

Lee, J. K., Jeong, E., Lee, J., Jung, M., Shin, E., hoon Kim, Y., Lee, K., Jung, I., Kim, D., Kim, S., & Kim, J. S. (2018). Directed evolution of CRISPR-Cas9 to increase its specificity. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-05477-x

Lee, R. G., Mazzola, A. M., Braun, M. C., Platt, C., Vafai, S. B., Kathiresan, S., Rohde, E., Bellinger, A. M., & Khera, A. V. (2023). Efficacy and Safety of an Investigational Single-Course CRISPR Base-Editing Therapy Targeting PCSK9 in Nonhuman Primate and Mouse Models. *Circulation*, *147*(3), 242–253. https://doi.org/10.1161/CIRCULATIONAHA.122.062132

Leenay, R. T., Aghazadeh, A., Hiatt, J., Tse, D., Roth, T. L., Apathy, R., Shifrut, E., Hultquist, J. F., Krogan, N., Wu, Z., Cirolia, G., Canaj, H., Leonetti, M. D., Marson, A., May, A. P., & Zou, J. (2019). Large dataset enables prediction of repair after CRISPR–Cas9 editing in primary T cells. *Nature Biotechnology*, *37*(9), 1034–1037. https://doi.org/10.1038/s41587-019-0203-2

Lei, Z., Meng, H., Lv, Z., Liu, M., Zhao, H., Wu, H., Zhang, X., Liu, L., Zhuang, Y., Yin, K., Yan, Y., & Yi, C. (2021). Detect-seq reveals out-of-protospacer editing and target-strand editing by cytosine base editors. *Nature Methods*, *18*(6), 643–651. https://doi.org/10.1038/s41592-021-01172-w

Leibowitz, M. L., Papathanasiou, S., Doerfler, P. A., Blaine, L. J., Sun, L., Yao, Y., Zhang, C. Z., Weiss, M. J., & Pellman, D. (2021). Chromothripsis as an on-target consequence of CRISPR–Cas9

genome editing. *Nature Genetics*, *53*(6), 895–905. https://doi.org/10.1038/s41588-021-00838-7

Lemaître, C., Grabarz, A., Tsouroula, K., Andronov, L., Furst, A., Pankotai, T., Heyer, V., Rogier, M., Attwood, K. M., Kessler, P., Dellaire, G., Klaholz, B., Reina-San-Martin, B., & Soutoglou, E. (2014). Nuclear position dictates DNA repair pathway choice. *Genes and Development*, *28*(22), 2450–2463. https://doi.org/10.1101/gad.248369.114

Lemaître, C., & Soutoglou, E. (2014). Double strand break (DSB) repair in heterochromatin and heterochromatin proteins in DSB repair. *DNA Repair*, *19*, 163–168. https://doi.org/10.1016/j.dnarep.2014.03.015

Lemos, B. R., Kaplan, A. C., Bae, J. E., Ferrazzoli, A. E., Kuo, J., Anand, R. P., Waterman, D. P., & Haber, J. E. (2018). CRISPR/Cas9 cleavages in budding yeast reveal templated insertions and strand-specific insertion/deletion profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(9), E2010–E2047. https://doi.org/10.1073/pnas.1716855115

Lessard, S., Francioli, L., Alfoldi, J., Tardif, J.-C., Ellinor, P. T., MacArthur, D. G., Lettre, G., Orkin, S. H., & Canver, M. C. (2017). Human genetic variation alters CRISPR-Cas9 on- and off-targeting specificity at therapeutically implicated loci. *Proceedings of the National Academy of Sciences*, *114*(52). https://doi.org/10.1073/pnas.1714640114

Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM* (arXiv:1303.3997). arXiv. http://arxiv.org/abs/1303.3997

Liang, P., Xie, X., Zhi, S., Sun, H., Zhang, X., Chen, Y., Chen, Y., Xiong, Y., Ma, W., Liu, D., Huang, J., & Songyang, Z. (2019). Genome-wide profiling of adenine base editor specificity by EndoV-seq. *Nature Communications*, *10*(1), 1–9. https://doi.org/10.1038/s41467-018-07988-z

Liu, M., Rehman, S., Tang, X., Gu, K., Fan, Q., Chen, D., & Ma, W. (2019). Methodologies for Improving HDR Efficiency. *Frontiers in Genetics*, *9*. https://doi.org/10.3389/fgene.2018.00691

Liu, R., Liang, L., Freed, E. F., & Gill, R. T. (2021). Directed Evolution of CRISPR/Cas Systems for Precise Gene Editing. *Trends in Biotechnology*, *39*(3), 262–273. https://doi.org/10.1016/j.tibtech.2020.07.005

Longo, G. M. C., Sayols, S., Kotini, A. G., Heinen, S., Möckel, M. M., Beli, P., & Roukos, V. (2024). Linking CRISPR-Cas9 double-strand break profiles to gene editing precision with BreakTag. *Nature Biotechnology*. https://doi.org/10.1038/s41587-024-02238-8

Makarova, K. S., Wolf, Y. I., Iranzo, J., Shmakov, S. A., Alkhnbashi, O. S., Brouns, S. J. J., Charpentier, E., Cheng, D., Haft, D. H., Horvath, P., Moineau, S., Mojica, F. J. M., Scott, D., Shah, S. A., Siksnys, V., Terns, M. P., Venclovas, Č., White, M. F., Yakunin, A. F., … Koonin, E. V. (2020). Evolutionary classification of CRISPR–Cas systems: A burst of class 2 and derived variants. *Nature Reviews Microbiology*, *18*(2), Article 2. https://doi.org/10.1038/s41579-019-0299-x

Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., Norville, J. E., & Church, G. M. (2013). RNA-Guided Human Genome Engineering via Cas9. *Science*, *339*(6121), 823–826. https://doi.org/10.1126/science.1232033

Malinin, N. L., Lee, G., Lazzarotto, C. R., Li, Y., Zheng, Z., Nguyen, N. T., Liebers, M., Topkar, V. V., Iafrate, A. J., Le, L. P., Aryee, M. J., Joung, J. K., & Tsai, S. Q. (2021). Defining genome-wide CRISPR-Cas genome-editing nuclease activity with GUIDE-seq. *Nature Protocols*, *16*(12), 5592–5615. https://doi.org/10.1038/s41596-021-00626-x

McKenna, A., & Shendure, J. (2018). FlashFry: A fast and flexible tool for large-scale CRISPR target design. *BMC Biology*, *16*(1), 74. https://doi.org/10.1186/s12915-018-0545-0

Mehryar, M. M., Shi, X., Li, J., & Wu, Q. (2023). DNA polymerases in precise and predictable CRISPR/Cas9-mediated chromosomal rearrangements. *BMC Biology*, *21*, 288. https://doi.org/10.1186/s12915-023-01784-y

Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J., & Soria, E. (2005). Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements. *Journal of Molecular Evolution*, *60*(2), 174–182. https://doi.org/10.1007/s00239-004-0046-3

Mojica, F. j. m., Ferrer, C., Juez, G., & Rodríguez-Valera, F. (1995). Long stretches of short tandem repeats are present in the largest replicons of the Archaea Haloferax mediterranei and Haloferax volcanii and could be involved in replicon partitioning. *Molecular Microbiology*, *17*(1), 85–93. https://doi.org/10.1111/j.1365-2958.1995.mmi_17010085.x

Mojica, F. J. M., Juez, G., & Rodriguez-Valera, F. (1993). Transcription at different salinities of Haloferax mediterranei sequences adjacent to partially modified PstI sites. *Molecular Microbiology*, *9*(3), 613–621. https://doi.org/10.1111/j.1365-2958.1993.tb01721.x

Mojica, F. J. M., & Rodriguez-Valera, F. (2016). The discovery of CRISPR in archaea and bacteria. *The FEBS Journal*, *283*(17), 3162–3169. https://doi.org/10.1111/febs.13766

Molla, K. A., & Yang, Y. (2020). Predicting CRISPR/Cas9-Induced Mutations for Precise Genome Editing. *Trends in Biotechnology*, *38*(2), 136–141. https://doi.org/10.1016/j.tibtech.2019.08.002

Montague, T. G., Cruz, J. M., Gagnon, J. A., Church, G. M., & Valen, E. (2014). CHOPCHOP: A CRISPR/Cas9 and TALEN web tool for genome editing. *Nucleic Acids Research*, *42*(W1), W401–W407. https://doi.org/10.1093/nar/gku410

Moore, R., Chandrahas, A., & Bleris, L. (2014). Transcription Activator-like Effectors: A Toolkit for Synthetic Biology. *ACS Synthetic Biology*, *3*(10), 708–716. https://doi.org/10.1021/sb400137b

Mosler, T., Conte, F., Longo, G. M. C., Mikicic, I., Kreim, N., Möckel, M. M., Petrosino, G., Flach, J., Barau, J., Luke, B., Roukos, V., & Beli, P. (2021). R-loop proximity proteomics identifies a role of DDX41 in transcription-associated genomic instability. *Nature Communications*, *12*(1), 7314. https://doi.org/10.1038/s41467-021-27530-y

Müthel, S., Marg, A., Ignak, B., Kieshauer, J., Escobar, H., Stadelmann, C., & Spuler, S. (2023). Cas9-induced single cut enables highly efficient and template-free repair of a muscular dystrophy causing founder mutation. *Molecular Therapy - Nucleic Acids*, *31*, 494–511. https://doi.org/10.1016/j.omtn.2023.02.005

Naeem, M., Majeed, S., Hoque, M. Z., & Ahmad, I. (2020). Latest Developed Strategies to Minimize the Off-Target Effects in CRISPR-Cas-Mediated Genome Editing. *Cells*, *9*(7), 1608. https://doi.org/10.3390/cells9071608

Nakade, S., Nakamae, K., Tang, T.-C., Yu, D., Sakuma, T., Yamamoto, T., & Lu, T. K. (2022). *Frame Editors for Precise, Template-Free Frameshifting* (p. 2022.12.05.518807). bioRxiv. https://doi.org/10.1101/2022.12.05.518807

Naso, M. F., Tomkowicz, B., Perry, W. L., & Strohl, W. R. (2017). Adeno-Associated Virus (AAV) as a Vector for Gene Therapy. *Biodrugs*, *31*(4), 317–334. https://doi.org/10.1007/s40259-017-0234-5

Pacesa, M., Lin, C.-H., Cléry, A., Saha, A., Arantes, P. R., Bargsten, K., Irby, M. J., Allain, F. H.-T., Palermo, G., Cameron, P., Donohoue, P. D., & Jinek, M. (2022). Structural basis for Cas9 off-target activity. *Cell*, *185*(22), 4067-4081.e21. https://doi.org/10.1016/j.cell.2022.09.026

Papapetrou, E. P., & Sadelain, M. (2011). Generation of transgene-free human induced pluripotent stem cells with an excisable single polycistronic vector. *Nature Protocols*, *6*(9), 1251–1273. https://doi.org/10.1038/nprot.2011.374

Picelli, S., Björklund, A. K., Reinius, B., Sagasser, S., Winberg, G., & Sandberg, R. (2014). Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Research*, *24*(12), 2033–2040. https://doi.org/10.1101/gr.177881.114

Pingoud, A., & Silva, G. H. (2007). Precision genome surgery. *Nature Biotechnology*, *25*(7), 743–744. https://doi.org/10.1038/nbt0707-743

Python Interface for H2O. module version 3.38.0.2. (2016). [Jupyter Notebook]. H2O.ai. https://github.com/h2oai/h2o-3 (Original work published 2014)

Qasim, W., Zhan, H., Samarasinghe, S., Adams, S., Amrolia, P., Stafford, S., Butler, K., Rivat, C., Wright, G., Somana, K., Ghorashian, S., Pinner, D., Ahsan, G., Gilmour, K., Lucchini, G., Inglott, S., Mifsud, W., Chiesa, R., Peggs, K. S., … Veys, P. (2017). Molecular remission of infant B-ALL after infusion of universal TALEN gene-edited CAR T cells. *Science Translational Medicine*, *9*(374), eaaj2013. https://doi.org/10.1126/scitranslmed.aaj2013

Rashmi, K. V., & Gilad-Bachrach, R. (2015). *DART: Dropouts meet Multiple Additive Regression Trees* (arXiv:1505.01866). arXiv. http://arxiv.org/abs/1505.01866

Riesenberg, S., Kanis, P., Macak, D., Wollny, D., Düsterhöft, D., Kowalewski, J., Helmbrecht, N., Maricic, T., & Pääbo, S. (2023). Efficient high-precision homology-directed repair-dependent genome editing by HDRobust. *Nature Methods*, *20*(9), 1388–1399. https://doi.org/10.1038/s41592-023-01949-1

Rouet, P., Smih, F., & Jasin, M. (1994). Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. *Molecular and Cellular Biology*, *14*(12), 8096–8106.

Saha, K. (2023). Accounting for diversity in the design and development of CRISPR-based therapeutic genome editing. *Nature Genetics*, *55*(1), 6–7. https://doi.org/10.1038/s41588-022-01272-z

Schep, R., Brinkman, E. K., Leemans, C., Vergara, X., van der Weide, R. H., Morris, B., van Schaik, T., Manzo, S. G., Peric-Hupkes, D., van den Berg, J., Beijersbergen, R. L., Medema, R. H., & van Steensel, B. (2021). Impact of chromatin context on Cas9-induced DNA double-strand break repair pathway balance. *Molecular Cell*, *81*(10), 2216-2230.e10. https://doi.org/10.1016/j.molcel.2021.03.032

Schmid-Burgk, J. L., Gao, L., Li, D., Gardner, Z., Strecker, J., Lash, B., & Zhang, F. (2020). Highly Parallel Profiling of Cas9 Variant Specificity. *Molecular Cell*, *78*(4), 794-800.e8. https://doi.org/10.1016/j.molcel.2020.02.023

Scully, R., Panday, A., Elango, R., & Willis, N. A. (2019). DNA double-strand break repair-pathway choice in somatic mammalian cells. *Nature Reviews Molecular Cell Biology*, *20*(11), 698–714. https://doi.org/10.1038/s41580-019-0152-0

Sfeir, A., & Symington, L. S. (2015). Microhomology-mediated end joining: A back-up survival mechanism or dedicated pathway? *Trends in Biochemical Sciences*, *40*(11), 701–714. https://doi.org/10.1016/j.tibs.2015.08.006

Shen, M. W., Arbab, M., Hsu, J. Y., Worstell, D., Culbertson, S. J., Krabbe, O., Cassa, C. A., Liu, D. R., Gifford, D. K., & Sherwood, R. I. (2018). Predictable and precise template-free CRISPR editing of pathogenic variants. *Nature*, *563*(7733), 646–651. https://doi.org/10.1038/s41586-018-0686-x

Sherkatghanad, Z., Abdar, M., Charlier, J., & Makarenkov, V. (2023). Using traditional machine learning and deep learning methods for on- and off-target prediction in CRISPR/Cas9: A review. *Briefings in Bioinformatics*, *24*(3), bbad131. https://doi.org/10.1093/bib/bbad131

Shi, X., Shou, J., Mehryar, M. M., Li, J., Wang, L., Zhang, M., Huang, H., Sun, X., & Wu, Q. (2019). Cas9 has no exonuclease activity resulting in staggered cleavage with overhangs and predictable di- and tri-nucleotide CRISPR insertions without template donor. *Cell Discovery*, *5*(1), 4–7. https://doi.org/10.1038/s41421-019-0120-z

Shou, J., Li, J., Liu, Y., & Wu, Q. (2018). Precise and Predictable CRISPR Chromosomal Rearrangements Reveal Principles of Cas9-Mediated Nucleotide Insertion. *Molecular Cell*, *71*(4), 498-509.e4. https://doi.org/10.1016/j.molcel.2018.06.021

Slaman, E., Lammers, M., Angenent, G. C., & de Maagd, R. A. (2023). High-throughput sgRNA testing reveals rules for Cas9 specificity and DNA repair in tomato cells. *Frontiers in Genome Editing*, *5*. https://doi.org/10.3389/fgeed.2023.1196763

Slesarenko, Y. S., Lavrov, A. V., & Smirnikhina, S. A. (2022). Off-target effects of base editors: What we know and how we can reduce it. *Current Genetics*, *68*(1), 39–48. https://doi.org/10.1007/s00294-021-01211-1

Stadtmauer, E. A., Fraietta, J. A., Davis, M. M., Cohen, A. D., Weber, K. L., Lancaster, E., Mangan, P. A., Kulikovskaya, I., Gupta, M., Chen, F., Tian, L., Gonzalez, V. E., Xu, J., young Jung, I., Joseph Melenhorst, J., Plesa, G., Shea, J., Matlawski, T., Cervini, A., … June, C. H. (2020). CRISPR-engineered T cells in patients with refractory cancer. *Science*, *367*(6481). https://doi.org/10.1126/science.aba7365

Stephenson, A. A., Raper, A. T., & Suo, Z. (2018). Bidirectional Degradation of DNA Cleavage Products Catalyzed by CRISPR/Cas9. *Journal of the American Chemical Society*, *140*(10), 3743–3750. https://doi.org/10.1021/jacs.7b13050

Swarts, D. C. (2019). Making the cut(s): How Cas12a cleaves target and non-target DNA. *Biochemical Society Transactions*, *47*(5), 1499–1510. https://doi.org/10.1042/BST20190564

Taheri-Ghahfarokhi, A., Taylor, B. J. M., Nitsch, R., Lundin, A., Cavallo, A.-L., Madeyski-Bengtson, K., Karlsson, F., Clausen, M., Hicks, R., Mayr, L. M., Bohlooly-Y, M., & Maresca, M. (2018). Decoding non-random mutational signatures at Cas9 targeted sites. *Nucleic Acids Research*, *46*(16), 8417–8434. https://doi.org/10.1093/nar/gky653

Tebas Pablo, Stein David, Tang Winson W., Frank Ian, Wang Shelley Q., Lee Gary, Spratt S. Kaye, Surosky Richard T., Giedlin Martin A., Nichol Geoff, Holmes Michael C., Gregory Philip D., Ando Dale G., Kalos Michael, Collman Ronald G., Binder-Scholl Gwendolyn, Plesa Gabriela, Hwang Wei-Ting, Levine Bruce L., & June Carl H. (2014). Gene Editing of CCR5 in Autologous CD4 T Cells of Persons Infected with HIV. *New England Journal of Medicine*, *370*(10), 901–910. https://doi.org/10.1056/NEJMoa1300662

Thomas, K. R., & Capecchi, M. R. (1987). Site-directed mutagenesis by gene targeting in mouse embryo-derived stem cells. *Cell*, *51*(3), 503–512. https://doi.org/10.1016/0092-8674(87)90646-5

Tsai, S. Q., Nguyen, N. T., Malagon-Lopez, J., Topkar, V. V., Aryee, M. J., & Joung, J. K. (2017). CIRCLE-seq: A highly sensitive in vitro screen for genome-wide CRISPR-Cas9 nuclease off-targets. *Nature Methods*, *14*(6), 607–614. https://doi.org/10.1038/nmeth.4278

Tsai, S. Q., Zheng, Z., Nguyen, N. T., Liebers, M., Topkar, V. V., Thapar, V., Wyvekens, N., Khayter, C., Iafrate, A. J., Le, L. P., Aryee, M. J., & Joung, J. K. (2015). GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nature Biotechnology*, *33*(2), 187–198. https://doi.org/10.1038/nbt.3117

Urnov, F. D., Miller, J. C., Lee, Y.-L., Beausejour, C. M., Rock, J. M., Augustus, S., Jamieson, A. C., Porteus, M. H., Gregory, P. D., & Holmes, M. C. (2005). Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature*, *435*(7042), 646–651. https://doi.org/10.1038/nature03556

U.S. Food and Drug Administration. (2023, December 8). *FDA Approves First Gene Therapies to Treat Patients with Sickle Cell Disease*. FDA; FDA. https://www.fda.gov/news-events/press-announcements/fda-approves-first-gene-therapies-treat-patients-sickle-cell-disease

Vakulskas, C. A., Dever, D. P., Rettig, G. R., Turk, R., Jacobi, A. M., Collingwood, M. A., Bode, N. M., McNeill, M. S., Yan, S., Camarena, J., Lee, C. M., Park, S. H., Wiebking, V., Bak, R. O., Gomez-Ospina, N., Pavel-Dinu, M., Sun, W., Bao, G., Porteus, M. H., & Behlke, M. A. (2018). A high-fidelity Cas9 mutant delivered as a ribonucleoprotein complex enables efficient gene editing in human hematopoietic stem and progenitor cells. *Nature Medicine*, *24*(8), 1216–1224. https://doi.org/10.1038/s41591-018-0137-0

van Overbeek, M., Capurso, D., Carter, M. M., Thompson, M. S., Frias, E., Russ, C., Reece-Hoyes, J. S., Nye, C., Gradia, S., Vidal, B., Zheng, J., Hoffman, G. R., Fuller, C. K., & May, A. P. (2016). DNA Repair Profiling Reveals Nonrandom Outcomes at Cas9-Mediated Breaks. *Molecular Cell*, *63*(4), 633–646. https://doi.org/10.1016/j.molcel.2016.06.037

Wagih, O. (2023). *ggseqlogo: A "ggplot2" Extension for Drawing Publication-Ready Sequence Logos* [R]. https://github.com/omarwagih/ggseqlogo (Original work published 2017)

Walton, R. T., Christie, K. A., Whittaker, M. N., & Kleinstiver, B. P. (2020). Unconstrained genome targeting with near-PAMless engineered CRISPR-Cas9 variants. *Science (New York, N.Y.)*, *368*(6488), 290–296. https://doi.org/10.1126/science.aba8853

Westermann, L., Neubauer, B., & Köttgen, M. (2021). Nobel Prize 2020 in Chemistry honors CRISPR: A tool for rewriting the code of life. *Pflugers Archiv*, *473*(1), 1–2. https://doi.org/10.1007/s00424-020-02497-9

Wienert, B., Wyman, S. K., Yeh, C. D., Conklin, B. R., & Corn, J. E. (2020). CRISPR off-target detection with DISCOVER-seq. *Nature Protocols*, *15*(5), 1775–1799. https://doi.org/10.1038/s41596-020-0309-5

Wong, C. (2023). UK first to approve CRISPR treatment for diseases: What you need to know. *Nature*, *623*(7988), 676–677. https://doi.org/10.1038/d41586-023-03590-6

Xue, C., & Greene, E. C. (2021). DNA Repair Pathway Choices in CRISPR-Cas9-Mediated Genome Editing. *Trends in Genetics*, *37*(7), 639–656. https://doi.org/10.1016/j.tig.2021.02.008

Yan, W. X., Mirzazadeh, R., Garnerone, S., Scott, D., Schneider, M. W., Kallas, T., Custodio, J., Wernersson, E., Li, Y., Gao, L., Federova, Y., Zetsche, B., Zhang, F., Bienko, M., & Crosetto, N. (2017). BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks. *Nature Communications*, *8*(May), 1–9. https://doi.org/10.1038/ncomms15058

Yang, Q., Abebe, J. S., Mai, M., Rudy, G., Kim, S. Y., Devinsky, O., & Long, C. (2023). *Phage DNA polymerase prevents deleterious on-target DNA damage and enhances precise CRISPR/Cas9 editing* (p. 2023.01.10.523496). bioRxiv. https://doi.org/10.1101/2023.01.10.523496

Yarrington, R. M., Verma, S., Schwartz, S., Trautman, J. K., & Carroll, D. (2018). Nucleosomes inhibit target cleavage by CRISPR-Cas9 in vivo. *Proceedings of the National Academy of Sciences*, *115*(38), 9351–9358. https://doi.org/10.1073/pnas.1810062115

Yau, E. H., & Rana, T. M. (2018). Next-Generation Sequencing of Genome-Wide CRISPR Screens. *Methods in Molecular Biology (Clifton, N.J.)*, *1712*, 203–216. https://doi.org/10.1007/978-1-4939-7514-3_13

Yilmaz, D., Furst, A., Meaburn, K., Lezaja, A., Wen, Y., Altmeyer, M., Reina-San-Martin, B., & Soutoglou, E. (2021). Activation of homologous recombination in G1 preserves centromeric integrity. *Nature*, *600*(7890), 748–753. https://doi.org/10.1038/s41586-021-04200-z

Zetsche, B., Gootenberg, J. S., Abudayyeh, O. O., Slaymaker, I. M., Makarova, K. S., Essletzbichler, P., Volz, S. E., Joung, J., van der Oost, J., Regev, A., Koonin, E. V., & Zhang, F. (2015). Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell*, *163*(3), 759–771. https://doi.org/10.1016/j.cell.2015.09.038

Zhang, S., Wang, Y., Mao, D., Wang, Y., Zhang, H., Pan, Y., Wang, Y., Teng, S., & Huang, P. (2023). Current trends of clinical trials involving CRISPR/Cas systems. *Frontiers in Medicine*, *10*. https://doi.org/10.3389/fmed.2023.1292452

Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., Weng, Z., Liu, Y., Mason, C. E., Alexander, N., Henaff, E., McIntyre, A. B. R., Chandramohan, D., Chen, F., Jaeger, E., Moshrefi, A., Pham, K., Stedman, W., Liang, T., … Salit, M. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data*, *3*(1), 160025. https://doi.org/10.1038/sdata.2016.25

Zook, J. M., McDaniel, J., Olson, N. D., Wagner, J., Parikh, H., Heaton, H., Irvine, S. A., Trigg, L., Truty, R., McLean, C. Y., De La Vega, F. M., Xiao, C., Sherry, S., & Salit, M. (2019). An open resource for accurately benchmarking small variant and reference calls. *Nature Biotechnology*, *37*(5), 561–566. https://doi.org/10.1038/s41587-019-0074-6

Zou, R. S., Liu, Y., Wu, B., & Ha, T. (2021). Cas9 deactivation with photocleavable guide RNAs. *Molecular Cell*, *81*(7), 1553-1565.e8. https://doi.org/10.1016/j.molcel.2021.02.007

Zuo, Z., & Liu, J. (2016). Cas9-catalyzed DNA Cleavage Generates Staggered Ends: Evidence from Molecular Dynamics Simulations. *Nature Publishing Group*, *November*, 1–9. https://doi.org/10.1038/srep37584

Zuris, J. A., Thompson, D. B., Shu, Y., Guilinger, J. P., Bessen, J. L., Hu, J. H., Maeder, M. L., Joung, J. K., Chen, Z. Y., & Liu, D. R. (2015). Cationic lipid-mediated delivery of proteins enables efficient protein-based genome editing in vitro and in vivo. *Nature Biotechnology*, *33*(1), 73–80. https://doi.org/10.1038/nbt.3081

Data Availability

All genomics data produced in this study have been deposited in GEO under accession number GSE223772

Code Availability

The BreakInspectoR pipeline and relevant bioinformatics pipelines used in this study can be found at https://github.com/roukoslab/breaktag and at https://github.com/roukoslab/breakinspectoR.

List of abbreviations

Abbreviation	Definition
1000G	1000 genomes project
ABE	Adenine base editor
BE	Base editor
BED	Brower extensible data
BER	Base excision repair
BLISS	Breaks Labeling In Situ and Sequencing
BLM	Bloom helicase
Cas	CRISPR associated protein
CBE	Cytosine base editor
CHANGE-seq	circularization for high-throughput analysis of nuclease genome-wide effects by sequencing
CIRCLE-seq	Circularization for In vitro Reporting of CLeavage Effects by sequencing
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
crRNA	CRISPR RNA
CTD	C-terminal domain
CtIP	C-terminal binding protein interacting protein
DDR	DNA damage response
DISCOVER-seq	Discovery of In Situ Cas Off-targets and VERification by Sequencing
DNA-PKcs	DNA-dependent protein kinase catalytic subunit
DSB	Double-strand break
EXO1	Exonuclease 1

FDR	False discovery rate
GIAB	Genome in a bottle
gRNA	Guide RNA
HDR	Homology-directed repair
HR	Homologous recombination
IVT	*In vitro* transcription
LigIV	Ligase IV
ML	Machine learning
MM	Mismatches
MMEJ	Microhomology-Mediated End Joining
MRN	Mre11-RAD50-NBS1
nCas9	Nickase Cas9
NGS	Next-generation sequencing
NHEJ	Non-homologous end joining
NT	Non-targeting
NTS	Non-target strand
NUC	Nuclease
PAM	Protospacer adjacent motif
pcgRNA	Photocleavale gRNA
PCR	Polymerase chain reaction
PE	Prime editor
pegRNA	Prime editor guide RNA
REC	Recognition
RNP	Ribonucleoprotein
RT	Reverse Transcription
sgRNA	Single guide RNA
SITE-seq	selective enrichment and identification of tagged genomic DNA ends by sequencing
SNP	Single-nucleotide polymorphism
SpCas9	*Straptoccocus pyogenes*
SSA	Single-strand annealing
ssDNA	single-stranded DNA
TALEN	Transcription activator-like effector nuclease
TS	Target strand
TTISS-seq	Tagmentation-based tag integration site sequencing
UDG	Uracil DNA glycosylase
UGI	Uracil Glycosilase inhibitor
UMAP	Uniform manifold approximation and projection
WGS	Whole genome sequencing
WRN	Werner helicase
WT	Wild type
XLF	XRCC4-like factor
XRCC4	X-ray cross complementing Group 4
ZFN	Zinc finger nuclease

List of figures

List of tables

Curriculum Vitae

Gabriel Mello da Cunha Longo

Date of Birth: 13.07.1996, Brazil
Address: Anni-Eisler-Lehmann-Straße 8 - 55122 Mainz, Germany.
Phone number: +49 1520 8230855
E-mail: cunhalongo@live.com | longo@imb-mainz.de

EDUCATION

2019-2024: Institute of Molecular Biology (IMB) - Johannes Gutenberg-Universität. Mainz, Germany. International PhD Programme in Molecular Biology

2014-2019: Federal University of Rio de Janeiro. Rio de Janeiro, Brazil. Bachelors of Science in Biology with emphasis in Genetics

RESEARCH EXPERIENCE

12.2019 - Present: PhD Student at Institute of Molecular Biology (IMB) – Johannes Gutenberg Universitaet Mainz, Germany.

Supervisor: Vassilis Roukos, PhD.

09.2019 – 12.2019: Research Assistant at Institute of Molecular Biology (IMB) – Johannes Gutenberg Universitaet Mainz, Germany.

Vassilis Roukos' research group.

12.2018 – 09.2019: Undergraduate Intern at Cellular and Molecular Hemato-oncology Laboratory - Bazilian National Institute of Cancer (INCA). Rio de Janeiro – Brazil.

Supervisor: Raquel Maia, MD, PhD.

12.2017 - 12-2018: Trainee at European Molecular Biology Laboratory - Genome Biology Unit. Korbel Group. Heidelberg, Germany.

Supervisor: Dr. Jan O. Korbel and Dr. Ashley Sanders.

07.2015 - 12-2017: Undergraduate Intern at Cellular and Molecular Hemato-oncology Laboratory - Bazilian National Institute of Cancer (INCA). Rio de Janeiro – Brazil.

Supervisor: Raquel Maia, MD, PhD.

08.2014 - 07-2015: Undergraduate Intern at Molecular Virology Laboratory - Institute of Microbiology Paulo de Goes, Federal University of Rio de Janeiro. Rio de Janeiro – Brazil.

Supervisor: Juliana Cortines, PhD.

PUBLICATIONS

Linking CRISPR-Cas9 double-strand break profiles to gene editing precision with BreakTag. **Longo GMC***, Sayols S*, Kotini AG, Heinen S, Möckel MM, Beli P, Roukos V **Nat Biotechnol**. 2024 May 13. doi: 10.1038/s41587-024-02238-8. Epub ahead of print. PMID: 38740992.

Territories or spaghetti? Chromosome organization exposed. **Longo GMC**, Roukos V. **Nat Rev Mol Cell Biol. 2021**;22(8):508. doi:10.1038/s41580-021-00372-8

Single-cell analysis of structural variations and complex rearrangements with tri-channel processing. Sanders AD, Meiers S, Ghareghani M, Porubsky D, Jeong H, van Vliet MACC, Rausch T, Richter-Pechańska P, Kunz JB, Jenni S, Bolognini D, **Longo GMC**, Raeder B, Kinanen V, Zimmermann J, Benes V, Schrappe M, Mardin BR, Kulozik AE, Bornhauser B, Bourquin JP, Marschall T, Korbel JO. **Nat Biotechnol**. 2020 Mar;38(3):343-354. doi: 10.1038/s41587-019-0366-x. Epub 2019 Dec 23. PMID: 31873213.

High clonal diversity and spatial genetic admixture in early prostate cancer and surrounding normal tissue. Zhang N, Harbers L, Simonetti M, Diekmann C, Verron Q, Berrino E, Bellomo SE, **Longo GMC**, Ratz M, Schultz N, Tarish F, Su P, Han B, Wang W, Onorato S, Grassini D, Ballarino R, Giordano S, Yang

Q, Sapino A, Frisén J, Alkass K, Druid H, Roukos V, Helleday T, Marchiò C, Bienko M, Crosetto N. **Nat Commun**. 2024 Apr 24;15(1):3475. doi: 10.1038/s41467-024-47664-z. PMID: 38658552; PMCID: PMC11043350.

R-loop proximity proteomics identifies a role of DDX41 in transcription-associated genomic instability. Mosler T, Conte F, **Longo GMC**, Mikicic I, Kreim N, Möckel MM, Petrosino G, Flach J, Barau J, Luke B, Roukos V, Beli P. **Nat Commun**. 2021 Dec 16;12(1):7314. doi: 10.1038/s41467-021-27530-y. PMID: 34916496; PMCID: PMC8677849.

RECON syndrome is a genome instability disorder caused by mutations in the DNA helicase RECQL1. Abu-Libdeh B*, Jhujh SS*, Dhar S, Sommers JA*, Datta A, **Longo GMC**, Grange LJ, Reynolds JJ, Cooke SL, McNee GS, Hollingworth R, Woodward BL, Ganesh AN, Smerdon SJ, Nicolae CM, Durlacher-Betzer K, Molho-Pessach V, Abu-Libdeh A, Meiner V, Moldovan GL, Roukos V, Harel T, Brosh RM Jr, Stewart GS. **J Clin Invest**. 2022 Jan 13:e147301. doi: 10.1172/JCI147301. Epub ahead of print. PMID: 35025765.

The LQB-223 Compound Modulates Antiapoptotic Proteins and Impairs Breast Cancer Cell Growth and Migration. Lemos, L**., Longo, G**., Mendonça, B., Robaina, M. C., Brum, M., Cirilo, C. A., Gimba, E., Costa, P., Buarque, C. D., Nestal de Moraes, G., & Maia, R. C. **International journal of molecular sciences**, 20(20), 5063. https://doi.org/10.3390/ijms20205063

LQB-118 compound inhibits migration and induces cell death in glioblastoma cells. Bernardo, P. S., Guimarães, G., De Faria, F., **Longo, G**., Lopes, G., Netto, C. D., Costa, P., & Maia, R. C**. Oncology reports**, 43(1), 346–357. https://doi.org/10.3892/or.2019.7402

Inhibition of O-GlcNAcylation Reduces Cell Viability and Autophagy and Increases Sensitivity to Chemotherapeutic Temozolomide in Glioblastoma. Amanda V Leonel, Frederico Alisson-Silva, Ronan C M Santos, Rodrigo P Silva-Aguiar, Julia C Gomes, **Gabriel M C Longo**, Bruna M Faria, Mariana S Siqueira, Miria G Pereira, Andreia Vasconcelos-Dos-Santos, Luciana B Chiarini, Chad Slawson, Celso Caruso-Neves, Luciana Romão, Leonardo H Travassos, Katia Carneiro, Adriane R Todeschini, Wagner B Dias. **Cancers (Basel).** 2023;15(19):4740. Published 2023 Sep 27. doi:10.3390/cancers15194740

PATENT

Method of parallel, rapid and sensitive detection of DNA double strand breaks: EP22189451 · Filed Aug 10, 2022 - WO2024033378A1

HONORS

2024: Poster prize - CRISPR and Beyond: Perturbations at Scale to Understand Genomes - Wellcome Sanger Institute.

Poster Title: Determinants of CRISPR/Cas9 Scission Profile for Precise, Predictable and Personalized Genome Editing.

2022: Science SLAM Winner - 1st prize at the 10th International PhD Programme symposium.

Title of the talk: CRISPR is a cut above the others!

2021: SFB1361 collaboration grant – Research grant for collaborating projects within the SFB1361 research initiative.

Received a research grant in collaboration with the Markus Loebrich Lab (TU Darmstadt) for the investigation of DNA repair dynamics using a simple qPCR approach.

2016: Best Oral Presentation – XVIII Jornada de Iniciação Científica e VIII Jornada de Pós-Graduação do INCA:

Best talk at the annual research symposium at the Brazilian National Institute of Cancer. Talk title: *O composto LQB-118 possui efeito antitumoral em linhagens de Câncer Colorretal cultivadas em monocamada e esferoides (Cultura 3D).*

PARTICIPATION IN CONFERENCES

2024: CRISPR and Beyond: Perturbations at Scale to Understand Genomes - Wellcome Sanger Institute, Cambridge, United Kingdom. Poster presentation: Determinants of CRISPR/Cas9 Scission Profile for Precise, Predictable and Personalized Genome

2023: Genome Engineering: CRISPR Frontiers - Cold Spring Harbor Laboratories, New York, United States of America.Poster presentation: Determinants of CRISPR/Cas9 Scission Profile for Precise, Predictable and Personalized Genome Editing

2023: Keystone meeting on Precision Genome Engineering joint with Genomic Instability and DNA Repair - Whistler, BC, Canada. Poster presentation: Determinants of CRISPR/Cas9 Scission Profile for Precise, Predictable and Personalized Genome Editing.

2022: Genome Engineering: CRISPR Frontiers - Cold Spring Harbor Laboratories, New York, United States of America. Poster presentation: Determinants of CRISPR/Cas9 Scission Profile for Precise, Predictable and Personalized Genome Editing

STUDENT SUPERVISION

01.2021 – 04.2021: Supervision of a BSc Student in Molecular Biology from the University of Patras, Rio, Greece at Institute of Molecular Biology (IMB Mainz), Germany.

TEACHING EXPERIENCE

January 2024: Research Methodologies lecture series of the Master's program in Molecular Biology at the University of Patras, Rio, Greece. Lectures taught: "*Introduction to Next-generation sequencing"* and "*Gene editing methods".*

August 2023: Research Methodologies lecture series of the Master's program in Molecular Biology at the University of Patras, Rio, Greece. Lectures taught: "*Introduction to Next-generation sequencing"* and "*Gene editing methods".*

January 2023: Research Methodologies lecture series of the Master's program in Molecular Biology at the University of Patras, Rio, Greece. Lectures taught: "*Introduction to Next-generation sequencing"* and "*Gene editing methods".*