Multiplex molecular recording of biological signals and events

Wei Chen

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:
Jay Shendure, Chair
Stanley Fields
Hao-yuan Kueh

Program Authorized to Offer Degree:
Molecular Engineering and Science

University of Washington

Abstract

Multiplex molecular recording of biological signals and events

Wei Chen

Chair of the Supervisory Committee:
Professor Jay Shendure
Genome science

Cells are essentially living computers, which they receive, integrate and respond to various signals. Although sharing an origin from a single zygote and a genome encoding the same genes, they differentiate into a myriad of cell types/states, with different functions. These differences may derive from different signals or heterogeneous responses to the same signal. Capturing the history of cells can help us understand their current state and predict or even shape their future behavior. Current methods of profiling cell states usually require destruction and only provide a snapshot of cell states, losing the information about their history. Can we have a device in the cells that runs autonomously to record cellular signals and events? Recently engineered genome engineering tools-CRISPR/Cas9 and its derivatives- provide a unique platform for such a purpose, allowing researchers to effectively alter/rewrite DNA sequence in a highly specific manner.

In my thesis, I focused on developing a DNA-based memory system where molecular signals and events could be recorded by alternating a pre-programmed stretch of DNA sequence, which we termed DNA TAPE. I describe experiments and methods to mainly address two questions: 1. How information can be encoded with the CRISPR system. 2. What information we can record.

We investigated the possibility of encoding information with the CRISPR system with three different approaches. We first used **CRISPR/Cas9 (cut)** to introduce pseudo-random

mutations and built a machine learning model **(Lindel)** to accurately predict the mutations and their frequencies. For each target in a DNA TAPE, this approach allows us to encode ∼3 bits of information (8 different states). We next used prime editing to introduce large deletions **(Prime-del)** or short insertions **(ENGRAM)**, which allows us to encode 1 bit and up to 20 bits of information per target in DNA TAPE.

We next sought to record various signals in cells. As most of the cell functions are executed by transcription, we recorded transcription activities of hundreds of enhancers and 3 different signals TetOn, NFκB and Wnt. We show that many signals can be recorded simultaneously with reasonable efficiency in a digital manner.

We also discussed future applications of recorders in understanding different biological functions.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

It has been a long journey since I started my scientific career, I would not make it this far without all the guidance and support from friends, mentors, and family.

I would like to first thank Xing Zhu and Keqiong Ye, who are my first scientific mentor in my career and introduced me to CRISPR in the summer of 2011. I would like to thank Dong Xing, Qian Su, and Yujie Sun. I worked closely with them for two years and learned many skills from them, including single-molecule techniques and building a super-resolution microscope. I would like to thank Jan Lammerding who is my master's advisor. I would like to thank Alex Rosenberg and Georg Seelig who introduced me to the field of single-cell genomics during my first rotation with them.

I would like to thank Ron Hause who is my rotation mentor in the Shendure lab. Ron is a patient mentor and an inspirational leader, who trained me to learn my computational skills. I would like to thank Aaron McKenna, who is my second mentor in the Shendure lab. Aaron nicely worked with me on my first project and walked me through a lot of analysis. I would like specially thank Junhong Choi, a mentor, a collaborator, and a lifelong friend, who I worked closely with on many projects and brainstormed many ideas in the lab. I also would like to thank all other Shendure lab members, especially Greg Findlay, Beth Martin, Nobu Hamazaki, Jean-Benit Lalanne, Sam Regalado, Silvia Domcke, Sanjay Srivatsan, Anh Leith, Charlie Lee, Vikram Agarwal, Jacob Tome, Lea Starita, and Ruolan Qiu.

I would like to thank my friends/colleagues Junyue Cao, Yi Yin, Xiaoyi Li, Xingfan Huang, Chengxiang Qiu, and Wei Yang. I would like to thank my friend Hongjie Chen, Mengying Zhang, Yifan Cheng, and Yiyuan Wang. I would like to thank my bros in my scientific journey, Zeyu Chen, Zeda Zhang, Shijie Zhao, and Zhilun Zhao, who I have known

for more than 10 years and who support each other on our scientific adventures.

I would like to thank my committee members, Stan Fields, Jesse Zalatan, and Haoyuan Kueh, for tolerating me being an outlier student, giving me valuable career and life advice, and supporting my decisions.

I would like to thank my advisor Jay Shendure. Jay is such a role model for my science career. I am really thankful that he took me as a student and allowed me to explore my own interests. No words would ever be enough to describe how amazing Jay is as a scientist and a mentor. I would say that Shendure lab is the best of the best place to work at any level of science. I would never forget his mentorship especially our overnight writing and submission marathon for the ENGRAM and DTT paper.

In the end, I would like to thank my parents for providing tremendous support during my growth, education, and training as a scientist. And for sure, their unconditional love.

DEDICATION

This dissertation is dedicated to my mother Xiaojiao, my father Xiaowu and my grandparents Jingdi and Bangqi, who provided unconditional support and love through out my life.

Chapter 1

INTRODUCTION

D'où venons-nous? Que sommes-nous? Où allons-nous?

Where Do We Come From? What Are We? Where Are We Going?

History is one of the most important aspects of human culture and society. Many methods during the short period of humanity have been developed to record and track history, including languages, paintings, writings, photography, and videos. A collective of historical events defines what we are today. Learning from history can not only help us understand where we come from and what we are but also guide us to make decisions for future directions. Similar principles can be applied to cells. For multicellular organisms, all cells are derived from the same zygote and share the same genome that encodes the same genes. Cells are constantly interacting with neighboring cells and the environment, receiving, integrating signals, and responding with molecular programs. A series of cellular events and signals activate different molecular programs to differentiate cells into various types and states. Recording cellular events and signals can help us understand what determines its current state and predict or even shape future behaviors. Yet, when the work of this dissertation began, there was no such recording device. To record biological events, we need a minimum system with two components: 1. a memory that is stable and easily accessible and 2. a writer that can efficiently record biological events to memory. In this introduction chapter, I will provide a brief review of the development of these two components.

1.1 DNA as memory device

Throughout evolution, cells developed many approaches that work as memories. For short-term storage, transcriptional circuits[1, 2, 3] and epigenetic modification[4, 5], can

have a memory life ranging from days to months. For long-term storage, cells use genomic DNA to store their genetic information, which can be stably preserved for decades to tens of thousands of years. We chose DNA as our recording media, as it is digital, present in every living cell, and can be easily accessed using DNA sequencing technology. To differentiate it from genomic DNA encoding genetic information, we term this DNA memory as DNA TAPE, which is programmable and integrated into the genome.

1.2 DNA writers

Various DNA altering proteins, including recombinase, integrase and nucleases, can be used as DNA writers. Here we briefly review their recording capability in terms of capacity and multiplexibility. Various DNA-altering proteins, including recombinase, integrase, and nucleases, can be used as DNA writers. An ideal DNA writer should be able to write with high precision, multiplexbility (how many events can be written simultaneously), and directionality (can we record orders of events). Here we briefly review the features of a few DNA writers.

1.2.1 Recombinase and Integrase

Both recombinase and integrase are able to alter DNA sequence in a site-specific manner. Recombinase usually refers to tyrosine recombinase such as Cre. Integrase usually refers to serine recombinase(integrase) such as Bxb1. They have distinct mechanistic and evolutionary features[6]. Tyrosine recombinase creates single-strand DNA nick with $3'$-phosphotyrosine bonds and two single-strand DNA can be rejoined through a holliday-junction structure. Due to the nature of tyrosine recombinase, excision is favored during recombination, generating only two states of DNA sequence (intact and excision, 1 bit). It has been mostly used to express reporter proteins[7, 8, 9]. Serine integrase recognizes a short attachment sequence (attB/attP) and generates a double-strand DNA break with a 2bp overhang. The break can be rejoined based on overhang compatibility. Thus, three states can be represented by one site (trits, intact, excision, and flip). 11 orthogonal integrase system has been identified by

mining the metagenomic database[10]. This allows the creation of a memory system with 1.375 bytes. Though impressive, this strategy is still limited by the ability of integrating individual integrase into the system of interest. More recently, instead of using orthogonal integrase, Chow *et.al*[11] developed a recording system (intMEMOIR) based on Bxb1 with orthogonal 2bp overhang, where they change the dinucleotide sequence in the attachment site to create a memory system with ∼15 bits (3^{10}. Serine integrase based recorders have demonstrated high specificity and reasonable capacity. However, individual events require individual enzymes to record specific information, limiting its multiplexibility. In addition, its possible but challenging to record the order of multiple events with integrase system[12]

1.2.2 CRISPR based system

CRISPR system is a versatile genome engineering tool and it has been used as a recording tool for lineage tracing[13, 14, 15], molecular events[16, 17, 18], or even short movies[19, 20].

Cas1-Cas2 is CRISPR spacer acquisition system that specifically inserts a 33bp DNA sequence (protospacer) to the target CRISPR array in an ordered manner. It has been adopted to record the exposures of chemicals and copper[16].

CRISPR/Cas9 system relies on guide RNA to target specific DNA sequences to record information. It has gone through a few generations, from traditional Cas9 cut with a double-strand break, to base editing, and to most recently prime editing. All three systems have been used for recording.

In 2016, McKenna *et.al* was the first to use CRISPR for lineage recording, where they designed a DNA TAPE with 10 sites that can be targeted by Cas9 (GESTALT). The double strand breaks introduced by Cas9 are mostly repaired by the Non-homologous end joining (NHEJ) pathway, resulting in imperfect repair with short insertion or deletions (indels). It was thought that indel generation is random, allowing the creation of large memory devices. However, more works[21, 22, 23, 24], including my own work described in **Chapter 2** suggested that this process is not random and is governed by the local sequence, limiting its capacity. Its possible but still very challenging, with careful target design (introducing mi-

crohomology flanking the cut site), to record the order of events using a Cas9-based system. In addition, the bigger issue of GESTALT is the generation of double-strand breaks, which could be toxic to sensitive cells and introduce information loss during DNA repair.

Base editing overcomes these challenges by replacing Cas9 with dCas9 fused with cytidine deaminase, which specifically converts CG pair to AT pair within a window without generating double-strand breaks. In two proof of concept studies, CAMERA and DOMINO[17, 18] demonstrated their capability of recording signal and its intensity. CAMERA and DOMINO can also record the order of two events. Though improving, the base editor based recording system is still limited by its capacity (1-2 bits/target) and multiplexity(not able to record multiple events simultaneously).

More recently, Liu group developed prime editing in which a nickase Cas9 (nCas9) is fused with reverse transcriptase (RT). In coupling with prime editing guide RNA (pegRNA), the prime editor can install programmed mutations (substitution, short insertion, and deletions) with high precision. Ideally, this would overcome all challenges mentioned above: 1. the edits are highly precise without generating double-strand breaks, 2. The information encoded is nearly unlimited. With the short insertion of specific sequences, in theory, prime editing is able to record thousands or even millions of events (10bp insertion, $4^10 = 20$bits). 3. The information encoded is programmed so that its possible to record the order of multiple events. In addition to the advantage mentioned here, I also demonstrated that prime editing can record the signal intensity with high sensitivity in **Chapter 4**.

1.3 Topics in this dissertation

The following chapters of this dissertation consist of 3 projects I worked on during my Ph.D. with one central topic-multiplex recording of biological events and signals. In **Chapter 2**, I describe a high throughput assay to profile repair outcomes of double-strand breaks induced by CRISPR/Cas9. I also trained a model **Lindel** that accurately predicts the mutation and its frequency. This allows us to predict the information content that can be stored in DNA TAPE. In **Chapter 3**, with Junhong Choi, we developed a highly precise genome

deletion tool using paired pegRNA prime editing. This tool can be used for recording or dissecting functions of DNA sequence in the genome. In **Chapter 4**, I lead the development of ENGRAM, a multiplex recording system that can efficiently capture and record transcription activity using prime editing. Finally, in **Chapter 5**, I close the dissertation with a prospective look towards future directions and applications of ENGRAM-based recorders. All chapters are modified from published (**Chapter 2,3**) or under revision work (**Chapter 4**).

Chapter 2

LINDEL: PREDICTING REPAIR OUTCOMES OF CAS9-MEDIATED DOUBLE-STRAND BREAK

This Chapter is adopted from published work with minimum changes

Chen, W., McKenna, A., Schreiber, J., Haeussler, M., Yin, Y., Agarwal, V., Noble, W.S. and Shendure, J., 2019. Massively parallel profiling and predictive modeling of the outcomes of CRISPR/Cas9-mediated double-strand break repair. *Nucleic acids research*.

Author contribution: Wei Chen designed and performed the experiments with the help from Aaron McKenna. Wei Chen analyzed the data with input from Aaron Mckenna, Vikram Agarwal, and Yi Yin. Wei Chen and Jacob Schreiber created the model Lindel with the input from William Nobel. Wei Chen and Jay Shendure wrote the manuscript.

A back story about Lindel:

We started to work on this project in the Fall/Winter of 2016. The beautiful work of GESTALT was published in the summer of 2016, and many opportunities were opened up. For example, from GESTALT data, we observed biased indel profiles from Cas9 induced double-strand break, which is similar to the observation from another paper that came out in the summer of 2016. We wonder if we can predict the indel profile for a random target sequence. I and Aaron quickly came up with the strategy to profile thousands of sequences in one experiment, in which gRNA and target sequence pairs were synthesized in the same DNA oligo. Aaron designed the target sequences and I performed the experiments. After collecting all the data, we reached out to Bill Nobel for modeling inputs. Jacob Shreiber from the Nobel lab kindly provided his expertise in model training. With the pressure of being scooped 2-3 times, we tried really hard to improve our model to outperform the other models (FORECasT and inDelphi).

2.1 Abstract

Non-homologous end-joining (NHEJ) plays an important role in double-strand break (DSB) repair of DNA. Recent studies have shown that the error patterns of NHEJ are strongly biased by sequence context, but these studies were based on relatively few templates. To investigate this more thoroughly, we systematically profiled ∼1.16 million independent mutational events resulting from CRISPR/Cas9-mediated cleavage and NHEJ-mediated DSB repair of 6,872 synthetic target sequences, introduced into a human cell line via lentiviral infection. We find that: 1) insertions are dominated by 1 bp events templated by sequence immediately upstream of the cleavage site, 2) deletions are predominantly associated with microhomology, and 3) targets exhibit variable but reproducible diversity with respect to the number and relative frequency of the mutational outcomes to which they give rise. From these data, we trained a model (Lindel) that uses local sequence context to predict the distribution of mutational outcomes. Exploiting the bias of NHEJ outcomes towards microhomology mediated events, we demonstrate the programming of deletion patterns by introducing microhomology to specific locations in the vicinity of the DSB site. We anticipate that our results will inform investigations of DSB repair mechanisms as well as the design of CRISPR/Cas9 experiments for diverse applications including genome-wide screens, gene therapy, lineage tracing and molecular recording.

2.2 Introduction

Genome engineering conventionally involves using a programmable endonuclease (*i.e.* a zinc finger nuclease (ZFN), transcription activator-like effector nuclease (TALEN) or RNA guided nuclease Cas9 (clustered regularly interspaced short palindromic repeats-CRISPR associated protein CRISPR/Cas9) to introduce a double-strand break (DSB) at a specific location in the genome. In mammalian cells, such DSBs are primarily repaired by one of two pathways – homology directed repair (HDR) and classical non-homologous end joining (c-NHEJ) [25, 26]. HDR uses homologous template sequences to repair the DSB, potentially

introducing programmed edits via the repair template. In contrast, c-NHEJ directly rejoins the broken ends, often perfectly but occasionally introducing errors, typically in the form of short insertions or deletions (indels) [27]. In addition to HDR and cNHEJ, there is evidence for an alternative NHEJ pathway (alt-NHEJ), also termed microhomology mediated end joining (MMEJ), wherein short, homologous sequences in the vicinity of the DSB are used to align the broken ends prior to joining, resulting in deletions or potentially more complex events [28]. Below, we use NHEJ to refer to both c-NHEJ and MMEJ/alt-NHEJ, *i.e.* template-free editing.

In recent years, CRISPR/Cas9 has emerged as a particularly versatile tool for genome editing. For many if not most applications of CRISPR/Cas9-mediated genome engineering, it is used in conjunction with the cells endogenous NHEJ machinery to introduce short indels in a targeted fashion [29, 30, 31], *e.g.* to disrupt the function of genes or regulatory elements [32, 33, 34] or to introduce irreversible changes that record cell lineage or molecular events [13, 14, 15]. However, despite NHEJs central role in this transformative tool, our understanding of the processes that determine the rate and patterns of NHEJ-mediated errors remains incomplete.

Recent studies have demonstrated that the error outcomes of NHEJ are strongly dependent on sequence context [21, 35]. Other studies show that the characteristics of the broken ends (blunt or staggered end; length of any overhang) also affect end-joining patterns both in vitro [36] and in vivo [36, 37]. However, a systematic profiling of the sequence determinants of NHEJ repair patterns has yet to be undertaken.

Here we profiled $\sim$1.16 million mutational events resulting from Streptococcus pyogenes Cas9 (SpCas9)-mediated cleavage and NHEJ-mediated DSB repair of 6,872 synthetic target sequences. From the resulting data, we identify the primary features of sequences adjacent to the sites of DSBs that shape the distribution and relative frequency of NHEJ-mediated mutational outcomes, *e.g.* nucleotide content and microhomology. We furthermore exploit microhomology to demonstrate the programming of deletion patterns. Finally, we develop a logistic regression model to predict insertions and deletions (**Lindel**) that result from

CRISPR/Cas9-mediated cleavage of an arbitrary sequence. A standalone Lindel webtool is freely available (https://shendurelab.github.io/Lindel/), and Lindel predictions have been integrated into the CRISPOR web tool (http://www.crispor.org) [38].

2.3 Results

2.3.1 Development of a massively parallel strategy to profile NHEJ-mediated genome edits

Toward a comprehensive understanding of the sequence determinants of NHEJ-mediated error patterns, we developed a strategy that would allow us to efficiently profile a large number of repair events from each of a large number of sequence contexts (**Figure 2.1**). In brief, we designed 70,000 targets balanced in nucleotide content and screened against the human genome for CRISPR/Cas9 single guide RNAs (sgRNAs). We then used array-based oligonucleotide synthesis to encode these targets in cis with their corresponding sgRNAs, separated by a 20 bp spacer. We then amplified and cloned these molecules to a lentiviral vector. In our initial experiments, the complexity of the resulting library of synthetic targets and their cognate sgRNAs was such that we obtained relatively few edited templates per target. Therefore, we re-cloned the library under bottlenecking conditions (Materials and Methods), reducing its complexity to 12,917 targets. We then proceeded with viral packaging and transduction, in triplicate, of a monoclonal human embryonic kidney (HEK) 293T cell line that stably expresses spCas9 (multiplicity of infection of 4-8). As such, within any given cell, only one or a few sgRNAs are expressed, and each one directs Cas9-mediated DSBs to a target located immediately adjacent to it. After five days to allow for the introduction of NHEJ-mediated errors at these targets, cells were harvested and genomic DNA isolated. We then PCR amplified the region comprising the targets and corresponding sgRNAs using unique molecular identifiers (UMIs) appended during the first extension cycle to distinguish whether identical edits were derived from the same cell or different cells.

Summing across the three replicates, we sequenced PCR amplicons to a depth of 148 million reads, which were reduced to 1.19 million reads after collapsing on the basis of

identical sequences and UMIs, and filtering of reads with evidence of lentivirus-mediated template switching [39, 40] or other unexpected sequences (*e.g.* synthesis or PCR errors). After further filtering of poorly represented targets (those represented by fewer than 10 UMIs), our dataset consisted of 1.16 million UMIs corresponding to 6,872 unique targets. On average, each target was represented by 168 UMIs and 24 alleles (where allele refers to a unique post-editing sequence of a given target). Each allele was aligned to its original sequence, known because the corresponding gRNA sequence is part of the same amplicon, using the Needleman-Wunsch algorithm [41]. Alleles were categorized as wild-type (*i.e.* unedited), a deletion, or an insertion.

Overall, targets were highly edited, with only 9.8% of UMIs corresponding to the wild-type allele. Of UMIs containing detectable mutations, 63.6% were deletions and 31.5% were insertions (**Figure 2.2a**). The remainder (4.9%) contained some combination of substitutions, insertions and deletions, and are excluded from all of our subsequent analyses. Deletions were dominated by small events; only 1.5% were 25 bp, although we note that deletions >150 bp are not captured by our assay [33, 42]. In contrast, although we believe that our assay should have been able to recover insertions up to 500 bp, the overwhelming majority of insertion events were of a single base pair.

2.3.2 Repair patterns are reproducible but exhibit highly variable entropy between targets

We sought to examine whether repair patterns for any given target were reproducible, as previously shown for a more limited set of templates [21]. For each target, we calculated the frequency of each non-wild-type allele. For any given target, the distribution of frequencies for its alleles were highly reproducible in pairwise comparisons of the three replicates (median Pearson's $r = 0.91, 0.93, 0.93$, **Figure 2.2b, left**). Meanwhile, if we permute the alleles in one replicate on a target-by-target basis and repeat the pairwise comparison, these correlations are greatly reduced (median Pearson's r = 0.20, **Figure 2.2b, right**).

Confirming the observations of [21], the diversity of mutations strongly varied from target to target. We calculated the Shannon entropy of mutational outcomes for any given target

as $-\sum p_i \times \log(p_i)$, where p_i is the frequency of ith indel of that target (**Figure 2.2c**). Entropy values for any given target were highly reproducible between replicates (**Figure 2.2d**) and only modestly correlated with sampling depth **(Figure 2.7)**. Of note, some targets consistently exhibited particularly diverse mutational outcomes consequent to NHEJ – that is, high entropy (*e.g.* **Figure 2.2e**, where the most frequently observed mutation occurs in only 10.1% of mutated templates). Other targets were strongly biased towards a more limited set of mutational outcomes – that is, low entropy (*e.g.* **Figure 2.2f**, where the most frequently observed mutation occurs in 80.4% of mutated templates).

2.3.3 Sequence context at the DSB site predicts the frequency of insertions

We next sought to investigate the determinants of insertions at the DSB, which were dominated by 1 bp events (**Figure 2.3a**). 84% of 1 bp insertions were predicted (and presumably templated) by the nucleotide immediately upstream of the cleavage site (*i.e.* the 17^{th} nucleotide in target sequence; **Figure 2.3b Figure 2.8**). Although it might have been expected that NHEJ-mediated repair would be symmetric with respect to the site of a DSB, we do not observe templating from the immediately downstream (18th) nucleotide (**Figure 2.3b**). Similarly, of 2 bp insertions, a substantially greater than expected proportion (41%) were templated by the sequence immediately upstream of the DSB (*i.e.* inserted sequence identical to the 16^{th} and 17^{th} nucleotides of the target sequence; **Figure 2.3c**). The asymmetric templating of NHEJ-mediated insertions was also described in two other recent studies based on data from yeast and mice [43, 15].

Because the ratio of insertions to deletions varied from target to target, we used kpLogo [44] to examine what local sequence features might shape this. We find that the presence of a T or A at the 17^{th} bp of the target was associated with insertion events, while a G or C at this position was associated with deletion events (**Figure 2.3d, left**). Additional analyses showed a TG dinucleotide flanking the cleavage site to be the most highly biased toward insertion (57% of events with that context are insertions), while a GA dinucleotide flanking the cleavage site was the most highly biased towards deletion (17% of events with

that context are insertions) (**Figure 2.3d, right**).

We split 2,680 targets associated with both insertion and deletion outcomes into training (n = 2,000) and test (n = 680) sets, and trained a linear regression model to predict the proportion of insertion events based on position-specific content of the hexamer centered on the DSB (single and dinucleotide k-mers; 104 binary features; **Figure 2.3e**). The model performs reasonably well (Pearson's r = 0.70).

Overall, these analyses confirm that local sequence around the DSB site plays an important role in shaping the outcome(s) of NHEJ-mediated errors. In particular, the asymmetry implied by the high rate of identity between 1-2 bp insertions and the nucleotides immediately upstream to the DSB, but not the nucleotides immediately downstream to the DSB (**Figure 2.3b-c**), suggests that not all CRISPR/Cas9-mediated cleavages are blunt-ended. Indeed, in vitro studies have shown that the non-complementary strand of the target can sometimes be cleaved by Cas9 at multiple sites upstream of the -3 bp position relative to the protospacer adjacent motif (PAM), while the complementary strand is cut only at that site, instances which would result in a $5'$ overhang [45, 46]. The preponderance of 1 bp insertions templated by the 17^{th} rather than 18^{th} base could be explained by fill-in of this overhang followed by blunt-ended ligation (and similarly for the preponderance of 2 bp insertions that are templated by the 16^{th} and 17^{th} bases, rather than the 18^{th} and 19^{th} bases).

To summarize, we propose a model (**Figure 2.3f**) where: 1) some proportion of cleavages of the non-complementary strand by Cas9 occur upstream of the -3 bp PAM cleavage site, while cleavage of the complementary strand always occurs between the 17^{th} and 18^{th} positions, resulting in a $5'$ overhang; 2) $5'$ overhangs are preferably repaired by gap-filling and ligation, resulting in the observed bias towards templating by the bases immediately upstream rather than downstream of the DSB; 3) local sequence context biases the pattern of cleavage on the non-complementary strand, resulting in different frequencies of blunt vs. $5'$ overhangs for different targets, which in turn biases the ratio of insertions vs. deletions. A similar model was recently proposed by Lemos *et al.* based on asymmetric templating of NHEJ-mediated insertions observed in yeast [43].

2.3.4 Extensive use of microhomology in NHEJ-mediated

We next examined patterns of deletion. Microhomology (MH) refers to the use of short regions of identical sequence (1 to 16 bp) that can mediate the alignment of broken ends (**Figure 2.4a**) and is relevant to both c-NHEJ and alt-NHEJ/MMEJ [47, 48, 28, 49]. Here, a deletion event is considered to be MH-mediated if the sequence at the 3′ of a rejoined end is identical to the 3′ end of the deleted sequence, and the size of the MH tract refers to the length of that identical sequence. By that definition, we found that over 75% of deletion events in our dataset are MH-mediated. The length of MH tracts ranged from 1-10 bp. Nearly all MH-mediated events (94.6%) involved relatively short tracts of microhomology, *i.e.* 1-4 bp. Longer MH tracts were observed more rarely **(Figure 2.9a)**, probably simply due to the relative paucity of opportunities in our set of target sequences.

The frequencies of tracts of various lengths consistent with MH usage were substantially higher than background expectation for all lengths except 1 bp, with that frequency increasing as a function of tract length (**Figure 2.4b**). We further investigated the relevance of 1 bp MH by comparing the proportion of 1 bp deletion events in targets with identical vs. non-identical nucleotides immediately spanning the cleavage site. We observe a 3-fold greater proportion of 1 bp deletion events when those nucleotides are identical than when they are not (**Figure 2.9b**), suggesting that 1 bp MH may play a role in aligning, stabilizing and rejoining the broken ends.

The lengths of MH vs. non-MH mediated deletions exhibited distinct distributions (**Figure 2.4c, d**). In particular, the distribution of deletion sizes for MH-mediated events peaks at both 1 bp and 5-6 bp, while an equivalent distribution for non-MH-mediated deletions peaks at both 1-2 bp and 8 bp. The frequency of longer deletions exhibits an exponential decay for both MH and non-MH mediated events. To investigate this further, we jointly analyzed the frequency of start and end points for deletion events, relative to the position of the canonical cleavage site (**Figure 2.4e, f**). Both MH and non-MH mediated deletions exhibited a preference for unidirectional events, *i.e.* either the start or end point is immediately

adjacent to the cleavage site, rather than the deletion spanning the cleavage site.

What explains the excess of deletion events of specific lengths? For MH-mediated events, the excess of 1 bp deletions may simply be attributable to the aforedescribed instances of identical nucleotides spanning the cleavage site **(Figure 2.9b)**. However, the excess of 5-6 bp MH-mediated events is clearly driven by events in the downstream direction (**Figure 2.4e**), *i.e.* deletions between the DSB and the PAM. A potential explanation is that the predilection of PAM-like sequences near the DSB for deletion events (*i.e.* a G nucleotide at the 17^{th} position or a CG dinucleotide at the $16^{th}/17^{th}$ position; **Figure 2.3d**), coupled with the consistent presence of the CGG PAM sequence at the 21st-23rd position, results in an excess of deletions mediated by CG (5 bp deletion) or G (5-6 bp deletion) microhomology (**Figure 2.4g**). Further work would be required to confirm this, as there may be other explanations.

For non-MH-mediated events, the excess of 8 bp events might be explained by the observation that in the dsDNA-sgRNA-Cas9 complex, the region 1-8 bp downstream of the cleavage site is occupied by Cas9, even after cleavage [50, 46]. Thus, the enrichment of non-MH deletions 8 bp from the cleavage site could simply correspond to the nearest position lacking Cas9 protection from endonucleases during repair (**Figure 2.4h**).

2.3.5 Generating predictable mutations by programming microhomology tracts

Since MH is widely used in deletion events, we reasoned that we could program a library of targets to generate predictable mutations by introducing MH proximal to the cleavage site. With the same basic experimental scheme (**Figure 2.1**), we tested a library of 1,000 targets and corresponding guides containing MH tracts of three different lengths (2, 4 or 6 bp) matching the sequence immediately upstream of the expected DSB site, and positioned 6 bp downstream of the cleavage site (**Figure 2.5a, b**). The resulting data were processed and analyzed similarly to the previous experiment.

Intentionally programming MH tracts resulted in a high proportion of events corresponding to the expected deletions (8, 10 and 12 bp deletions for 2, 4 and 6 bp MH tracts, respec-

tively; **Figure 2.5b, c, Figure 2.10**). We also observe that the ratio of the programmed deletion increases as a function of length of the MH tract (**Figure 2.5c**). However, despite the greater predictability of which MH-mediated outcome would occur, the relative proportion of MH-mediated deletions increased only slightly from 76% to 82% (**Figure 2.5d**). Furthermore, we did not observe an excess of imperfect MH-mediated events, *e.g.* an excess of 11 bp or 13 bp deletions in targets for which a 12 bp deletion was expected (**Figure 2.10**). Nonetheless, the results show how targets that would result in diverse editing outcomes can be strongly biased towards a specific outcome by the presence of MH tracts (**Figure 2.5e,f**).

2.3.6 A machine learning model to predict editing patterns

The above results suggest that the NHEJ-mediated repair outcomes for any given target sequence are both reproducible and dependent on sequence context. Accordingly, we next sought to train a machine learning model to predict these outcomes and their relative frequencies. We began by filtering out target sequences that were either poorly reproducible (low correlation between replicates, mainly due to low UMI counts; **Figure 2.11a**) or poorly edited, resulting in a dataset of $\sim$1 million UMIs representing 4,790 target sequences. On average, each target in this subset of the data used for modeling was represented by 204 UMIs and 28 alleles.

Because larger events are rare in our data, we focused on predicting deletion events $<$ 30 bp in length, as well as all possible 1-2 bp insertion events at the DSB. Across all targets, we identified 557 event classes. The vast majority of CRISPR/NHEJ-mediated indels arising from any given target sequence should fall into one of these 557 event classes. We therefore framed our machine learning task as one of predicting, for an arbitrary target sequence, the relative frequency of CRISPR/NHEJ-mediated indels falling into each of these 557 event classes. These included 536 deletions (defined solely by their start/end points), all 4 possible single nucleotide insertions, all 16 possible dinucleotide insertions, and finally, a single event class for insertions greater than 2 bp in length. Of note, the 536 deletion event classes comprise almost all of the 550 possible combinations of start/end positions, with

the constraints that deletions must be less than 30 bp and overlap with the -3/+2 window around the cleavage site. The 14 potential deletions that satisfy these constraints but were not observed in the modeling dataset were mainly large deletions.

We also defined 3,033 binary features to characterize the target sequence for which repair outcomes are being predicted. These are 1) Sequence features: 384 binary features corresponding to one-hot encoded sequence, including 80 for single nucleotide content (4 nucleotides $\times$ 20 positions) and 304 for dinucleotide content (16 dinucleotides $\times$ 19 positions); 2) Microhomology features: 2,649 binary features corresponding to MH tracts; specifically, for each of the possible deletion event class, we defined 2 to 5 binary features (depending on the size of deletion) corresponding to the length of the MH tract ([0-4 bp $\times$ 519] + [0-3 bp $\times$ 7] + [0-2 bp$\times$6] + [0-1 bp$\times$4] deletion event classes = total of 2,649 binary features) (**Figure 2.6a**).

We split the 4,790 target sequences in our modeling dataset into subsets of 3,900 (for training), 450 (for validation) and 440 (for testing).

Our model consists of three components: (1) predicting the ratio of insertions to deletions; (2) predicting the distribution of 536 classes of deletion events; (3) predicting the distribution of 21 classes of insertion events. The overall distribution of predicted outcomes is then determined by intersecting these three components. Of note, because we define deletion classes using the deletion start site and deletion length, two or more classes can effectively represent identical outcomes due to microhomology, but in a sequence specific manner (**Figure 2.11c**). To address this, while evaluating performance, we simply collapsed identical outcomes.

We trained predictors for each of the three components independently using logistic regression with a varied number of features and varied strength of L1 or L2 penalties. All of the models were trained on the training set using cross-entropy loss and evaluated on the validation set using the mean squared error (MSE). For the indel ratio predictor, we predicted that microhomology features and sequence context would both be important for prediction. However, including microhomology features did not improve the performance compared to using one-hot encoded sequence alone (MSE = 0.0203 and 0.0201 respectively, **Figure 2.11d**).

For the insertion predictor, it has been shown above that most insertions were templated by the sequence upstream of the cleavage site. We reasoned that sequence context around the cleavage site (3 bp) should be sufficient to predict insertions. Consistent with this, including the full 20 bp target worsened performance, increasing MSE from 0.00666 to 0.00711 (**Figure 2.11e**). For the deletion predictor, we compared the performance of models using sequence features only, microhomology features only or all features. The model with all-features performed the best (MSE of 0.000204, as compared with 0.000271 for sequence-only and 0.000208 for microhomology-only) (**Figure 2.11f**). However, the nearly identical performance of the all-features vs. microhomology-only models for predicting deletions is notable. We used the best performing model for each component to build a predictor for the overall distribution of outcomes (**Figure 2.6a**).

Applying this model to the test set of 440 target sequences, which had been entirely held out from the training and validation steps, we compared the observed versus predicted frequencies of indels falling into various event classes. Observations and predictions were well matched for most targets, with a MSE of 0.000172 (**Figure 2.6b**). As a baseline, we also generated a set of predictions based simply on the aggregate frequencies of event classes in the training and validation datasets; as expected, these predictions performed more poorly (MSE of 0.000359; **Figure 2.6b**), confirming the improvement conferred by the model. Poorly predicted targets tended to be those with relatively shallower sampling of editing events, *i.e.* where our observed frequencies are noisier (**Figure 2.11b**).

2.3.7 Comparison to other models

While this manuscript was in preparation, several similar studies were published [22, 51, 52]. Together with this manuscript, all four studies profiled repair outcomes of Cas9-induced DSBs at large numbers of endogenous [52] or synthetic [22, 51] targets (**Table 2.1**). The primary conclusions, *e.g.* that sequence context around the DSB, together with MH, are the major determinants of repair outcomes, are consistent between these studies as well as with earlier studies [21, 35]. In addition, Shen et al. built inDelphi and Allen et al.

built ForeCasT, as models that predict NHEJ repair outcomes, analogous to the Lindel model described here. The ForeCasT model predicts deletions similarly to Lindel, as well as all possible 1-2 bp insertions (**Table 2.1**). In contrast, the inDelphi model predicts the frequency of three classes of indels independently, using a neural network model for MH-mediated deletions (90 classes) and non-MH deletions (59 classes corresponding to 1-59 bp deletions, without prediction of location), and k-nearest neighbors model for 1 bp insertions (4 classes) (**Table 2.1**).

We compared the three models by measuring the MSE on 440 targets in our test set as well as 4,298 targets in the ForeCasT test set (of note, the predicted probabilities of event classes not predicted by inDelphi were simply set to 0). Our model performed the best on our test set (MSE = 0.000172, 0.000225, 0.000212 for Lindel, ForeCasT and inDelphi, respectively), while ForeCasT performed the best on its test set (MSE = 0.000173, 0.000152, 0.000182, for Lindel, ForeCasT and inDelphi, respectively). We then combined our training set (3,900 sequences) with ForeCast training set (10,725 sequences), resulting in a total of 14,625 sequences. We trained on this aggregated training set using the Lindel modeling approach, which resulted in the best overall performance (MSE = 0.000165 on our test set and 0.000125 on ForeCasT test set; **Figure 2.6c, d**). This final Lindel model is more accurate at predicting the ratio of insertions to deletions than ForeCasT (MSE = 0.01 and 0.02 for final Lindel model and ForeCasT, respectively; **Figure 2.12a, b**). We further investigated the source of the errors for these models. Despite the fact that they implement different modeling approaches, both Lindel and ForeCasTs mispredictions primarily lie with small deletions and 1 bp insertions (**Figure 2.12c, f**).

As a common use of CRISPR/Cas9 in conjunction with NHEJ is to introduce frameshifting mutations, we also assessed the observed vs. predicted ratios of frameshifting indels for each of the 440 targets in our test set and 4,298 targets in ForeCasT test set, and found them to be reasonably correlated (Pearson's r = 0.707, MSE = 0.0122 on our test set and Pearson's r = 0.676, MSE = 0.0098 on ForeCasT test set; **Figure 2.6e**; **Figure 2.12g**). This result compares very favorably with the predictions of a previously published tool that we tested

on this same task (Pearson's r = 0.283, MSE = 0.0431 on our test set and Pearson's r = 0.455, MSE = 0.0315 on the ForeCasT test set; **Figure 2.6f, Figure 2.12h**) [53].

2.4 Discussion

In summary, we developed an assay to systematically profile the diversity and relative frequencies of mutational events resulting from CRISPR/Cas9-mediated cleavage and NHEJ-mediated DSB repair of thousands of synthetic sequences. In applying this assay and analyzing the editing outcomes associated with 6,872 target sequences, we confirm that CRISPR/NHEJ-mediated repair outcomes for any given target sequence are reproducible, predictable, and largely shaped by the sequence context around the cleavage site [21, 35].

Our results also provide further insights into NHEJ-mediated repair of CRISPR/Cas9-mediated DSBs in human cell lines. First, we observe that insertion events are dominated by 1-2 bp insertions templated by the sequence immediately upstream of the cleavage site. Together with in vitro data from the literature [46, 45], the data supports a model in which the sequence context around the DSB biases the extent to which cleavages are blunt-ended vs. include a 1-2 bp 5′ overhang. Such 5′ overhangs are repaired by gap-filling and ligation, resulting in asymmetrically templated 1-2 bp insertions. Second, we observe extensive usage of 1-4 bp microhomology in mediating deletion events, and furthermore show that repair outcomes can be strongly biased towards predictable outcomes by intentionally introducing MH tracts at specific distances from the DSB. Notably, however, the introduction of MH tracts did not substantially increase the proportion of MH-mediated events. Third, both MH and non-MH-mediated deletions were overwhelmingly unidirectional (*i.e.* extending either upstream or downstream from the DSB, rather than spanning it).

Our assay has two main limitations. First, because of the locations of the PCR primer sites, we are only able to recover small deletions, and may be missing the rare, large deletion events that we and others have described [33, 42]. Greater knowledge of the frequency and determinants of large events is necessary to enable their prediction. Second, the lentiviral-based assay that we used fails to capture the influence of chromatin state on editing efficiency

and repair outcomes. Lentivirus integrates to diverse locations across the genome, such that we are effectively observing an average, but integrations are biased towards open chromatin [54, 55]. As such, the patterns that we observe and model may be biased to this compartment. Furthermore, we are varying the sgRNA spacer sequence within the context of constant neighboring sequence, *i.e.* the lentiviral backbone. To the extent that this sequence biases nucleosome positioning, and that nucleosome positioning in turn influences Cas9 binding and cleavage [56, 57], the patterns that we observe and model may be additionally biased. Additional systematic profiling of repair outcomes, in different compartments (*e.g.* open vs. closed chromatin) and in different sequence contexts, will be necessary to understand the magnitude and nature of each of these potential sources of chromatin-mediated bias [58, 52]

In addition to insights into NHEJ-mediated repair of CRISPR/Cas9-mediated DSBs, our study also provides a new tool for sgRNA design for diverse goals. First, an important application of CRISPR/Cas9 is to achieve gene knockouts, a goal that depends on the efficient introduction of frameshifting indels. Dual cleavage with a variety of different nuclease systems has previously been shown to be an effective strategy for introducing frameshifting mutations [59, 60, 61, 62, 63]. Our models accuracy for predicting which sgRNAs/targets are likely to result in a high proportion of frameshifting indels will improve the viability of single cleavage with CRISPR/Cas9 for this same goal. Of note, our approach will not obviate the need for downstream validation to identify clones bearing the intended mutation, although it may reduce the number of clones that need to be screened. Second, for applications focused on mutation correction (*e.g.* using CRISPR/NHEJ to correct pathogenic mutations), the model may be useful for identifying sgRNAs/targets for which the desired outcome is predicted to occur at a high or sufficient frequency. Third, we and others have recently repurposed CRISPR/Cas9 as a tool for lineage tracing and/or molecular recording [13, 14, 15]. For some goals (*e.g.* lineage tracing), the identification of high entropy targets may critically enable the diversity necessary to uniquely label millions or billions of cells. For other goals (*e.g.* molecular recording), the design of low entropy targets may facilitate predictable sequential editing. More generally, a deeper understanding of CRISPR/NHEJ-mediated mutations will

strengthen our ability to precisely orchestrate not only the locations but also the outcomes of genome editing.

2.5 Materials and Methods

2.5.1 sgRNA and target pair library design

To generate a library of CRISPR/Cas9 targets that could safely be characterized within human cells, we evaluated 1 million random 20mer crRNA sequences, scoring them against the human genome (version hg19) for off-target effects using FlashFry [64]. We excluded guides with an exact match or up to two mismatches against any potential target in the human genome, or those with an off-target score less than 90 [65], resulting in a modest bias towards targets containing CpG dinucleotides (**Figure 2.13a,b**), and then selected a final library of 70,000 top scoring guides for synthesis. The resulting sgRNA sequence and their corresponding targets were separated by a common 20 bp spacer sequence and ordered as an Agilent SureGuide Unamplified Custom CRISPR Library array (**Figure 2.1**A).

To analyze the potential impact of programmed microhomology, we selected a subset of 1,000 sgRNA-target pairs from the library above and introduced microhomology with different lengths (2 bp, 4 bp, and 6 bp) matching the last 2, 4, and 6 nucleotides upstream of the cleavage site. Each design was assigned a 4 bp barcode, indicating its programmed microhomology pattern (**Figure 2.5a,b**). This library of microhomology sequences was ordered as an oligo pool from Twist Biosciences.

2.5.2 Library Cloning

The lentiGuide-Puro (Addgene #52963) vector was modified with two rounds of PCR to remove the existing tracrRNA and filler sequence (primer P1, P2), and to incorporate two BsmBI restriction site for integration of sgRNA-target pairs (primer P3, P4). The modified vector was digested with BsmBI (NEB, Buffer 3.1) at 55 for 3h and gel purified with Monarch DNA Gel Extraction Kit (NEB). This digested and purified vector was used

for all downstream cloning.

Oligos with sgRNA-target pairs from Agilent or Twist Bioscience were both resuspended to 10ng/μL. The oligo pool was PCR amplified using KAPA Biosystems HiFi HotStart ReadyMix 2x using primers P5 and P6 and cleaned with the DNA Clean&Concentrator kit (Zymo Research). The purified PCR product was then digested with BsmBI (NEB, buffer 3.1) at 55°C for 1h to generate compatible sticky ends matching the modified lentiGuide-Puro above, and subsequently cleaned with DNA Clean&Concentrator (Zymo Research). Digested vector and insert were ligated with T4 ligase (NEB) with a molar ratio of 1:3. Ligation products were transformed into Stable Competent E.coli (NEB C3040H). Transformed cells were cultured at 30°C overnight and plasmid DNA was prepared using a ZymoPURE II Plasmid Kit. The subsampled library with 12,917 targets was bottlenecked by seeding transformed cells on plate. Colonies on plates were transferred to liquid medium to expand them. The precision of the number 12,917 follows from the fact that we can simply count the number of unique guide sequences present in deep sequencing of PCR amplicons from cells.

2.5.3 Cell Culture and lentivirus transduction

We generated a mono-clonal 293T cell line expressing Cas9 by transduction of Cas9-blast lentivirus particles (Addgene plasmid #52962). Cells were cultured in DMEM High glucose (GIBCO) supplemented with 10% Fetal Bovine Serum (Rocky Mountain Biologicals) and 1% penicillin-streptomycin (GIBCO) and grown with 5% CO2 at 37°C.

All lentivirus libraries were produced by the Fred Hutchinson Cooperative Center for Excellence in Hematology Vector Production core facility. HEK293T cells were transduced and media was changed to virus free media at 24 hours post-transduction. Cells were passed every 48h with a split ratio of 1:6. Cells were harvested at day 5 after transduction.

2.5.4 Sequencing Library Generation

Genomic DNA was extracted with DNeasy Blood & Tissue Kit (Qiagen) following the manufacturer's protocol. 15 bp unique molecular identifiers (UMIs) were added by one initial round of linear PCR using a primer containing a 5′ sequencing adaptor (P7). For each reaction we used 250ng of genomic DNA, 0.2μL 100mM primer and 25 μL HiFi HotStart ReadyMix 2x (KAPA Biosystems). PCR reaction were performed as follows: 95°C 3 mins, 98°C 20 s, 5 cycles of 65°C 1 min and 72°C 2 min, 98°C 20 s, 5 cycles of 65°C 1 min and 72°C 2 min. The subsequent PCR product was cleaned with 1.8x AMPure XP beads (Beckman Coulter) and resuspended in 25μL of elution buffer. A second round of amplification was performed using primers targeting the 5′ sequencing adaptor (P8) and 50 bp downstream of the cleavage site (P9) for 20 cycles. The resulting PCR product was then size selected using a dual size-selection cleanup of 0.4x and 0.8x AMPure XP beads (Beckman Coulter) to remove genomic DNA and small fragments (<200 bp) respectively. This size-selected product was subsequently re-amplified to add the 3′ sequencing adaptor with primer P8 and P10 for an additional five cycles. The final PCR product was cleaned with 0.75x AMPure XP beads (Beckman Coulter) and was re-amplified to add flow-cell adaptor and sample index for 5 cycles. All PCR reactions used HiFi HotStart ReadyMix 2x (KAPA Biosystems) with the manufacturer's recommended conditions. The library was sequenced on an Illumina NextSeq 500 sequencer using paired-end 150 cycle reads. All primers used are listed in **Table 2.2**. Sequence data and associated data files are deposited in Figshare with a doi link: https://doi.org/10.6084/m9.figshare.7374155,

2.5.5 Sequence processing pipeline

Across three replicates, we sequenced a total of 148 million paired-end reads on an Illumina NextSeq 500. We first clustered these paired-end reads by their 15 bp UMI sequence and then filtered out reads with less than 90% identity within their representative UMI clusters. Sequence identity was identified using edlib [66]. UMIs with fewer than 10

reads were excluded from downstream analysis. This yielded 4,405,379 UMIs (91,325,700 reads), representing 61.8% of our sequencing data (**Table 2.3**). We then selected the most common forward and reverse read sequence for each UMI for further processing. These forward and reverse reads were merged into a single read using PEAR [67] and aligned in a two step process as follows. First, we sought to identify the reference sequences for each programmed array sequence. We aligned the merged reads to a backbone sequence where the guides and targets were represented by Ns using EMBOs needleall software [41] with the following scoring matrix: match=5, mismatch=-4, gap-open=-20, gap-extension=-0.5. The mismatch penalty for Ns was set to 0. The sequence over the guide region was then extracted and matched against the list of programmed array sequences. Guide sequences with more than 2 mismatches to the designed guides were excluded, with edit distances assessed with UMI-tools [68]. Second, merged reads were aligned to their discovered reference, in which Ns were replaced by the guide/target sequence identified from the first step, using Biopython.pairwise2 [69] with the following scoring matrix: match=5, mismatch=-4, gap-open=-13, gap-extension=-0.5. All indels were then right aligned (*e.g.* **Figure 2.9a**). Aligned reads with indels within -3/+2 bp of the cleavage site were assigned to their indel class. Aligned reads were excluded for downstream analysis if the sgRNA and target sequence didnt match, the result from template switch during lentivirus transduction [40, 39], or unexpected mutations introduced during synthesis, cloning, and PCR. A final library of 1.19 million unique reads (UMIs) were identified. Our library of 1,000 microhomology sequences were processed by this same pipeline, yielding a final library of 249,039 UMIs from 31,239,645 paired-end sequencing reads. Scripts and other software are available from our GitHub repository: https://github.com/shendurelab/Lindel.

2.5.6 Data processing and analysis

kpLogo Analysis: Sequence motif analysis was conducted with kpLogo [44] using default settings with a specified k-mer length of 1 or 2. Input sequences were weighted by the frequency of insertion. Microhomology identification: For n from 1 to 10 nucleotides, the last

n nucleotides upstream of each deletion were compared to the last n nucleotides of the deleted sequence (as the deletion is right aligned. **Figure 2.9a**). The length of microhomology was identified as the largest n nucleotides match in sequence.

2.5.7 Machine learning modeling

We phrased our problem of predicting repair outcomes and their frequencies as that of a classification task with 557 classes. Because large mutation events are rare, we limited our classification effort to deletion events < 30 bp, and we grouped insertions ≥ 3 bp into one class. In total, we defined 557 classes of indels. These classes include 536 deletion alleles, 4 possible single nucleotide insertion, and 16 possible dinucleotide insertion and insertions $\geq$ 3 bp. There are a total of 550 potential deletion events that are both < 30 bp in length and overlap with the -3/+2 window around the cleavage site. We captured 536 deletion alleles in our data; the missing 14 classes are mainly large deletions. As input to our model, we defined 3,033 binary features. These are 1) Sequence features: 384 binary features corresponding to the one-hot encoded target sequence (excluding the PAM region), including 80 for single nucleotide content (4 nucleotides$\times$20 positions) and 304 for dinucleotide content (16 dinucleotides$\times$19 positions); 2) Microhomology features: 2,649 binary features corresponding to MH tracts; specifically, for each of the possible deletion event class, we defined 5 binary features (or 2-4, depending on the size of deletion) corresponding to the length of the MH tract, if any (0-4 bp $\times$ 519 + 0-3 bp $\times$ 7 + 0-2 bp $\times$ 6 + 0-1 bp $\times$ 4 deletion event classes = total 2,649 binary features. Our 4,790 programmed sequences were randomly partitioned into a training set of 3,900 sequences, a validation set of 450 sequences, and a test set of 440 sequences.

We trained the logistic regression in a standard manner for machine learning models. However, because each target sequence can generate many possible repair outcomes, we trained our models using soft labels that correspond to the probability that each class is observed, rather than hard labels that force each input to correspond exclusively to one class. Each model was trained using the Adam optimizer [70] with a learning rate of 0.001 and

a categorical cross-entropy loss. Training proceeded for a maximum of 100 epochs with a patience of 1, meaning that training was stopped after two epochs with no improvement in validation set performance. All initializations and the hyperparameters for the Adam optimizer were set to the defaults in Keras v2.1.3 [71] with a backend of Theano v1.0.1 [72]. We selected the best model based on performance on the validation set according to the coefficient of determination using grid search over hyperparameters. This search involved separate scans over regularization strengths for L1-regularization and L2-regularization individually with a range of 10^{-10} to 10^{-1}(**Figure 2.11d-f**).

2.5.8 Model comparison

We compared our model to two other models (ForeCasT and inDelphi) in the same setting. All models used a 60 bp sequence centered at the cleavage site as an input, while trying to predict the frequency of 557 classes of indels that we defined above. As both Lindel and ForeCasT are predicting all possible unique repair outcomes, its straightforward to compare them directly. inDelphi only predicts 90 classes deletions with locations, and 1-59 bp deletions regardless of their locations. We used the classes that inDelphi predicts that overlap with the ForeCasT and Lindel classes, which included all 90 classes using microhomology, 1 bp deletion (assuming its located the cleavage site), and 1 bp insertions. All classes that inDelphi is not predicting were assigned as 0. The performance were measured using MSE on all 450 unique classes for each sequence in our test set (n = 440) and the ForeCasT test set (n = 4,298)

2.6 Figures and Tables

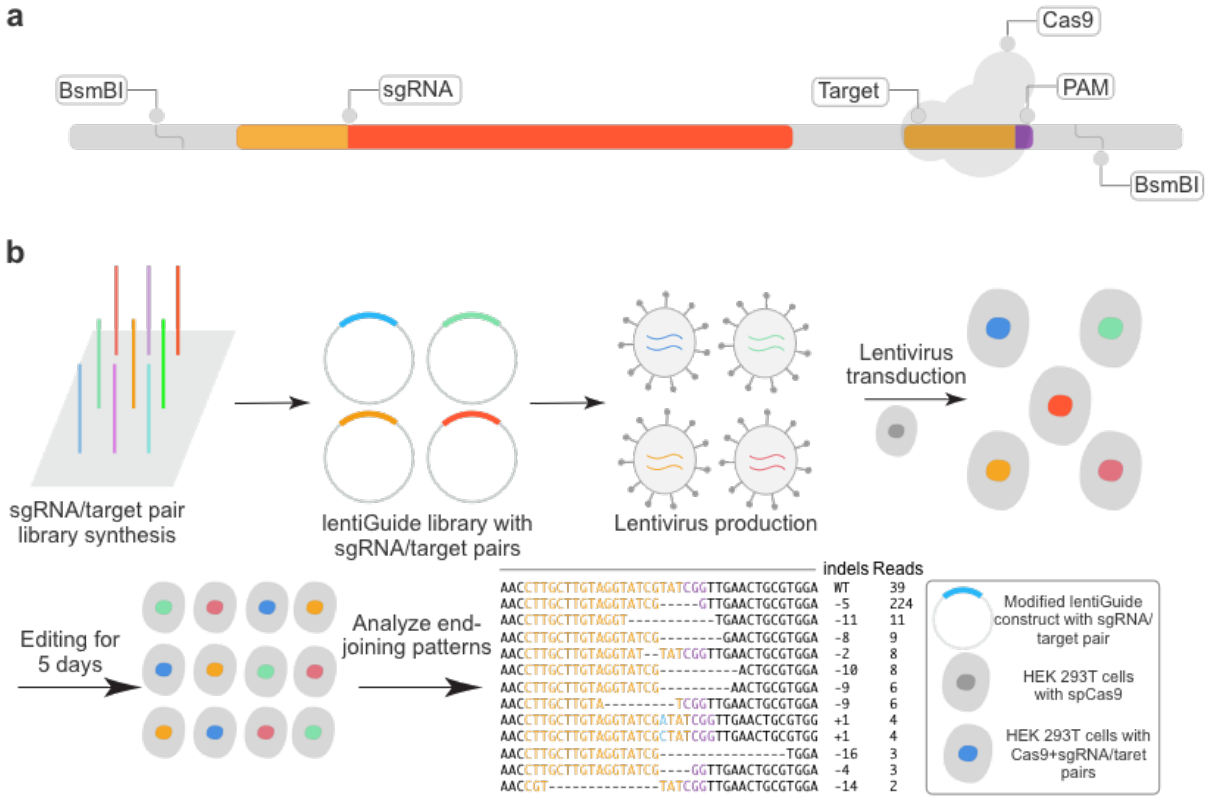

Figure 2.1: An assay for massively parallel profiling of the outcomes of CRISPR/Cas9-mediated double-stranded DNA break repair. **a.** Schematic of library of 200 bp oligonucleotides encoding sgRNAs (red) targeting a large number of designed 20 bp spacers, with their matched target sequence encoded in cis (yellow: target; PAM: purple). In our primary experiment, 70,000 target sequences were designed and cloned. **b.** After array-based synthesis and PCR amplification of the library, BsmBI restriction sites at either end were used for cloning into a modified lentiviral construct. The library was bottlenecked to 12,286 targets to facilitate greater coverage of independent NHEJ-mediated events corresponding to each target. Monoclonal HEK293T cells expressing Cas9 were transduced with packaged lentivirus. Cells were harvested at 5 days after transduction, and a region including both the spacer and the target was PCR amplified from genomic DNA for high-throughput sequencing. The sequences of mutated targets were aligned to their corresponding unmutated reference, assigned based on the spacer sequence (yellow: target; PAM: purple; green: inserted bases; dashes: deleted bases).

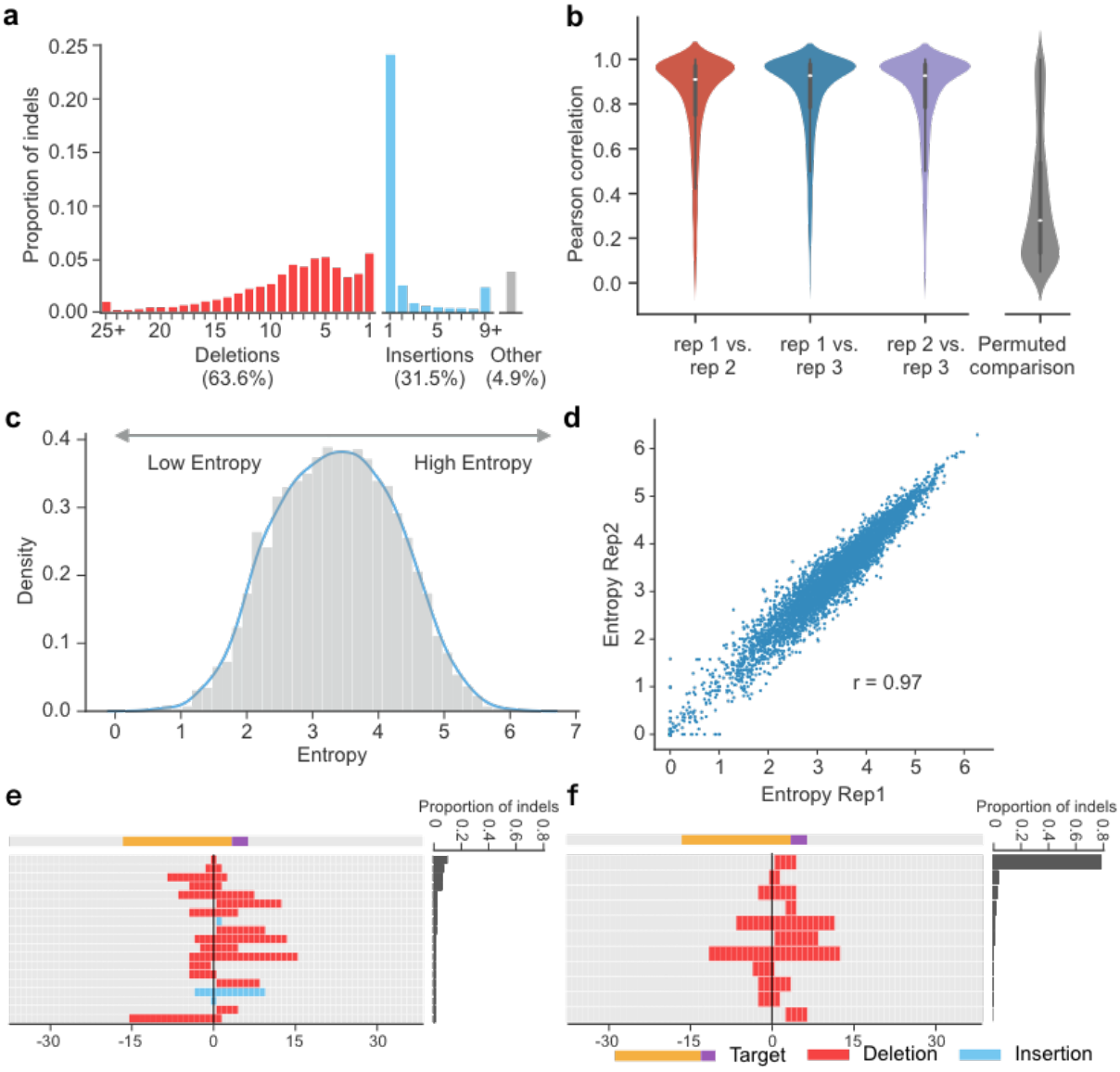
a
Proportion of indels
0.25
0.20
0.15
0.10
0.05
0.00
25+ 20 15 10 5 1
1 5 9+
Deletions
(63.6%)
Insertions
(31.5%)
Other
(4.9%)
b
Pearson correlation
1.0
0.8
0.6
0.4
0.2
0.0
rep 1 vs.
rep 2
rep 1 vs.
rep 3
rep 2 vs.
rep 3
Permuted
comparison
c
Low Entropy
High Entropy
Density
0.4
0.3
0.2
0.1
0.0
0 1 2 3 4 5 6 7
Entropy
d
Entropy Rep2
6
5
4
3
2
1
0
r = 0.97
0 1 2 3 4 5 6
Entropy Rep1
e
Proportion of indels
0 0.2 0.4 0.6 0.8
-30 -15 0 15 30
f
Proportion of indels
0 0.2 0.4 0.6 0.8
-30 -15 0 15 30
Target
Deletion
Insertion

Figure 2.2: Mutation patterns resulting from DSB repair vary greatly between targets, but are highly reproducible for individual targets. a. Overview of indel profiles. The histogram represents the indel rate per target, based on aggregated data from three replicates. The x-axis corresponds to the size of insertion or deletion events. Of all detectably mutated targets, 63.6% were deletions (red) and 31.5% were insertions (blue). The remainder (4.9%) contained some combination of substitutions, insertions and deletions, and are excluded from all subsequent analyses. **b.** End-joining patterns were highly reproducible for the same target between replicates. Left: violin plot of distribution of correlation coefficients for pairwise comparison of individual targets between replicates. Right: Permuting the allele counts for each target in one replicate and repeating the pairwise comparison greatly reduces the observed correlations. **c.** Entropy quantifies the diversity of NHEJ outcomes from individual targets. Targets were separated into low, medium and high entropy classes. **d.** Estimated entropy for individual targets was highly reproducible between replicates (rep1 vs. rep2 shown). **(e,f.)** Example of targets with high and low entropy. High entropy targets had diverse outcomes at appreciable frequencies **e.** while low entropy targets were dominated by a single outcome **f**).

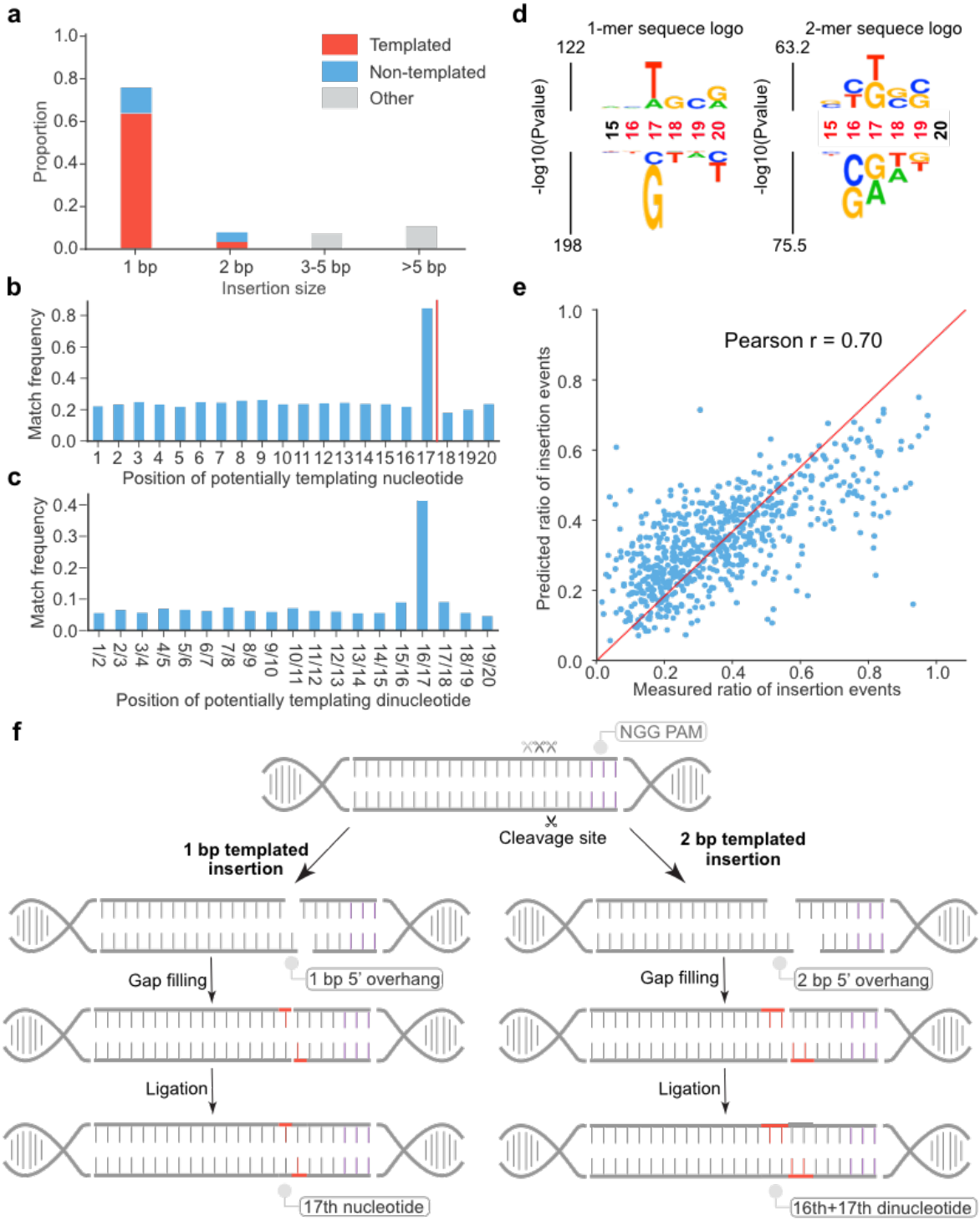
a
Templated
Non-templated
Other
Proportion
1 bp
2 bp
3-5 bp
>5 bp
Insertion size
b
Match frequency
Position of potentially templating nucleotide
c
Match frequency
Position of potentially templating dinucleotide
d
1-mer sequece logo
2-mer sequece logo
-log10(Pvalue)
122
198
63.2
75.5
e
Pearson r = 0.70
Predicted ratio of insertion events
Measured ratio of insertion events
f
NGG PAM
Cleavage site
1 bp templated insertion
2 bp templated insertion
Gap filling
1 bp 5' overhang
Ligation
17th nucleotide
Gap filling
2 bp 5' overhang
Ligation
16th+17th dinucleotide

Figure 2.3: A model for asymmetric templating of NHEJ-mediated insertion events at sites of CRISPR/Cas9-mediated DSBs. a. 75.3% of the insertions were 1 bp. Of 1 bp insertions, 85% appear to be templated. **b,c.** Histogram of the number of 1 bp **b.** or 2 bp **c.** insertion events where the inserted base or dinucleotide is identical to the base at a specific position in the target. The canonical DSB site is between 17^{th} and 18^{th} of the target sequence (red line). The result suggests many 1 bp insertions are templated by the nucleotide at the 17^{th} position but not the 18^{th} position **b.** and many 2 bp insertions are templated by dinucleotide at the 16^{th} and 17^{th} positions **c.**. **d.** The immediate sequence context surrounding the DSB strongly biases the proportion of NHEJ-mediated outcomes that result in insertions vs. deletions. The 1-mer sequence logo (left) shows that the presence of a T and A at the 17^{th} position increased the ratio of insertions. The 2-mer sequence logo (right) shows that the presence of a TG dinucleotide at the $17^{th}/18^{th}$ position increased the ratio of insertions, while a CG dinucleotide at the $16^{th}/17^{th}$ position, or a GA dinucleotide at the $17^{th}/18^{th}$ position, decreased the ratio of insertions. Significant positions are colored in red. **e.** A regression model using the nucleotide content of a 6 bp window centered on the DSB site predicted the ratio of insertion-to-deletion events. **f.** A model for how insertions at CRISPR/Cas9-mediated DSBs are asymmetrically biased by local sequence context. Local sequence context biases the pattern of cleavage of the non-complementary strand to the sgRNA, resulting in different frequencies of blunt vs. $5'$ overhangs for different targets. This in turn biases the ratio of insertions vs. deletions, as $5'$ overhangs are preferably repaired by gap-filling (red) and ligation, resulting in the observed preponderance of 1 bp or 2 bp templated insertions (red).

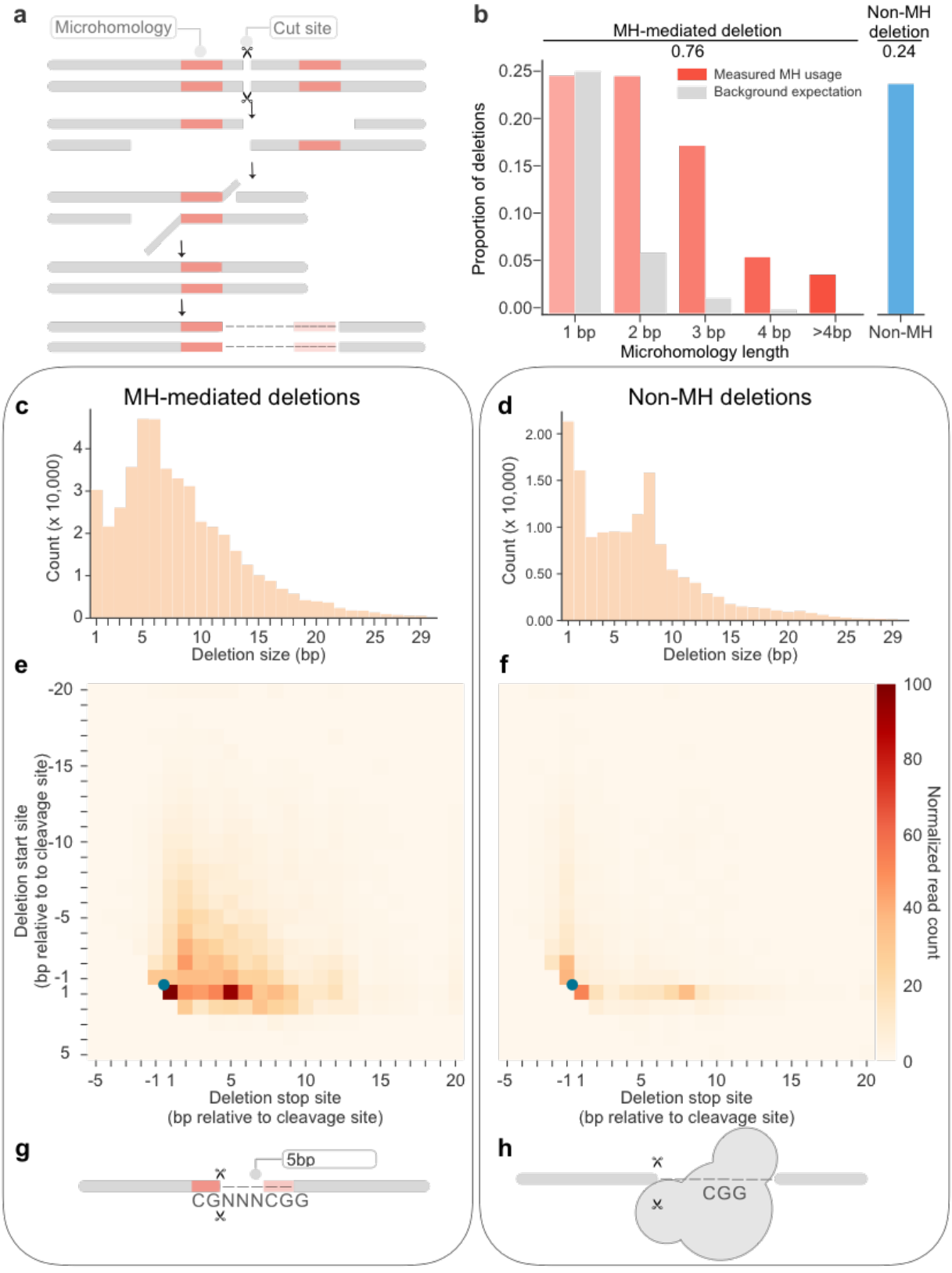
a
Microhomology
Cut site
b
MH-mediated deletion
0.76
Non-MH deletion
0.24
Measured MH usage
Background expectation
Proportion of deletions
0.25
0.20
0.15
0.10
0.05
0.00
1 bp
2 bp
3 bp
4 bp
>4bp
Non-MH
Microhomology length
c
MH-mediated deletions
Count (x 10,000)
Deletion size (bp)
d
Non-MH deletions
Count (x 10,000)
Deletion size (bp)
e
Deletion start site
(bp relative to to cleavage site)
Deletion stop site
(bp relative to cleavage site)
f
Deletion stop site
(bp relative to cleavage site)
Normalized read count
g
5bp
CGNNNCGG
h
CGG

Figure 2.4: Extensive use of microhomology in NHEJ-mediated deletion events. **a.** Schematic of microhomology (MH) usage in end-joining repair. Tracts of MH (red) in the vicinity of the DSB are used to align the broken ends. The unannealed overhang is cleaved by endonuclease and the gap filled by polymerase. Here, a deletion event is defined as MH-mediated deletion if the sequence at the $3'$ of a rejoined end (red, left) is identical to the $3'$ end of the deleted sequence (red, right). The size of the MH tract refers to the length of that identical sequence. **b.** Length distribution of MH tracts in observed MH-mediated events. With the exception of 1 bp deletions, all MH tract lengths occured at substantially greater than expected frequencies. **(c,d.)** Distribution of deletion sizes of MH-mediated **c.** and non-MH **d.** events. **(e, f.)** Heatmap of showing frequency of start/stop sites of MH-mediated **e.** and non-MH **f.** deletion events. The Y and X axes correspond to the start and stop sites of deletion events, respectively, with positions shown relative to the canonical DSB site (blue dot). Both MH-mediated and non-MH deletions were primarily unidirectional relative to the DSB site, rather than spanning it. **g.** Schematic of potential explanation for the observed excess of 5-6 bp MH-mediated deletions. PAM-like sequences near the DSB are biased towards deletion events. Microhomology between a G at the17^{th} position or a CG at the $16^{th}/17^{th}$ position with corresponding sequences in the PAM result in an excess of 5-6 bp deletions. **h.** Schematic of potential explanation for the observed excess of 8 bp non-MH deletions. In the dsDNA-sgRNA-Cas9 complex, the region 1-8 bp downstream of the cleavage site is occupied by Cas9. The enrichment of non-MH deletions 8 bp from the cleavage site could simply correspond to the nearest position lacking Cas9 protection from endonucleases during repair.

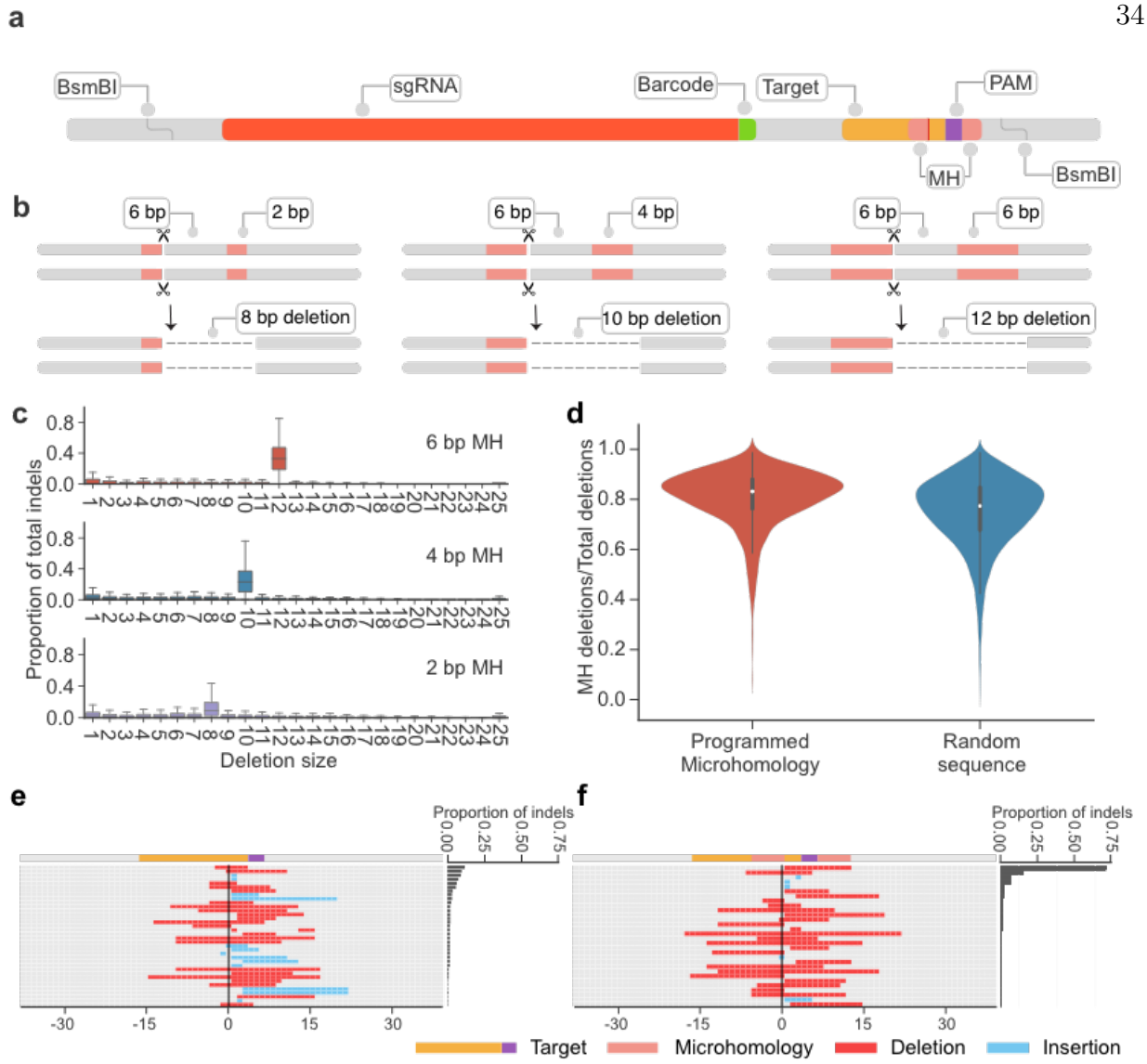

Figure 2.5: Programming microhomology tracts into targets increases predictability of repair outcomes. a,b. Schematic of programmed MH tract designs, which starts immediately downstream of PAM sequence (purple), and expected deletion sizes. The distance between the regions of MH (pink) was consistently 6 bp, while the MH tracts were 2 bp, 4 bp or 6 bp, such that the expected deletion sizes were 8 bp, 10 bp and 12 bp, respectively. **c.** Distribution of observed deletion sizes for 1,000 targets with programmed MH tracts of various lengths, based on a single replicate/experiment. Each boxplot summarizes the ratio of certain deletions for each target. The box represents 25^{th} percentile, 50^{th} percentile and 75^{th} percentile and whiskers represent 1.5x of the inter-quartile range (IQR). We observe a strong bias towards deletions of the expected lengths, with the proportion increasing for longer MH tracts (median 0.080, 0.209, and 0.318 for 2 bp MH, 4 bp MH and 6 bp MH, respectively). **d.** MH usage in sequences with (left) or without (right) programmed MH. Despite the strong bias towards intended deletions when MH occurred, the proportion of MH events only slightly increased from 76% to 82%. **e,f.** Example of a sequence that shows diverse editing outcomes **e.**. However, when a 6 bp MH tract is introduced onto this sequence backbone, the programmed 12 bp deletion comprises nearly 75% of the editing outcomes **f**.

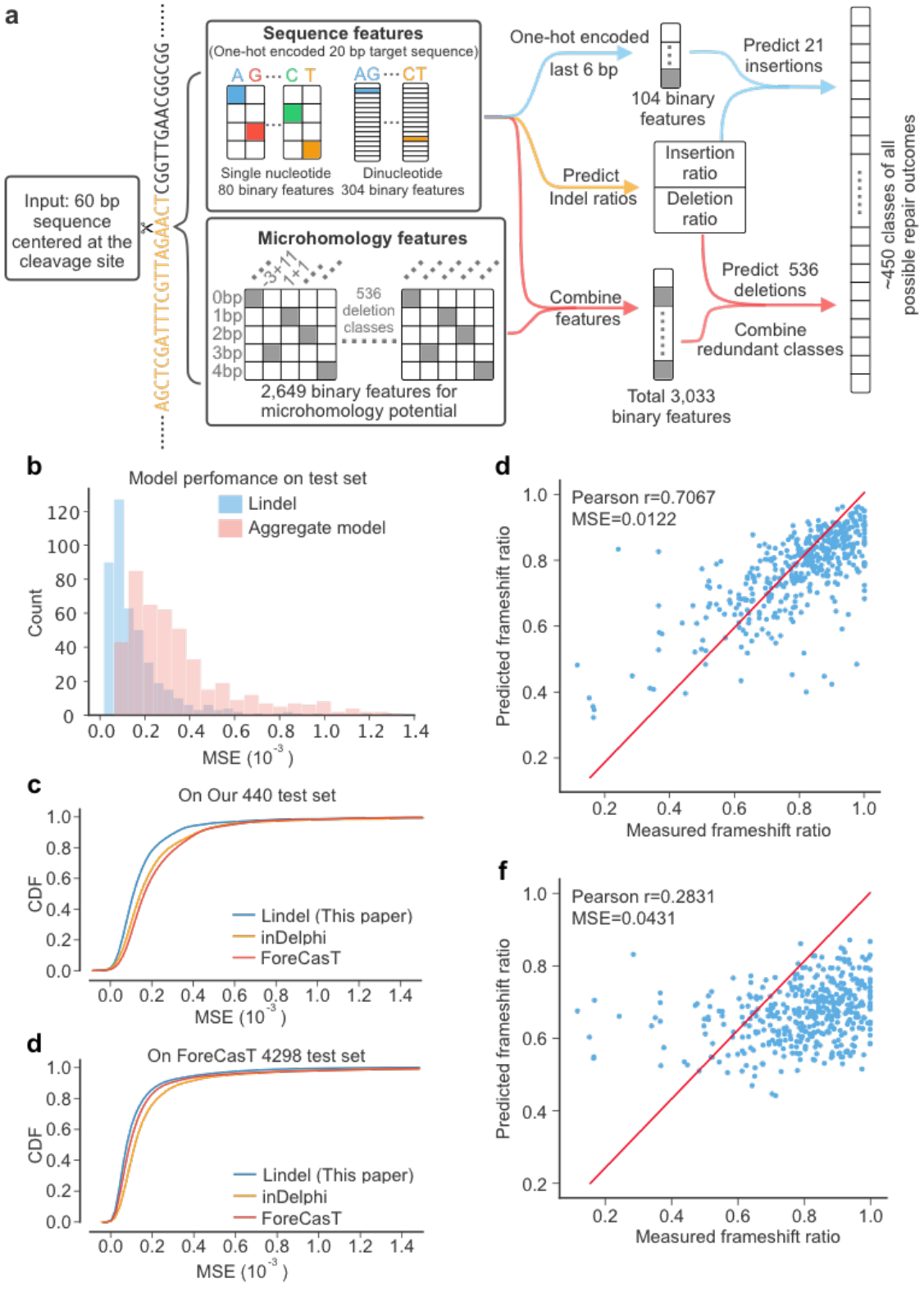
a
Input: 60 bp sequence centered at the cleavage site
AGCTCGATTTCGTTAGAACTCGGTTGAACGGCGG
Sequence features
(One-hot encoded 20 bp target sequence)
A G ··· C T
AG ··· CT
Single nucleotide 80 binary features
Dinucleotide 304 binary features
Microhomology features
-3+1
1+1
0bp
1bp
2bp
3bp
4bp
536 deletion classes
2,649 binary features for microhomology potential
One-hot encoded last 6 bp
104 binary features
Predict 21 insertions
Predict Indel ratios
Insertion ratio
Deletion ratio
Combine features
Predict 536 deletions
Combine redundant classes
Total 3,033 binary features
~450 classes of all possible repair outcomes
b
Model perfomance on test set
Lindel
Aggregate model
Count
MSE (10^-3)
c
On Our 440 test set
CDF
Lindel (This paper)
inDelphi
ForeCasT
MSE (10^-3)
d
On ForeCasT 4298 test set
CDF
Lindel (This paper)
inDelphi
ForeCasT
MSE (10^-3)
d
Pearson r=0.7067
MSE=0.0122
Predicted frameshift ratio
Measured frameshift ratio
f
Pearson r=0.2831
MSE=0.0431
Predicted frameshift ratio
Measured frameshift ratio

Figure 2.6: End joining patterns are accurately predicted by Lindel. a. Schematic of machine learning framework for Lindel. A 60 bp sequence (± 30 bp around the cleavage site) is used as the input to the model. A total of 3,033 binary features – 2,649 corresponding to MH potential and 384 to one-hot encoded mono- and di-nucleotide content of the 20 bp target – are extracted. One-hot encoded sequence features were used to predict the overall ratio of insertion to deletion events. One-hot encoded sequences corresponding to the last 6 bp of the target sequence were used to predict 21 insertion classes. Both sequence features and microhomology features were used to predict deletion classes. Probabilities of redundant deletion classes were combined in a sequence specific manner. **b.** Performance of Lindel on the test dataset. The distribution of MSE values for the 440 test targets is shown (blue). Poorly predicted targets largely correspond to those that were poorly sampled (see **Figure 2.11**b). As a baseline to illustrate the improvement conferred by Lindel, we show a similar distribution for the aggregate model, in which the predicted frequencies of 557 indel classes are simply taken from the aggregate frequency at which each is observed in the training and validation datasets (red). **(c, d.)** The performance of Lindel trained using training data aggregated from this study and the ForeCasT study (3,900 sequences from this study and 10,725 from (32)). A Lindel model that was trained on the aggregated training datasets performed best on both the test sets from this study (440 sequences) and the ForeCasT study (4,298 sequences). **(e, f.)** Lindel **e.** compared favorably to Microhomology Predictor [53] **f.** in predicting the ratio of frameshifting mutations for each of the 440 targets in the test set.

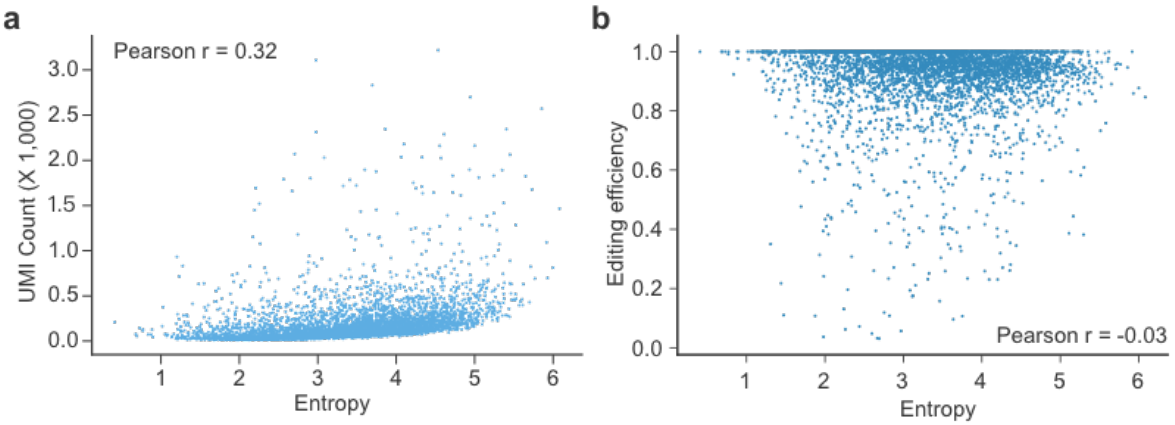

Figure 2.7: Correlation between entropy vs. UMI count or editing efficiency. a. Estimates of target-specific entropy (x-axis) are only modestly correlated with UMI counts (y-axis). Pearson's $r = 0.32$. **b.** Estimates of target-specific entropy (x-axis) are not correlated with editing efficiency (y-axis). Pearson's r = -0.03.

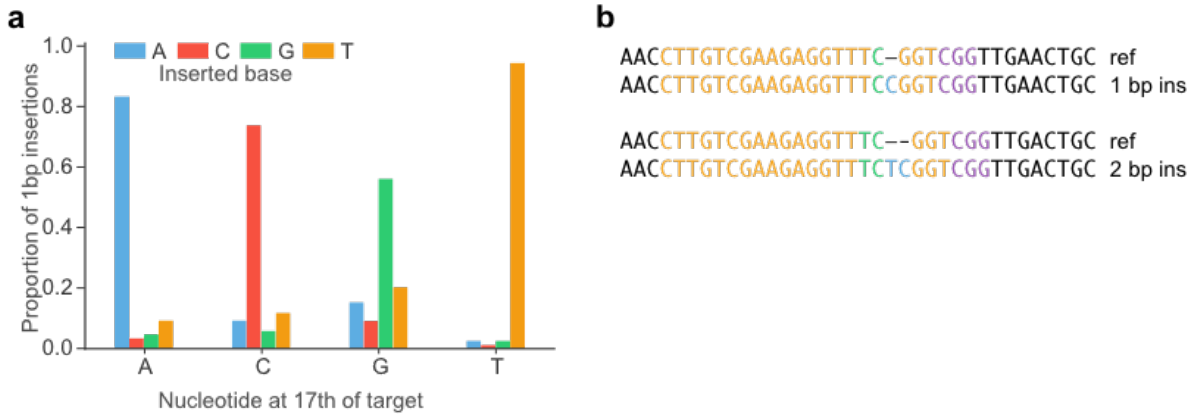

Figure 2.8: 1-2 bp insertion events are templated by the nucleotides upstream of the cleavage site. a. Most 1 bp insertions were predicted, and presumably templated, by the identity of the 17^{th} nucleotide of the target sequence. **b.** Example of insertions templated by the 17^{th} (top) or 16^{th} and 17^{th} (bottom) position. Template nucleotides are shown in green and inserted nucleotides are shown in blue.

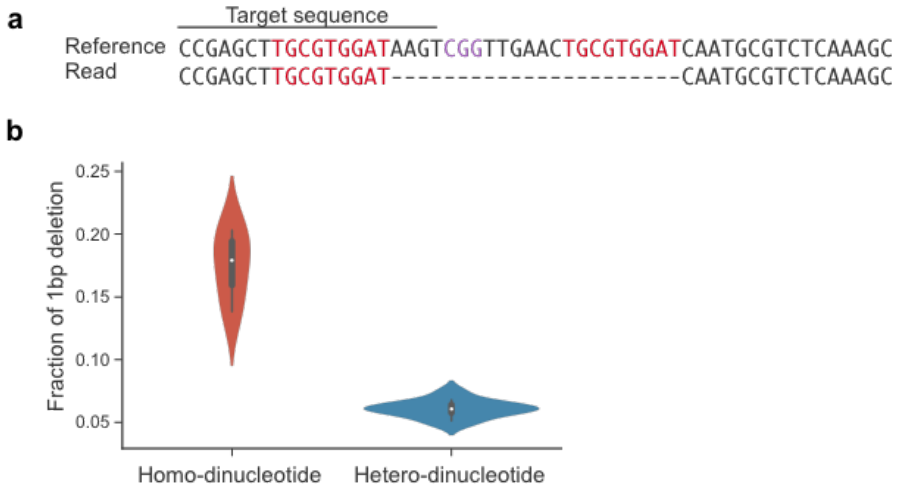

Figure 2.9: Examples of microhomology usage. a. An observed example of a long MH tract mediating a deletion event. PAM and microhomology are shown in purple and red, respectively. This particular outcome, involving a 9 bp MH tract, represented 9% of indel events associated with this target. **b.** Targets with identical nucleotides (*i.e.* homo-dinucleotide) spanning the cleavage site exhibit a much higher proportion of 1 bp deletions than targets with non-identical nucleotides (*i.e.* hetero-dinucleotide) spanning the cleavage site, suggesting that 1 bp microhomology may help mediate 1 bp deletion events.

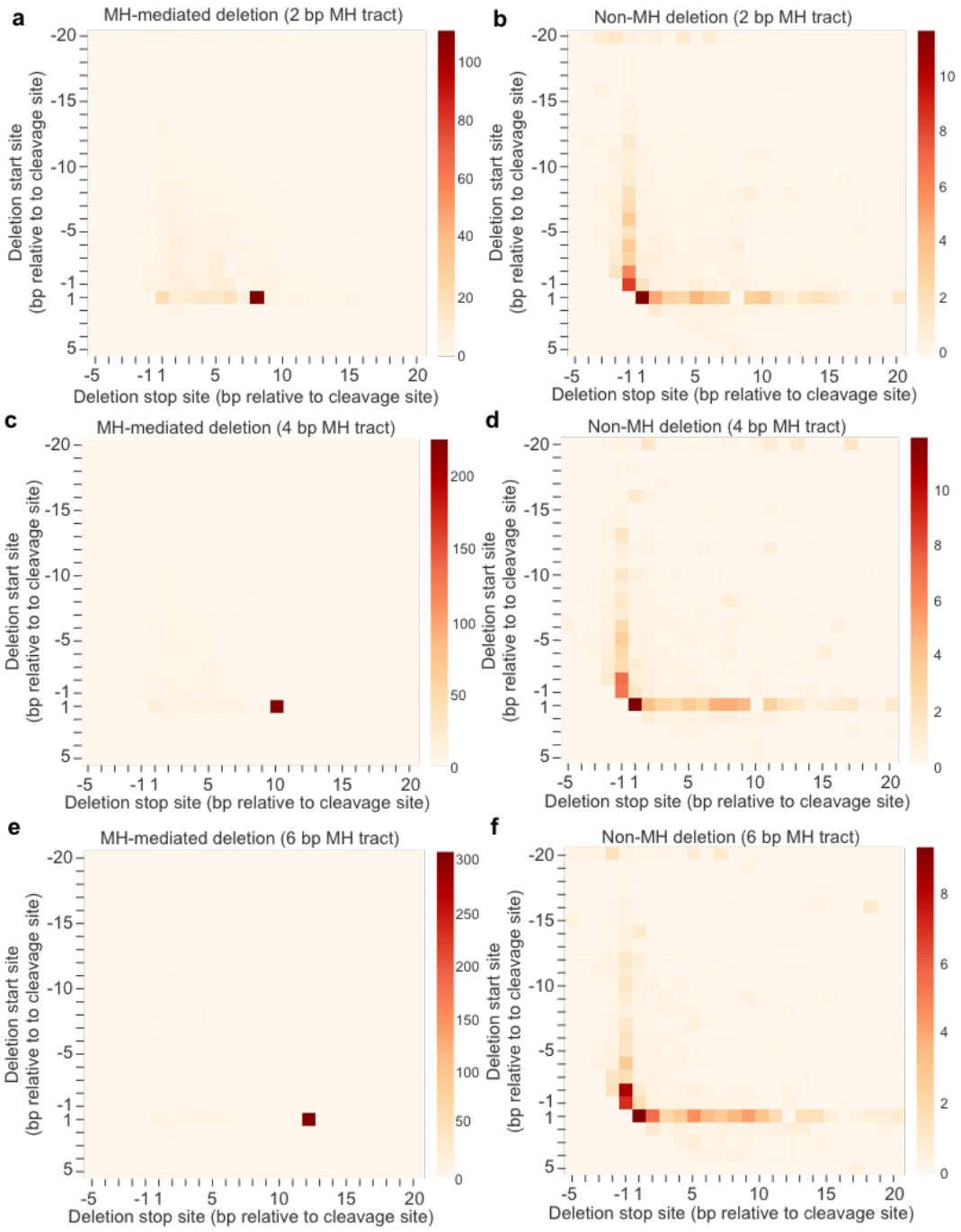

Figure 2.10: Heatmap of deletions with microhomology design. a, c, e. Heatmap of showing frequency of start/stop sites of MH-mediated deletions with 2 bp, 4 bp, 6 bp programmed microhomology, respectively. **(b, d, f.)** Heatmap of showing frequency of start/stop sites of non-MH deletions with 2 bp, 4 bp, 6 bp programmed microhomology, respectively.

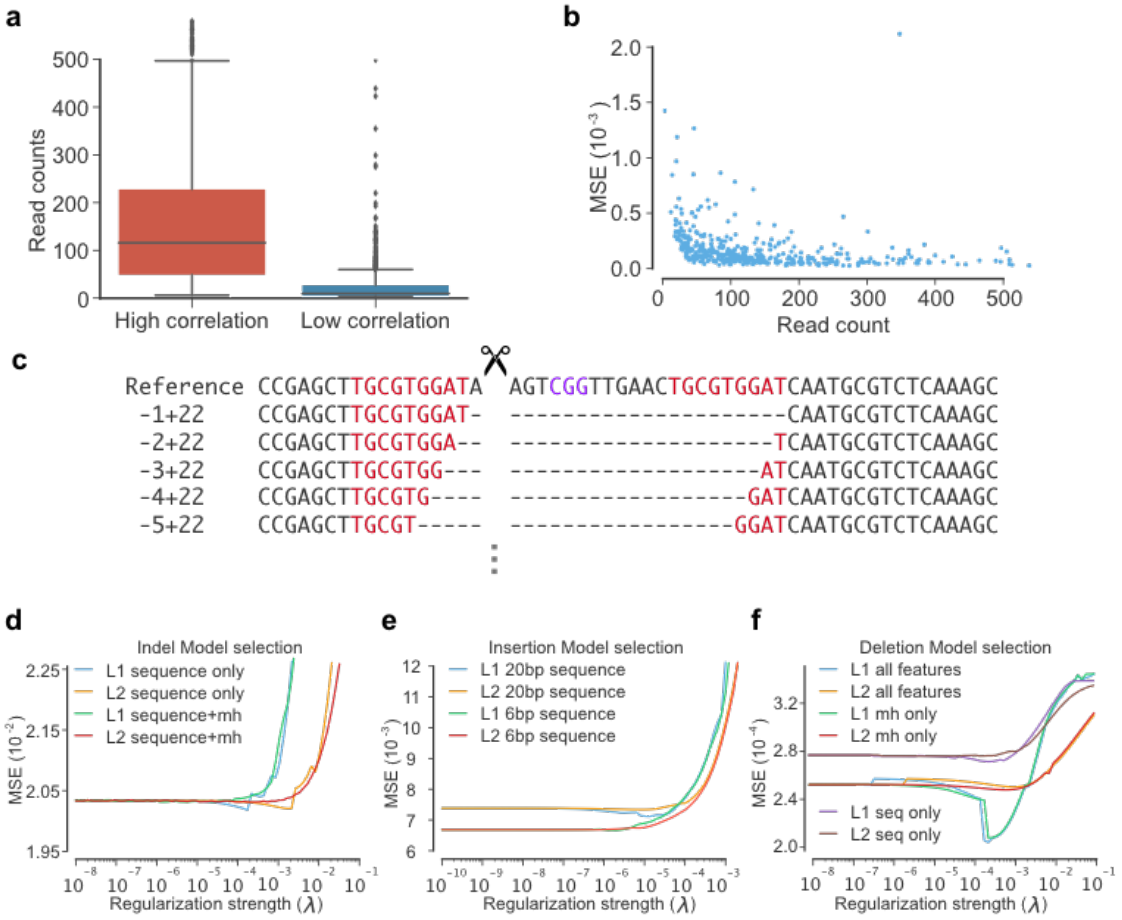

Figure 2.11: Machine learning model selection and performance. a. Read counts for targets exhibiting high (r>0.75) vs. low (r<0.75) correlation between replicate experiments. The median read count for the two groups are 117 and 23, respectively. Targets with low correlation between replicates (r<0.75) were excluded from model training/validation/testing. **b.** Poorly predicted targets (high MSE) largely corresponded to those that were poorly sampled. **c.** Example of redundant deletion classes. We define deletion classes using the deletion start site and deletion length (*e.g.* -1 + 22 where -1 is the location relative to the cleavage site and 22 is the deletion length). For any given sequence, there may be several deletion classes that represent identical outcomes due to microhomology (red). We collapsed the probability of these classes in prediction. PAM is colored purple. **(d, f.)** Model selection for indel ratio prediction, insertion prediction and deletion prediction. Hyperparameter search involved separate scans over regularization strengths for L1-regularization and L2-regularization individually with a range of 10^{-10} to 10^{-1}. MSE on the validation set is plotted and was used to pick the best performing model.

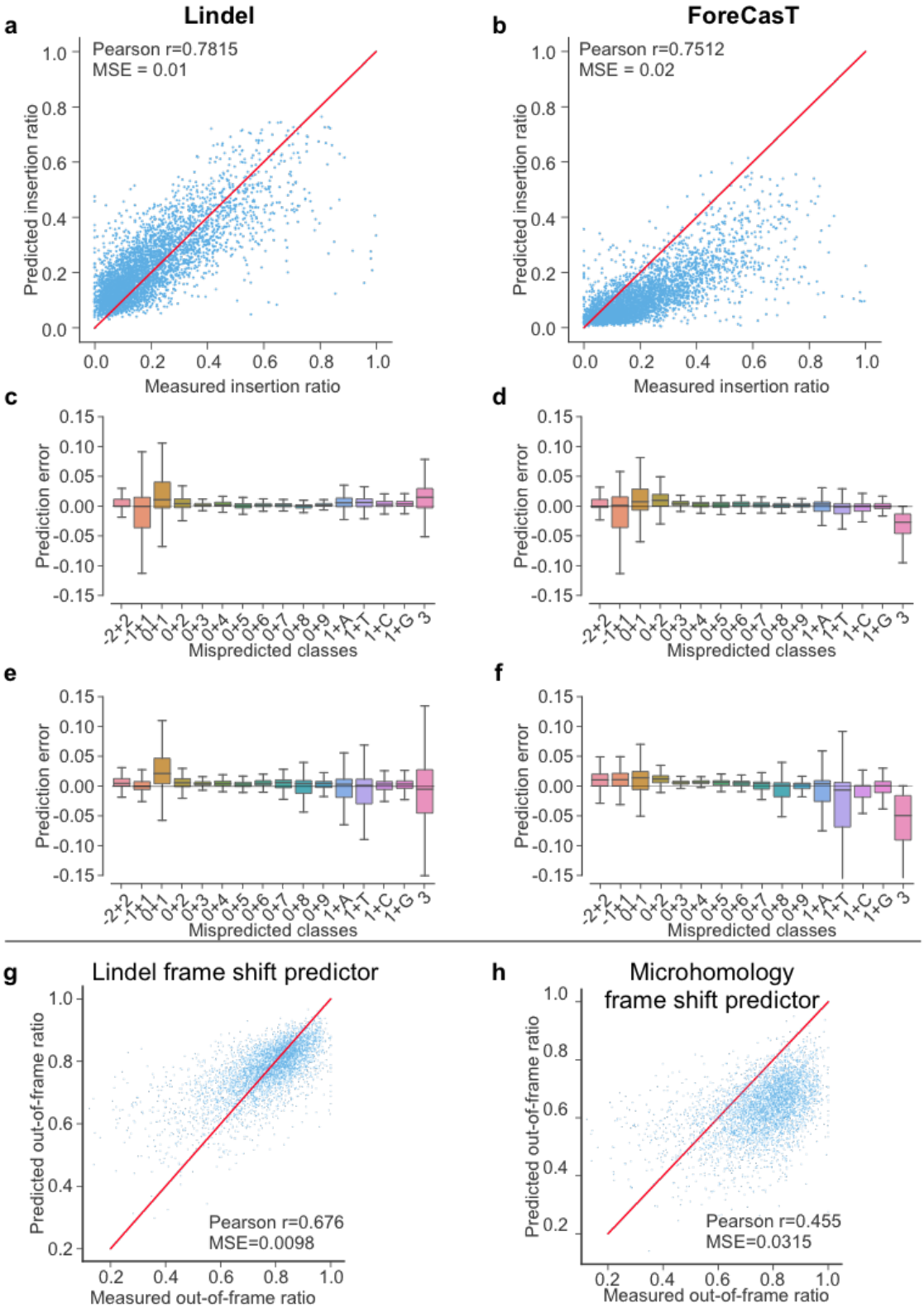
a
Lindel
Pearson r=0.7815
MSE = 0.01
Predicted insertion ratio
Measured insertion ratio
b
ForeCasT
Pearson r=0.7512
MSE = 0.02
Predicted insertion ratio
Measured insertion ratio
c
Prediction error
Mispredicted classes
d
Prediction error
Mispredicted classes
e
Prediction error
Mispredicted classes
f
Prediction error
Mispredicted classes
-2+2
-1+1
0+1
0+2
0+3
0+4
0+5
0+6
0+7
0+8
0+9
1+A
1+T
1+C
1+G
3
g
Lindel frame shift predictor
Pearson r=0.676
MSE=0.0098
Predicted out-of-frame ratio
Measured out-of-frame ratio
h
Microhomology
frame shift predictor
Pearson r=0.455
MSE=0.0315
Predicted out-of-frame ratio
Measured out-of-frame ratio

Figure 2.12: Performance of Lindel on ForeCast test I actuset. a, b. Lindel **a.** compared favorably to ForeCasT **b.** in predicting the overall indel ratio for each of the 4,298 targets. **(c, f.)** Mispredicted classes in Lindel and ForeCast on both test sets. (**c, e.** Lindel on ForeCast test set and our test set; **d, f** ForeCasT on their test set and our test set). Each boxplot summarized the error of certain classes. The box represents 25^{th} percentile, 50^{th} percentile and 75^{th} percentile and whiskers represents 1.5x of the inter-quartile range (IQR). Small deletions around the cleavage site and 1 bp are difficult to predict accurately. **(g, h.)** Lindel **(g)** compared favorably to Microhomology Predictor (34) **h.** in predicting the ratio of frameshifting mutations for each of the 4,298 targets.

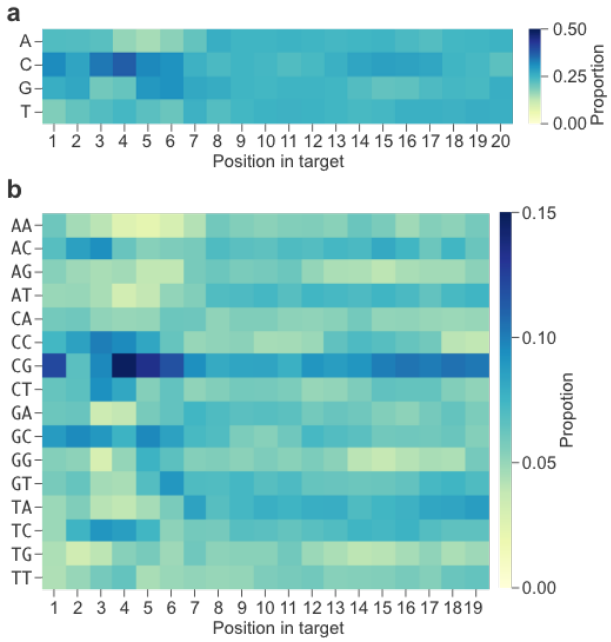

Figure 2.13: Sequence content of synthetic sgRNA-target library. a, b. Heatmaps of mononucleotide **a.** and dinucleotide **b.** balance within the final subsampled library of 6,872 well-represented CRISPR/Cas9 targets on which most analyses were performed. Each column sums to 1. Although initially designed sequences were balanced in mono/dinucleotide content, the overrepresentation of CG dinucleotides was likely introduced by how we screened these initial designs to remove on-target or off-target matches against the human genome (*i.e.* thereby subtly selecting in favor of designs containing CG dinucleotides, which are underrepresented in the human genome).

	Total targets tested	Endogenous/ genomic targets	Synthetic targets	Cell line(s)	Indels predicted
This study	6,872	0	6,872	HEK293T	536 classes of deletions (∼420 unique) 21 classes of insertions
Shen et al. (inDelphi)	1,872	0	1,872	HEK293, K562, HCT116, mESCs and U2OS	∼90 classes of MH deletion 59 classes of Non-MH deletion 4 classes of 1bp insertion
Allen et al. (ForeCasT)	41,630	6,654	27,906 unique + others	K562, RPE-1, iPSC, CHO, HAP1 and mESCs	∼420 classes of deletions (slightly varied by sequence) 20 classes of insertions
Chakrabarti et al.	1,492	1,492	0	HepG2	Not available

Table 2.1: Comparison of the design and results of different profiling data.

Name	Sequence	Usage
P1	5' AAGCTTGGCGTAACTAGATCTTGAGACAAA 3'	Backbone cloning
P2	5' ATTTACAACCGTCTCCGGTGTTTCG 3'	Backbone cloning
P3	5' TTGAGACATTGGTGGACGCGTCGTCTCAAAGCTTG GCGTAACTAGATC 3'	Backbone cloning
P4	5' ACGCGTCCACCAATGTCTCAAATTTACAACCGTCTC CGGTGTTTCG 3'	Backbone cloning
P5	5' GAGCAGCTCGTCTCTCACC 3	Oligo amp
P6	5' GCAAGCTTTGAGACGCATTG 3'	Oligo amp
P7	5' GCGTCAGATGTGTATAAGAGACAGNNNNNNNNNNNN NNNGGCTTTATATATCTTGTGGAAAGGACGAAACACCG 3'	UMI annealing
P8	5' GCGTCAGATGTGTATAAGAGACAG 3'	Genomic DNA amp up
P9	5' TTCAGACGTGTGCTCTTCCGATCTGTCTCAAGATCTA GTTACGCCAAGCTTTGAGACGC 3'	Genomic DNA amp up
P10	5' TTCAGACGTGTGCTCTTCCGATCTGTGGATGAATACT GCCATTTGTCTC 3'	Genomic DNA amp up
P11	5' AATGATACGGCGACCACCGAGATCTACACNNNNNNN NNNTCGTCGGCAGCGTCAGATGTGTATAAGAGACAG 3'	Sequencing adaptor Nextseq I5
P12	5' CAAGCAGAAGACGGCATACGAGATNNNNNNNNNNG TGACTGGAGTTCAGACGTGTGCTCTTCCGATCT 3'	Sequencing adaptor Trueseq I7

Table 2.2: Primers.

Sequence library	Reads	UMIs	UMI pass filter	Reads pass filter	UMI with designed sequence	Template switch	Other mutations	Final UMI count
Replicate 1	42,796,114	8,393,602	1,327,837	27,896,784 (65%)	1,005,606	27.15%	37.28%	357,671 (35.57%)
Replicate 2	60,898,224	12,094,298	1,783,500	39,107,335 (64%)	1,354,159	26.98%	37.27%	484,056 (35.75%)
Replicate 3	44,075,664	9,794,807	1,294,042	24,321,581 (55%)	982,804	27.92%	36.56%	349,066 (35.52%)
Total of Replicates 1-3	147,770,002							1,190,615
MH design	31,239,645	6,266,832	659,220	20,810,016 (67%)	445,440	8.05%	36.04%	249,039 (55.91%)

Table 2.3: Statistics of sequencing runs.

Chapter 3

PRIME-DEL: PRECISE GENOMIC DELETIONS USING PAIRED PRIME EDITING

This Chapter is adopted from published work with minimum changes

Choi, J.*, **Chen, W**.*, Suiter, C.C., Lee, C., Chardon, F.M., Yang, W., Leith, A., Daza, R.M., Martin, B. and Shendure, J., 2022. Precise genomic deletions using paired prime editing. Nature Biotechnology, 40(2), pp.218-226.

Author contribution: Junhong Choi designed and performed the experiments with the help from Wei Chen. Junhong Choi analyzed the data. Wei Chen created the design tool for paired pegRNA. Junhong Choi and Jay Shendure wrote the manuscript.

A back story about Prime-del: As I discussed in the back story in the chapter of ENGRAM, Prime-del is a co-effort between me and Junhong Choi. The idea actually stemmed from a discussion with another grad student Chase Suiter in the lab: what would happen if we use two pegRNA instead of pegRNA+sgRNA in the architecture of PE3. Junhong took the idea one step further by designing paired pegRNAs to introduce homology arms that are complementary to the 5 or 3 of the target of the other pegRNA. In this way, we are able to program the precise deletion of sequences between two pegRNA target sites. To facilitate the design process, I developed the algorithm and website for paired pegRNA design.

3.1 Abstract

Technologies that precisely delete genomic sequences in a programmed fashion can be used to study function as well as potentially for gene therapy. The leading contemporary method for programmed deletion uses CRISPR-Cas9 and pairs of guide RNAs (gRNAs) to generate two nearby double-strand breaks, which is often followed by deletion of the intervening sequence during DNA repair. However, this approach can be inefficient and imprecise, with errors including small indels at the two target sites as well as unintended large deletions and more complex rearrangements. Here we describe a prime editing-based method that we term *PRIME-Del*, which induces a deletion using a pair of prime editing gRNAs (pegRNAs) that target opposite DNA strands, effectively programming not only the sites that are nicked but also the outcome of the repair. We demonstrate that *PRIME-Del* achieves markedly higher precision than CRISPR-Cas9 and gRNA pairs in programming deletions up to 10 kb. We also show that *PRIME-Del* can be used to couple genomic deletions with short insertions, enabling deletions whose junctions do not fall at protospacer-adjacent motif (PAM) sites. Finally, we demonstrate that lengthening the time window of expression of prime editing components can substantially enhance efficiency without compromising precision. We anticipate that *PRIME-Del* will be broadly useful in enabling precise, flexible programming of genomic deletions, including in-frame deletions, as well as for epitope tagging and potentially for programming rearrangements.

3.2 Introduction

The ability to precisely manipulate the genome can critically enable investigations of the function of specific genomic sequences, including genes and regulatory elements. Within the past decade, CRISPR-Cas9-based technologies have proven transformative in this regard, allowing precise targeting of a genomic locus, with a quickly expanding repertoire of editing or perturbation modalities[73]. Among these, the precise and unrestricted deletion of specific genomic sequences is particularly important, with critical use cases in both functional

genomics and gene therapy.

Currently, the leading method for programming genomic deletions uses a pair of CRISPR guide RNAs (gRNAs) that each target a protospacer-adjacent motif (PAM) sequence, generating a pair of nearby DNA double-strand breaks (DSBs). Upon simultaneous cutting of two sites, cellular DNA damage repair factors often ligate two ends of the genome without the intervening sequence[74] through non-homologous end joining (NHEJ) (**Figure 3.1a**). Although powerful, this approach has several limitations: 1) An attempt to induce a deletion, particularly a longer deletion, often results in short insertions or deletions (indels; typically less than 10-bp) near one or both DSBs, with or without the intended deletion[75, 76, 77]; 2) Other unintended mutations including large deletions and more complex rearrangements can frequently occur, and go undetected for technical reasons[77, 78, 79, 80]; 3) DSBs are a cytotoxic insult[81]; and 4) The junctions of genomic deletions programmed by this method are limited by the distribution of naturally occurring PAM sites. Notwithstanding these limitations, various studies have employed this strategy to great effect, e.g. to investigate the function of genes and regulatory elements[82, 83, 77], as well as towards gene therapy[84, 85]. However, limited precision, DSB toxicity and the inability to program arbitrary deletions have handicapped the utility of CRISPR-Cas9-induced deletions in functional and therapeutic genomics.

Recently, Liu and colleagues described prime editing, which expands the CRISPR-Cas9 genome editing toolkit in critical ways[86]. Prime editing utilizes a Prime Editor-2 enzyme, which is a Cas9 nickase (Cas9 H840A) fused with a reverse-transcriptase, and a 3-extended gRNA (prime-editing gRNA or pegRNA). The Prime Editor-2 enzyme and pegRNA complex can nick one strand of the genome and attach a 3 single-stranded DNA flap to the nicked site following the template RNA sequence in the pegRNA molecule. By including homologous sequences to the neighboring region, DNA damage repair factors can incorporate the 3-flap sequence into the genome. The incorporation rate can be further enhanced using an additional gRNA, which makes a nick on the opposite strand, boosting DNA repair with the 3-flap sequence but often with a decrease in precision (strategy referred to as PE3/PE3b)[86]

(**Figure 3.1b**). The principal advantage of prime editing lies with its encoding of both the site to be targeted and the nature of the repair within a single molecule, the pegRNA. In addition to demonstrating many other classes of precise edits, Anzalone et al. used the PE3 strategy to show that a single pegRNA/gRNA pair could be used to program deletions ranging from 5 to 80 bp achieving high efficiency (52-78%) with modest precision (on average, 11% rate of unintended indels)[86]. However, even the PE3 strategy could face difficulties in programming deletions larger than 100 bp, as at least in plants, observed efficiencies fall precipitously for deletions larger than 20 bp[87].

We reasoned that a pair of pegRNAs could be used to specify not only the sites that are nicked but also the outcome of the repair, potentially enabling programming of deletions longer than 100 bp (**Figure 3.1c**). Here we demonstrate that this strategy, which we call *PRIME-Del*, induces the efficient deletion of sequences up to 10 kb in length with much higher precision than observed or expected with either the Cas9/paired-gRNA or PE3 strategies. We furthermore show that *PRIME-Del* can concurrently program short insertions at the deletion site. Concurrent deletion/insertion can be used to introduce in-frame deletions, to introduce epitope tags concurrently with deletions, and, more generally, to facilitate the programming of deletions unrestricted by the endogenous distribution of PAM sites. By filling these gaps, *PRIME-Del* expands our toolkit to investigate the biological function of genomic sequences at single nucleotide resolution.

3.3 Results

3.3.1 PRIME-Del induces precise deletions in episomal DNA

We first tested the feasibility of the *PRIME-Del* strategy by programming deletions to an episomally encoded eGFP gene. We designed pairs of pegRNAs specifying 24-, 91- and 546-bp deletions within the eGFP coding region of the pCMV-PE2-P2A-GFP plasmid (Addgene #132776) (**Figure 3.1d**). We cloned each pair of pegRNAs into a single plasmid with separate promoters, the human U6 and H1 sequences[77]. We transfected HEK293T cells with

eGFP-targeting paired-pegRNA and pCMV-PE2-P2A-GFP plasmids. We harvested DNA (including both genomic DNA and residual plasmid) from cells 4-5 days after transfection and PCR amplified the eGFP region. We then sequenced PCR amplicons to quantify the efficiency of the programmed deletion as well as to detect unintended edits to the targeted sequence.

We calculated deletion efficiency as the number of reads aligning to a reference sequence of the intended deletion, out of the total number of reads aligning to reference sequences either with or without the deletion. Estimated deletion efficiencies ranged from 38% (24-bp deletion) to 77% (546-bp deletion), and were consistent across replicates (note: throughout the paper, the term replicate is used to refer to independent transfections) (**Figure 3.1e**). This result clearly indicates that the *PRIME-Del* strategy outlined in Fig. 1c can work. However, we were initially concerned that these were overestimates of efficiency due to the shorter, edited templates being favored by both PCR and Illumina-based sequencing, particularly for the 546-bp deletion, because it has the largest difference between amplicon sizes (766-bp vs. 220-bp for wild-type and deletion amplicons, respectively). To address this, we repeated the amplification on DNA from the 546-bp deletion experiment with a two-step PCR, first adding 15 bp unique molecular identifiers (UMIs) via linear amplification before a second, exponential phase. The addition of UMIs via linear PCR was intended to minimize PCR and sequencing biases in our estimates of deletion efficiencies[88]. *PRIME-Del* efficiency was assessed based on the sequencing data after collapsing of reads with identical UMIs, as well as on the product size distribution (Agilent TapeStation). We observed a slight decrease in deletion efficiency after duplicate removal, from 73% to 66%, comparable to the 70% efficiency measured on the TapeStation (**Figure 3.1f**). These results suggest that our initial estimates of efficiency are only modestly impacted by size-dependent biases.

For most of these sequencing data, we had only a single read extending over the intended deletion site. As such, it was difficult to distinguish unintended editing outcomes (e.g. indels at the nick sites) from PCR or sequencing errors. To address this in part, we plotted frequencies of different classes of errors (substitutions, insertions, deletions) for sequences

aligning either to the unedited sequence (**Figure 3.1g, top**) or the intended deletion (**Figure 3.1g, bottom**), along the length of the sequencing read. For all replicates of the three deletion experiments (**Figure 3.6**), these profiles showed low rates of substitutions and indels, with nearly identical profiles and no consistent increase in the rate of any class of error at either the positions of the Prime Editor-2 enzyme nick sites or 3 flap ends above 1%, particularly after collapsing by UMI (**Figure 3.1g**, **Figure 3.6e**) or repeating sequencing with longer, paired-end sequencing reads (**Figure 3.1h**).

3.3.2 Simultaneous deletion and short insertion using PRIME-Del

We reasoned that because the homology sequences in the 3′-flaps program the deletion, we could potentially use *PRIME-Del* to concurrently introduce a short insertion at the deletion junction (**Figure 3.2a**). The desired insertion would be encoded into the pair of pegRNAs in a reverse complementary manner, just 5′ to the deletion-specifying homology sequences. With the conventional strategy for programming deletions, i.e. with Cas9 and paired gRNAs, the deletion junctions are determined by the gRNA targets, the selection of which is limited by the natural distribution of PAM sites (**Figure 3.2b**). Simultaneous deletion and short (less than 100 bps) insertion with *PRIME-Del* would offer at least three advantages over this conventional strategy. First, an arbitrary insertion of 1-3 bases could enable a reading frame to be maintained after editing, e.g. for deletions intended to remove a protein domain. Second, an arbitrary insertion could be used to effectively move one or both deletion junctions away from the cut-sites determined by the PAM, increasing flexibility to program deletions with base-pair precision. Third, insertion of functional sequences at the deletion junction could allow genome editing with *PRIME-Del* to be coupled to other experimental goals (e.g. protein tagging or insertion of a transcriptional start site).

To test this concept, we designed pegRNA pairs encoding five insertions ranging from 3 to 30 bp at the junction of a 546-bp programmed deletion within eGFP (**Figure 3.2c**). While our main objective was to test the effect of insertion length on deletion efficiency, we chose insertion sequences for their importance in molecular biology: The 3-bp insertion sequence

generates an in-frame stop codon. The 6-bp insertion sequence includes the start codon with the surrounding Kozak consensus sequence. The 12-bp insertion sequence includes tandem repeats of m6A post-transcriptional modification consensus sequence of GGACAT[89]. The 21-bp insertion sequence includes T7 RNA polymerase promoter sequence. The 30-bp insertion sequence encodes for the in-frame FLAG-tag peptide sequence when translated. The estimated efficiencies for simultaneous short insertion and deletion within the episomal eGFP gene in HEK293T cells were comparable to the 546-bp deletion alone, ranging from 83% to 90% for the various programmed insertions (**Figure 3.2d**). Also, insertion, deletion and substitution error rates at deletion junctions and across programmed insertions were comparable to the background error frequencies (**Figure 3.2e**, **Figure 3.7a**). As expected, the vast majority (>99%) of reads containing the programmed deletion also contained the insertion (**Figure 3.2f**), indicating that the full lengths of the pair of 3′-DNA flaps generated following the programmed pegRNA sequences specify the repair outcome (**Figure 3.2a**).

3.3.3 PRIME-Del induces precise deletions in genomic DNA

Encouraged by our initial results on editing episomal DNA, we next tested *PRIME-Del* on a copy of the eGFP gene integrated into the genome. We first generated the polyclonal HEK293T cells that carry the eGFP gene by lentiviral transduction, followed by flow-sorting to select GFP-positive cells (**Figure 3.3a**). We then tested the same pairs of pegRNAs encoding concurrent deletion and insertions (546-bp deletion with or without short insertions at the deletion junction) by transfecting pegRNAs and Prime Editor-2 enzyme without eGFP (pCMV-PE2; Addgene #132775) to these cells. Although editing efficiencies decreased substantially in comparison to episomal eGFP (7-17%; **Figure 3.3b**), we remained unable to detect errors that were clearly associated with editing (**Figure 3.3c**, **Figure 3.7b**). Specifically, there was no consistent pattern of error classes above background level accumulating at the nick-site or 3′-DNA-flap incorporation sites. Also, as previously, the vast majority of reads with the 546-bp deletion also contained programmed insertions (**Figure 3.7c**).

To test *PRIME-Del* on native genes, we designed two pairs of pegRNAs that respectively

specified 118 and 252-bp deletions within exon 1 of *HPRT1* (**Figure 3.3d**). We have previously performed a scanning deletion screen across the *HPRT1* locus using a Cas9/paired-gRNA strategy[77]. To directly compare *PRIME-Del* with Cas9/paired-gRNAs in programming genomic deletions, we attempted the same deletions with the same guides but substituting Prime Editor-2 enzyme with Cas9 in transfection of HEK293T cells. We quantified the resulting deletion efficiencies using two independent methods: First, we used the aforedescribed strategy of appending 15-bp unique molecular identifier (UMI) sequence via linear PCR step, before the standard PCR and sequencing readout. Resulting sequencing reads are collapsed by shared UMIs to minimize possible biases introduced in the PCR amplification and sequencing cluster generation steps. Second, we used droplet-digital PCR (ddPCR), which partitions genomic DNA into emulsion droplets before PCR amplification and fluorescence read-out of TaqMan probes within each droplet. We designed our probe to bind at the deletion junction, which would generate fluorescence signals specifically in the presence of the deletion. Our design of reporter probe aims to quantify the precise editing efficiencies, as errors introduced at the deletion junction are less likely to induce efficient binding of the probe during PCR[90]. Signals from deletions were normalized to the reference signal from detecting the copy-number of *RPP30* gene, which has been previously characterized and often used as a standard in ddPCR assay[90]. At exon 1 of *HPRT1*, we observed comparable deletion efficiencies for the *PRIME-Del* and Cas9/paired-gRNA strategies in HEK293T, ranging from 5% to 30% efficiencies for 118-bp and 252-bp deletions (**Figure 3.3e**). Of note, we observed consistently lower efficiencies with the ddPCR assay compared to the UMI-based sequencing assay. While this could be due to overestimation of efficiencies by the UMI-based approach, we also note that PCR amplification of the target region may be inefficient in the ddPCR assay based on the lack of clear separation of fluorescence intensities between positive and negative droplets (**Figure 3.8c,d**).

As is well established[75, 76, 77], the Cas9/paired-gRNA strategy often resulted in errors (mostly short deletions), whether with or without the intended deletion (**Figure 3.3f,g**; **Figure 3.8a**). Of reads lacking the intended 118-bp or 252-bp deletions, 12% or 12%

also contained an unintended indel at the observable target site, respectively (these are underestimates, because they only account for one of two target sites) (**Figure 3.3f, top**). Of reads containing the intended 118-bp or 252-bp deletions, 38% or 34% also contained an unintended indel at the deletion junction, respectively (**Figure 3.3f**, bottom). Such junctional errors are an established consequence of error-prone repair by NHEJ. In contrast, unintended indels were far less common with *PRIME-Del* (**Figure 3.3g**; **Figure 3.8b**). Of reads lacking the intended 118-bp or 252-bp deletions, 1.1% or 0.5% also contained an unintended short indel at the observable target site, respectively (**Figure 3.3g, top**). Of reads containing the intended 118-bp or 252-bp deletions, 12% or 2.7% also contained an unintended indel at the deletion junction, respectively (**Figure 3.3g, bottom**). The pattern of higher correct editing efficiencies for *PRIME-Del* over the Cas9/paired-gRNA strategy is also suggested by the ddPCR measurements, where the *PRIME-Del* reports a nearly 2-fold higher precisely edited population for both deletions.

For *PRIME-Del*, especially with the 118-bp deletion on *HPRT1*, the observation of an appreciable rate of insertions at the deletion junction in association with intended deletions (**Figure 3.3g, bottom**; **Figure 3.8b**) contrasts with our earlier observations at eGFP, where these rates were consistently equivalent to background. Further investigation of the error mode revealed that these errors corresponded to long insertions (mean 47-bp +/- 12-bp; **Figure 3.9**). The most frequent long insertion at the 118-bp deletion junction was 55-bp, a chimeric sequence between two 32-bp 3′-DNA flap sequences, overlapping at a GCCCT sequence, suggesting its origin from the annealing of GC-rich ends of 3′-DNA flaps. Similar chimeric sequences were observed as insertions at the 252-bp deletion junction, overlapping at GCCG within their 3′-DNA flaps. Nonetheless, even with these long insertions, 82% and 91% of all reads containing an indel matched the intended deletion exactly with *PRIME-Del*, but only 38% and 49% with the Cas9/paired-gRNA strategies (**Figure 3.3h**). Indel errors from the Cas9/paired-gRNA strategy are likely underestimated, because errors at only one of two Cas9 cut-sites are captured by our sequencing strategy.

The structure of the observed insertions and the lack of similar errors in applying *PRIME-*

Del to the eGFP locus suggested that this issue might be addressable through alternative pegRNA designs. As one approach, we either shortened or lengthened the RT template portion of both pegRNAs. For 118-bp deletion that used 32-bp RT template lengths for both pegRNAs, we shortened to either 17- and 25-bp long homology arms or lengthened to 42- and 46-bp long homology arms (**Figure 3.10a**). Both lengthening and shortening homology arms resulted in decreased deletion efficiencies (29% and 26% of the efficiencies observed with the standard designs for short and long homology arms, respectively) (**Figure 3.10b**). However, among deleted products, lengthening the homology arms also tended to decrease the long-insertion error frequency (to 30% of the standard design), while shortening the homology arms increased the insertion error frequency (to 129% of the standard design) (**Figure 3.10d**). Similar trends was observed with the 252-bp deletion, where shortening or lengthening homology arms decreased the deletion efficiency (**Figure 3.10c**), while lengthening the homology arm increased precision (**Figure 3.10e**). As a further control, substituting the sequence of the RT template to that used for programming a 546-bp deletion at eGFP failed to induce deletions for both 118-bp and 252-bp constructs targeting *HPRT1* (**Figure 3.10b,c**), fortifying the conclusion that *PRIME-Del* deletions are specific to DNA repair guided by the homology arm sequences.

We further applied genomic deletion using *PRIME-Del* at additional native loci, altogether testing 10 different deletions at 7 loci (**Figure 3.3h**). We performed all deletions in HEK293T cells, quantified deletion efficiencies and error frequencies using UMI-based sequencing assay, and directly compared *PRIME-Del* with the Cas9/paired-gRNA method (i.e. using the same guides but substituting in Cas9). Deletion sizes ranged from 118 bp at *HPRT1* exon 1 to 710 bp at e-NMU (enhancer for NMU gene) locus. In all 10 cases, we observe substantially lower error rates with *PRIME-Del* compared to the Cas9/paired-gRNA method. In five out of ten cases, we observe that the precise deletion is more efficient with *PRIME-Del* compared to the Cas9/paired-gRNA method, suggesting that higher precision does not compromise the deletion efficiencies in general. We did not observe a strong relationship between the deletion size and efficiency in this range (118 to 710 bps) for either

method.

Inversion of the sequence between two DSBs is a well-documented phenomenon when using the Cas9/paired-gRNA method[75, 91] (**Figure 3.11a**). To understand the frequency of inversion events using *PRIME-Del*, we aligned sequencing reads to a reference that was generated by inverting the sequence between two nick-sites. Across 10 deletions in 7 loci at which we performed *PRIME-Del*, we observed that virtually no reads aligned to the inverted reference (**Figure 3.11b**), while for Cas9/paired-gRNA controls, inversions were detected up in up to 2% of reads (**Figure 3.11b**).

To evaluate the length limits of *PRIME-Del*, we designed two additional deletions, sized 1,064 bps (1 kb) and 10,204 bps (10 kb) at the *HPRT1* locus. Since our sequencing-based assay is not well suited to detect amplicons greater than 1 kb, we used sequencing to quantify error frequencies in the deletion product alone, and ddPCR to measure the efficiency of precise deletion, again comparing Prime Editor-2 and Cas9 side-by-side. We observed that while deletion efficiencies between *PRIME-Del* and the Cas9/paired-gRNA method were comparable in HEK293T cells (**Figure 3.3i**), *PRIME-Del* achieves much higher precision, consistent with our observations while inducing shorter deletions. For the 1-kb deletion, both *PRIME-Del* and the Cas9/paired-gRNA method achieved nearly 3% deletion efficiency. For the 10-kb deletion, *PRIME-Del* and the Cas9/paired-gRNA method achieved 0.8% and 1.6% deletion efficiency, respectively. Upon sequencing amplicons derived from a PCR specific to the post-deletion junction, 98% and 97% of reads lacked indel errors at the junction with *PRIME-Del* for the 1-kb and 10-kb deletions, respectively, while only 47% and 42% of reads lacked indel errors with the Cas9/paired-gRNA strategy (**Figure 3.3j**).

To test whether the *PRIME-Del* can be multiplexed, we pooled plasmids encoding paired-pegRNAs programming four different but overlapping deletions (118, 252, 469 and 1064 bps) at the *HPRT1* locus, and transfected HEK293T cells with these together with a plasmid encoding the Prime Editor-2 enzyme. After incubating cells for 4 days and extracting genomic DNA, we used sequencing-based quantification to estimate 5.1%, 8.5% and 2.8% efficiencies for the 118-, 252-, and 469-bp deletions, and ddPCR to estimate 2% efficiency for the 1064-bp

deletion (**Figure 3.12**). Altogether, we estimate that 18% of *HPRT1* loci carry one of the four programmed deletions, which is comparable to the averaged efficiency of four deletions performed by transfecting a single construct of paired-pegRNA plasmid separately (12%). Our result suggests that *PRIME-Del* can be used to concurrently program multiple deletions by using pooled paired-pegRNA constructs similar to Cas9/paired-gRNA method[82, 83, 77].

3.3.4 Extending the editing time window enhances prime editing and PRIME-Del efficiency

In contrast with Cas9-mediated DSBs followed by NHEJ, both prime editing and *PRIME-Del* have high editing precision, producing an intended edit or conserving the original editable sequence. We reasoned that if the editing efficiencies of prime editing and *PRIME-Del* are limited by the transient availability of PE2/pegRNA molecules in the cell, extending Prime Editor-2 enzyme and pegRNA expression through stable genomic integration or, alternatively, repetitive transfection, would boost the rates of successful editing over time, particularly if uneditable dead ends outcomes are not concurrently accruing.

To facilitate prolonged expression, we generated monoclonal HEK293T and K562 cell lines expressing Prime Editor-2 enzyme (termed HEK293T(PE2) and K562(PE2), respectively). Because the Prime Editor-2 enzyme gene was larger than the lentiviral vectors typical limit, we cloned Prime Editor-2 enzyme into the piggyBAC cargo, transfected it along with the piggyBAC transposase, and identified a monoclonal cell line with active PE2. To continuously express pegRNAs in addition to Prime Editor-2 enzyme, we generated lentiviral vectors with pegRNAs and transduced them into both HEK293T(PE2) and K562(PE2) cells (**Figure 3.4a**). We tested two different deletions at *HPRT1* using *PRIME-Del* (the aforedescribed 118-bp and 252-bp deletions at exon 1), along with standard prime editing to insert 3-bp (CTT) into the synthetic HEK3 target sequence[86]. In K562(PE2), we observed a steady increase of the correctly edited population over time, both for CTT-insertion using prime editing and for 118- or 252-bp deletions using *PRIME-Del*. The end-point prime editing efficiencies for the CTT-insertion were very high, reaching 90% of targets with correct edits by 19 days after the first transduction of pegRNA into K562(PE2) cells (**Figure 3.4b**). The

rate of precise deletions using *PRIME-Del* also reached nearly 50% and 25% for the 118-bp and 252-bp deletions, respectively, by 19 days. In HEK293T(PE2) cells, we observed lower CTT-insertion efficiencies for the first 10 days, but eventually reaching 80-90% by day 19 (**Figure 3.4c**). Unexpectedly, we observed the near-absence of *PRIME-Del*-induced deletions in HEK293T(PE2) cells (**Figure 3.4c**). However, the same HEK293T(PE2) cell line showed modest increases in editing to 5 - 50% when we attempted multiple transfections of either PE2/pegRNA without additional stable integration or Prime Editor-2 enzyme alone after stable integration of piggyBAC-pegRNA, over four weeks (**Figure 3.13a,b**). We repeated the experiment with the same HEK293T(PE2) cell line but with a different batch of pegRNA-encoding lentivirus, which resulted in better efficiencies for the 118-bp deletion over time (**Figure 3.13c**), suggesting that *PRIME-Del* is more sensitive to both pegRNA and Prime Editor-2 enzyme expression level differences between transfection and lentiviral transduction than standard prime editing, presumably because the *PRIME-Del* requires two concurrent PE2 actions. Together, our results confirm that extended expression of prime editing or *PRIME-Del* components can boost efficiency.

3.3.5 Potential applications of PRIME-Del

Here we introduce *PRIME-Del*, a paired pegRNA strategy for prime editing, and demonstrate that it achieves high precision for programming deletions, both with or without short programmed insertions. We tested deletions ranging from 20 to 10,000-bp in length at episomal, synthetic genomic, and native genomic loci. The editing efficiency on native genes ranged from 1-30% with a single round of transient transfection in HEK293T cells, although we also observed that prolonged, high expression of prime editing or *PRIME-Del* components enhanced editing efficiency in K562 cells. For 12 deletions at seven genomic loci targeted with *PRIME-Del*, we observed high precision of editing except at *HPRT1* exon 1, where long insertions were sometimes observed at the deletion junction (5% of total reads). The GC-rich ends of 3′-DNA flap sequences of the pegRNA pairs used at *HPRT1* exon 1 appear to underlie the long insertions. Optimizing pegRNA design may be able to eliminate this error

mode, and we show that lengthening homology arms tends to decrease the frequency of long insertion errors. To facilitate avoidance of this particular error mode, we have developed an accompanying Python-based webtool for designing *PRIME-Del* paired-pegRNA sequences, which notifies the user if such sequences are present in designed pegRNA pairs.

However, even with these insertion errors, *PRIME-Del* consistently demonstrated higher precision than the Cas9/paired-gRNA strategy, i.e. for all 12 genomic deletions tested here, *PRIME-Del* resulted in fewer erroneous outcomes. For these same 12 cases, *PRIME-Del* exhibited markedly higher precise-deletion efficiencies for five (greater than a factor of two), comparable efficiencies for five (within a factor of two), and markedly lower efficiencies for two (less than half), compared to the Cas9/paired-gRNA method. Overall, these observations support the view that *PRIME-Del* achieves higher precision than the Cas9/paired-gRNA method without compromising editing efficiency.

A potential design-related limitation of *PRIME-Del* is that relative to the conventional Cas9/paired-gRNA strategy, it constrains the useable pairs of genomic protospacers, as they need to occur on opposing strands with the PAM sequences oriented towards one another (**Figure 3.1c**). However, the development and optimization of a near-PAMless[92] prime editing enzyme[93] would relax this constraint. A further limitation is that because of their longer length, cloning a pair of pegRNAs in tandem is more challenging than cloning gRNA pairs. Each pegRNA used here is 135 to 140 bp in length, such that synthesizing their unique components in tandem as a single, long oligonucleotide approaches the limits of conventional DNA synthesis technology, particularly for goals requiring array-based synthesis of paired pegRNA libraries.

Notwithstanding these limitations, *PRIME-Del* offers significant advantages over alternatives across several potential areas of application (**Figure 3.5**). Most straightforwardly, *PRIME-Del* can be used for precise programming of deletions up to 10 kb; we have yet to attempt deletions longer than 10 k**b.** In addition to the much lower indel error rate observed at the deletion junction compared to the Cas9/paired-gRNA strategy, inducing paired nicks is less likely to result in large, unintended deletions locally, rearrangements genome-

wide (chromothripsis)[94], or off-target editing[86, 95, 79, 96, 97]. These characteristics are advantageous for developing therapeutic approaches, e.g. where the *PRIME-Del* deletes pathogenic regions such as CGG-repeat expansions in 5-UTR of FMR1, without undesired perturbation of nearby or distant sequences[84, 85].

PRIME-Del also allows simultaneous insertion of short sequences at the programmed deletion junction without substantially compromising its efficiency or precision. Inserting short sequences allows for precise deletions of protein domains while preserving the native reading frame, i.e. avoiding a premature stop codon that might otherwise elicit a complex nonsense-mediated decay (NMD) response[98, 99]. Furthermore, inserting biologically active sequences upon deletion is likely to be advantageous in coupling *PRIME-Del* with technologies, i.e. by inserting epitope tags or T7 promoter sequences that can be used as molecular handles within edited genomic loci.

We also expect less toxicity via DNA damage by prime editing-based *PRIME-Del* than with the conventional Cas9/paired-gRNA strategy, which may facilitate multiplexing of programmed genomic deletions for frameworks such as scanDel and crisprQTL[78, 77]. For studying the non-coding elements in transcription, efficient and precise deletions up to 10 kb complements the current use of deactivated Cas9-tethered KRAB domain for CRISPR-interference (CRISPRi), which cannot control the range of epigenetic modifications around target regions. As such, we anticipate that *PRIME-Del* could be broadly applied in massively parallel functional assays to characterize native genetic elements at base-pair resolution.

3.4 Materials and Methods

3.4.1 pegRNA/gRNA design

For pegRNA/gRNA design, we initially used CRISPOR[100] to select for 20-bp CRISPR-Cas9 spacers within a given region of interest. We avoided spacers annotated as inefficient, including U6/H1 terminator and GC-rich sequences, and generally selected spacers that had higher predicted efficiencies (Doench scores for U6 transcribed gRNAs[101]). The length of

the RT-template portion of a pegRNA was initially set to 30-bp and extended by 1 to 2-bp if it ended in G or C[86, 102].

3.4.2 Web tool for PRIME-Del paired-pegRNA design

To facilitate *PRIME-Del* paired-pegRNA design, we developed a Python-based web tool that automates the design process. The software takes a FASTA-formatted sequence file as the input, identifies all possible PAM sequences within the provided region, and initially generates all potential paired pegRNA sequences to program deletions. The software can also optionally take as input scored gRNA files generated using Flashfry[103], CRISPOR[100] or GPP sgRNA designer[100]; this is highly recommended to identify effective CRISPR-Cas9 spacers. For FlashFry and CRISPOR, gRNA spacers with MIT specificity scores[104] below 50 are filtered out as recommended by CRISPOR. From initially generated pegRNA pairs, the software selects relevant ones based on additional user-provided design parameters. For example, the user can define the deletion size range. The user can also define the start and end position of desired deletion, and the software will filter to pegRNA pairs present windows centered at those coordinates. pegRNAs for deletions whose junctions do not fall at PAM sites can be designed using the option –precise (-p), which adds insertion sequences to both pegRNAs to facilitate the desired edit.

The *PRIME-Del* design software also enables additional design constraints to be specified. The pegRNA RT-template length (also known as the homology arm) is set to 30-bp by default, unless specified otherwise by the user. The pegRNA PBS length is set to 13-bp from the PE2 nick-site by default, unless specified otherwise by the user. The nick position relative to the PAM sequence is predicted using previously identified parameters (Lindel[24]), and RT-template length is adjusted accordingly if the predicted likelihood of generating a nick at a non-canonical position is greater than 25%. PegRNA sequences that include RNA polymerase III terminator sequences (more than four consecutive Ts) are filtered out. The software generates warning messages if more than 4 out of 5 bp in either 3-DNA-flap are either G or C. Code is available at https://github.com/shendurelab/*PRIME-Del*, and interactive

webpage is available at https://primedel.uc.r.appspot.com/.

3.4.3 pegRNA cloning

After designing pegRNA pairs, we followed the Golden-Gate cloning strategy outlined by Anzalone et al.[86], assembling three dsDNA fragments and one plasmid backbone. The first dsDNA fragment contains the pegRNA-1 spacer sequence, annealed from two complementary synthetic single-strand DNA oligonucleotides (IDT) with 4-bp 5′-overhangs. The second dsDNA fragment contains the pegRNA-1 gRNA scaffold sequence, annealed from two DNA oligonucleotides with 5′-end phosphorylation at the end of 4-bp overhang. The third dsDNA fragment contains the pegRNA-1 RT template sequence and primer binding sequence (PBS), pegRNA-1 terminator sequence (six consecutive Ts), and pegRNA-2 sequence with H1 promoter sequence. This was generated by appending pegRNA-1 portion and pegRNA-2 portion to two ends of gene fragments (purchased as gBlocks from IDT) by PCR amplification. The gene fragments contained the pegRNA-1 terminator sequence, H1 promoter sequence, pegRNA-2 spacer sequence, and pegRNA-2 gRNA scaffold sequences. The forward primer included the BsmBI or BsaI restriction site, pegRNA-1 RT template sequence and PBS. The reverse primer included pegRNA-2 RT template, PBS, and BsmBI or BsaI restriction site. PCR fragments (sized between 300 and 400 bp) were purified using 1.0X AMPure (Beckman Coulter) and mixed with two other dsDNA fragments and linearized backbone vector with corresponding overhangs for Golden-Gate-based assembly mix (BsmBI or BsaI golden-gate assembly mix from New England Biolabs). For the pegRNA cloning backbone, we used either the GG-acceptor plasmid (Addgene #132777) or piggyBAC-cargo vector that carries the blasticidin-resistance gene. Each construct plasmid was transformed into Stbl Competent E. coli (NEB C3040H) for amplification and purified using a miniprep kit (Qiagen). Cloning was verified using Sanger sequencing (Genewiz).

3.4.4 Tissue culture, transfection, lentiviral transduction, and monoclonal line generation

HEK293T and K562 cells were purchased from ATCC. HEK293T cells were cultured in Dulbeccos modified Eagles medium with high glucose (GIBCO), supplemented with 10% fetal bovine serum (Rocky Mountain Biologicals) and 1% penicillin-streptomycin (GIBCO). K562 cells were cultured in RPMI 1640 with L-Glutamine (Gibco), supplemented with 10% fetal bovine serum (Rocky Mountain Biologicals) and 1% penicillin-streptomycin (GIBCO). HEK293T and K562 cells were grown with 5% CO_2 at 37 $°C$.

For transient transfection, about 50,000 cells were seeded to each well in a 24-well plate and cultured to 70-90% confluency. For prime editing, 375 ng of Prime Editor-2 enzyme plasmid (Addgene #132775) and 125 ng of pegRNA or paired-pegRNA plasmid were mixed and prepared with transfection reagent (Lipofectamine 3000) following the recommended protocol from the vendor. For deletion using Cas9/paired-gRNA, 375 ng of Cas9 plasmid (Addgene #52962) was used instead of Prime Editor-2 enzyme plasmid. Cells were cultured for four to five days after the initial transfection unless noted otherwise, and its genomic DNA was harvested either using DNeasy Blood and Tissue kit (Qiagen) or following cell lysis and protease protocol from Anzalone et al.[86].

For lentiviral generation, about 300,000 cells were seeded to each well in a 6-well plate and cultured to 70-90% confluency. Lentiviral plasmid was transfected along with the ViraPower lentiviral expression system (ThermoFisher) following the recommended protocol from the vendor. Lentivirus was harvested following the same protocol, concentrated overnight using Peg-it Virus Precipitation Solution (SBI), and used within 1-2 days to transduce either K562 or HEK293T cells without a freeze-thaw cycle.

For transposase integration, 500 ng of cargo plasmid and 100 ng of Super piggyBAC transposase expression vector (SBI) were mixed and prepared with transfection reagent (Lipofectamine 3000) following the recommended protocol from the vendor. Prime Editor-2 enzyme-expressing single-cell clones were generated by integrating PE2 using piggyBAC transposase system, selected by marker (puromycin resistance gene), single-cell sorted into 96-well plates

using flow-sort apparatus, cultured for 2-3 weeks until confluency, and screened for PE activity by transfecting CTT-inserting pegRNA alone (Addgene #132778) and sequencing the HEK3-target loci.

3.4.5 DNA sequencing library preparation

To quantify programmed deletion efficiency and possible errors generated by *PRIME-Del*, we amplified the targeted region from purified DNA (200 to 1000 bp in length) using two-step PCR and sequenced using Illumina sequencing platform (NextSeq or MiSeq) (**Figure 3.6 a**). Each purified DNA sample contains wild-type and edited DNA molecules, which were amplified together using the same pairs of primers through each PCR reaction. For the PCR-amplification, we designed a pair of primers for each genomic locus (amplicon) where entire amplicon sizes, with or without deletion, were greater than 200 bp to avoid potential problems in PCR-amplification, in purifying of PCR products, and in clustering onto the sequencing flow-cell.

The first PCR reaction (KAPA Robust) included 300 ng of purified genomic DNA or 2 μL of cell lysate, 0.04 to 0.4 μM of forward and reverse primers in a final reaction volume of 50 μL. We programmed the first PCR reaction to be: 1) 3 minutes at $95°C$, 2) 15 seconds at $95°C$, 3) 10 seconds at $65°C$, 4) 45 seconds at $72°C$, 25-28 cycles of repeating step 2 through 4, and 5) 1 minute at $72°C$. Primers included sequencing adapters to their 3-ends, appending them to both termini of PCR products that amplified genomic DNA. After the first PCR step, products were assessed on 6% TBE-gel and purified using 1.0X AMPure (Beckman Coulter) and added to the second PCR reaction that appended dual sample indexes and flow cell adapters. The second PCR reaction program was identical to the first PCR program except we run 5-10 cycles. Products were again purified using AMPure and assessed on the TapeStation (Agilent) before denatured for the sequencing run. For long deletions that generate amplicons sized 200 to 300 bp, we used Miseq sequencing platform at low (8 pM) input DNA concentration to minimize the short amplicons replacing the long

amplicons during clustering, aiming cluster density of 300-400 k/mm^2. Denatured libraries were sequenced using either Illumina NextSeq or MiSeq instruments following the vendor protocols.

For appending 15-bp unique molecular identifiers (UMI), we performed the first PCR reaction in two-steps: First, genomic DNA was linearly amplified in the presence of 0.04 to 0.4 μM of single forward primer in two PCR cycles using KAPA Robust polymerase. We programmed the UMI-appending linear PCR reaction to be: 1) 3 minutes and 15 seconds at $95°C$, 2) 1 minute at $65°C$, 3) 2 minutes at $72°C$, 5 cycles of repeating step 2 and 3, 4) 15 seconds at $95°C$, 5) 1 minute at $65°C$, 6) 2 minutes at $72°C$, and another 5 cycles of repeating step 5 and 6. This reaction was cleaned up using 1.5X AMPure, and subject to the second PCR with forward and reverse primers. In this case, the forward primer anneals to the upstream of UMI sequence and is not specific to the genomic loci. After PCR amplification, products were cleaned up and added to another PCR reaction that appended dual sample indexes and flow cell adapters, similar to other samples.

3.4.6 Sequencing data processing and analysis

We designed the sequencing layout to cover at least 50-bp away from the deletion junction in each direction (**Figure 3.6 a**). In case of the paired-end sequencing, PEAR[105] was used to merge the paired-end reads with default parameters and -e flag to disable the empirical base frequencies. When 15-bp UMI was present in the sequencing reads, we used a custom Python script to find all reads that share the same UMI, and collapsed into a single read with the most frequent sequence. The resulting sequencing reads were aligned to two reference sequences (with or without deletion) generally using the CRISPResso2 software[106]. Default alignment parameters were used in CRISPResso2, with the gap-open penalty of -20, the gap-extension penalty of -2, and the gap incentive value of 1 for inserting indels at the cut/nick sites. The minimum homology score for a read alignment was explored between 50 and 95 for different amplicon length. Custom python and R scripts were used to analyze the alignment results from CRISPResso2.

3.4.7 Droplet digital PCR (ddPCR) assay

Alignment was done using two reference sequences (wild-type and deletion) of same sequence length, generating two sets of reads with respective reference sequences. Deletion efficiencies were calculated as the fraction of total number of reads aligning to the reference sequence with deletion over the total number of reads aligning to either references. Genome editing has three types of error modes: substitution, insertion, and deletion. Each error frequency was plotted across two reference sequences, highlighting in each such plot the Cas9(H840A) nick-site and the 3′-DNA flap incorporation sites.

We designed ddPCR probes following the recommended parameters by Bio-Rad Laboratories. We purchased pre-mixed reference probes and primers for the *RPP30* gene from Bio-Rad Laboratories. Probes and PCR primers were purchased from Integrated DNA Technologies (IDT). Probes were modified with FAM on their 5′-ends and included double quenchers (IDT PrimeTime qPCR probes). Probe sequences were specifically designed to cover the deletion junction for detecting precise deletion products[90]. For detecting each deletion, we prepared a 20X primer mix composed of 18 μM forward-primer, 18 μM reverse-primer, and 5 μM FAM-labeled probe in 50 mM Tris-HCl buffer (pH 8.0 at room temperature). 25 μL of ddPCR reaction mixes were composed of 12.5 μL of 2X Supermix for Probes (no dUTP) (Bio-Rad Laboratories), 1.25 μL of 20X HEX-modified *RPP30* reference mix (Bio-Rad Laboratories), 1.25 μL of 20X FAM-modified primer mix, 0.5 μL of cell lysate containing genomic DNA, and 9.5 μL of DNAse-free water. We added 20 μL of ddPCR reaction mix to 70 μL of Droplet generation oil for probes and used QX200 Droplet generator (Bio-Rad Laboratories) to generate droplets. Droplets were transferred to ddPCR 96-well plates (Bio-Rad Laboratories) and run on 96-well thermocyclers (Eppendorf) with the following program: 1) 10 minutes at $95°C$, 2) 30 seconds at $94°C$, 3) 1 minute at $50°C$, 41 cycles of repeating step 2 and 3, 4) 10 minutes on $98°C$, and 5) cooled down to $4°C$ before loading to QX200 Droplet reader. Temperature ramps were limited to $1°C$ per second on all steps on thermocyclers. We used QX200 Droplet reader and Bio-Rad QuantaSoft Pro

software to visualize and analyze ddPCR experiments. The deletion efficiencies were taken from the ratio of FAM+ (precise-deletion) over HEX+ (*RPP30* reference for genomic DNA loading) events.

3.5 *Figures*

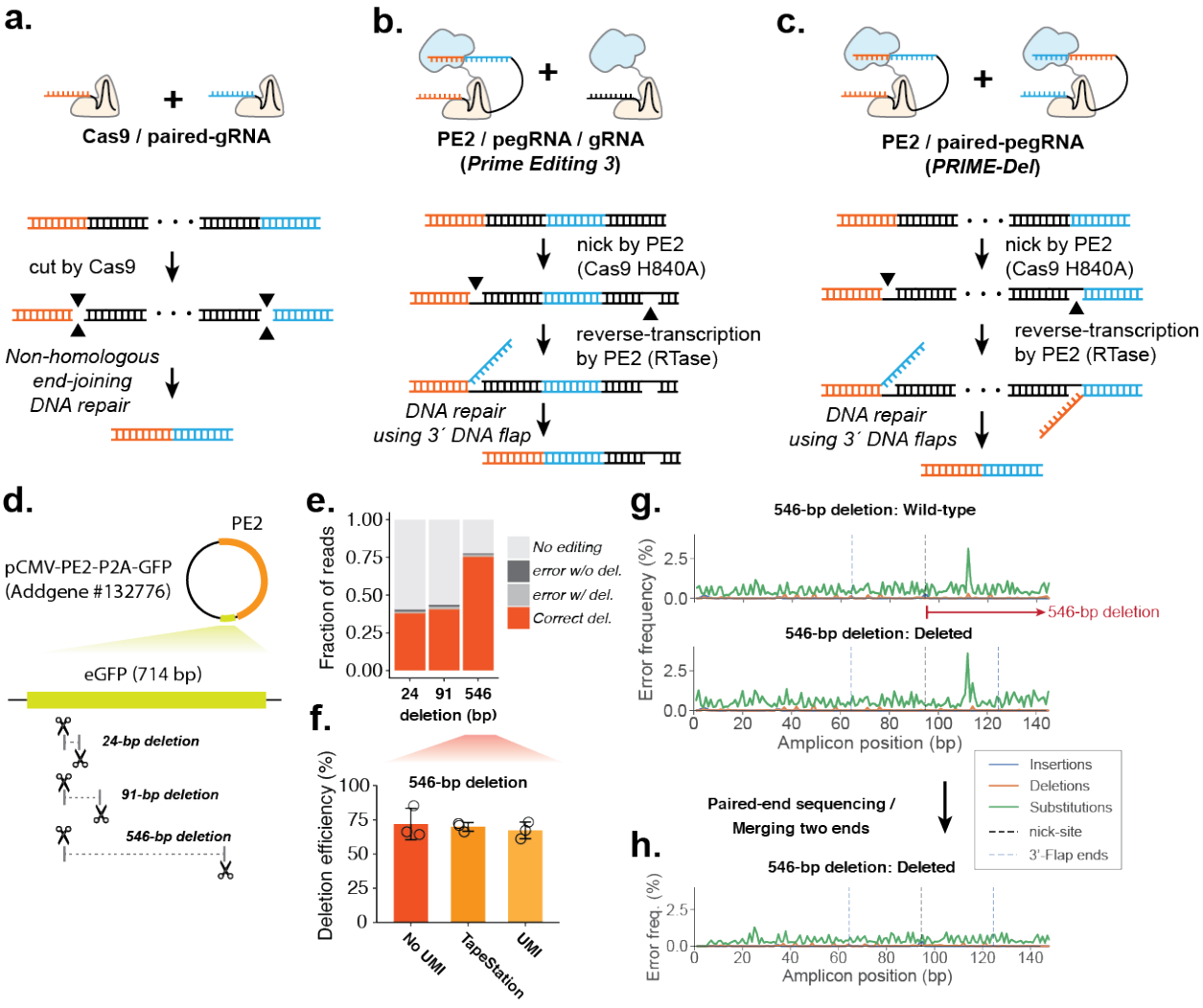
a.
Cas9 / paired-gRNA
cut by Cas9
Non-homologous
end-joining
DNA repair
b.
PE2 / pegRNA / gRNA
(Prime Editing 3)
nick by PE2
(Cas9 H840A)
reverse-transcription
by PE2 (RTase)
DNA repair
using 3´ DNA flap
c.
PE2 / paired-pegRNA
(PRIME-Del)
nick by PE2
(Cas9 H840A)
reverse-transcription
by PE2 (RTase)
DNA repair
using 3´ DNA flaps
d.
PE2
pCMV-PE2-P2A-GFP
(Addgene #132776)
eGFP (714 bp)
24-bp deletion
91-bp deletion
546-bp deletion
e.
Fraction of reads
No editing
error w/o del.
error w/ del.
Correct del.
24
91
546
deletion (bp)
f.
546-bp deletion
Deletion efficiency (%)
No UMI
TapeStation
UMI
g.
546-bp deletion: Wild-type
546-bp deletion
546-bp deletion: Deleted
Error frequency (%)
Amplicon position (bp)
Paired-end sequencing /
Merging two ends
Insertions
Deletions
Substitutions
nick-site
3'-Flap ends
h.
546-bp deletion: Deleted
Error freq. (%)
Amplicon position (bp)

Figure 3.1: Precise episomal deletions using *PRIME-Del*. **a.** Schematic of Cas9/paired-gRNA deletion strategy. **b.** Schematic of PE3 strategy, wherein the Prime Editor-2 enzyme and gRNA complex induce a nick (denoted as a gap in the bottom DNA strand), even after the correct editing event. **c.** Schematic of *PRIME-Del* using pairs of pegRNAs that target opposite DNA strands. Each pegRNA encodes the sites to be nicked at each end of the intended deletion, as well as a 3 flap that is complementary to the region targeted by the other pegRNA. **d.** Cartoon representation of deletions programmed within the episomally-encoded eGFP gene (not drawn to a scale). **e.** *PRIME-Del*-mediated deletion efficiencies and error frequencies (with or without intended deletion) were measured for 24-bp, 91-bp, and 546-bp deletion experiments in HEK293T cells (averaged over replicates; n = 5). Sequencing reads were classified as without indel modifications (No editing), indel errors without the intended deletion, indel errors with the intended deletion, and correct deletion without error. **f.** *PRIME-Del*-mediated deletion efficiency was measured for the 546-bp deletion experiment using three methods. Error bars represent standard deviation for three replicates. **g.** Insertion, deletion and substitution error frequencies across sequencing reads from 546-bp deletion experiment. Reads were aligned to reference sequence either without (top) or with (bottom) deletion. Plots are from single-end reads with collapsing of UMIs to reduce sequencing errors; also shown with additional replicates and error-class-specific scales in **Figure 3.6 e**. Note that only one of the two 3-DNA-flaps is covered by the sequencing read in amplicons lacking the deletion (labeled as wild-type). **h.** Insertion, deletion and substitution error frequencies across the amplicons from 546-bp deletion experiment after merging paired-end sequencing reads.

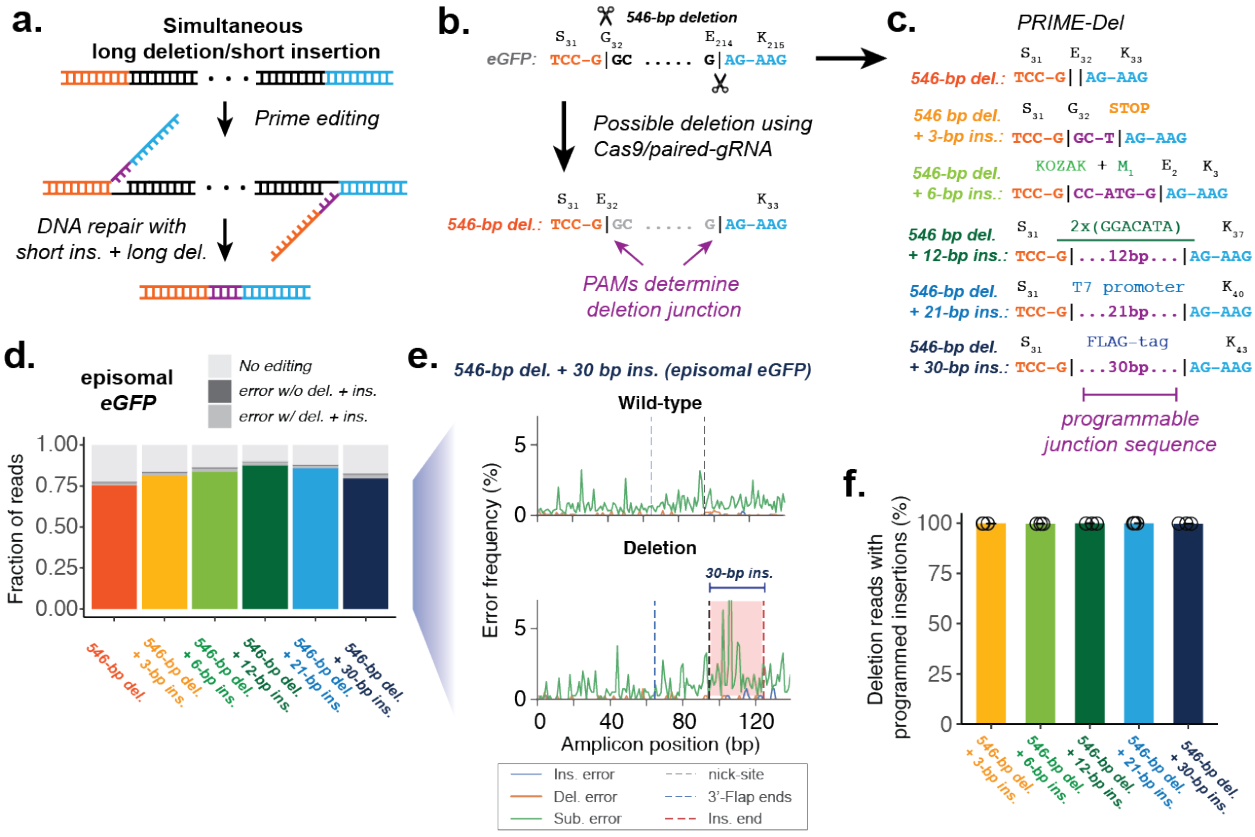

Figure 3.2: Concurrent programming of deletion and insertion using *PRIME-Del*. **a.** Schematic of strategy, with reverse complementary sequences corresponding to the intended insertion in purple. b. Conventional strategy for deletion with Cas9 and pairs of gRNAs. Potential deletion junctions are restricted by the natural distribution of PAM sites. **c.** Pairs of pegRNAs were designed to encode five insertions, ranging in size from 3 to 30 bp, together with a 546 bp deletion in eGFP. **d.** Estimated deletion efficiencies and indel error frequencies (with or without intended deletion) in using these pegRNA pairs to induce concurrent deletion and insertion in HEK293T cells (averaged over replicates; n = 3). **e.** Representative insertion, deletion and substitution error frequencies plotted across sequencing reads from concurrent 546-bp deletion and 30-bp insertion condition. Plots are from single-end reads without UMI correction. Note that only one of the two 3′-DNA-flaps is covered by the sequencing read in amplicons lacking the deletion (labeled as wild-type). **f.** The percentage of reads containing the programmed deletion that also contain the programmed insertion. Error bars represent standard deviation for at least three replicates.

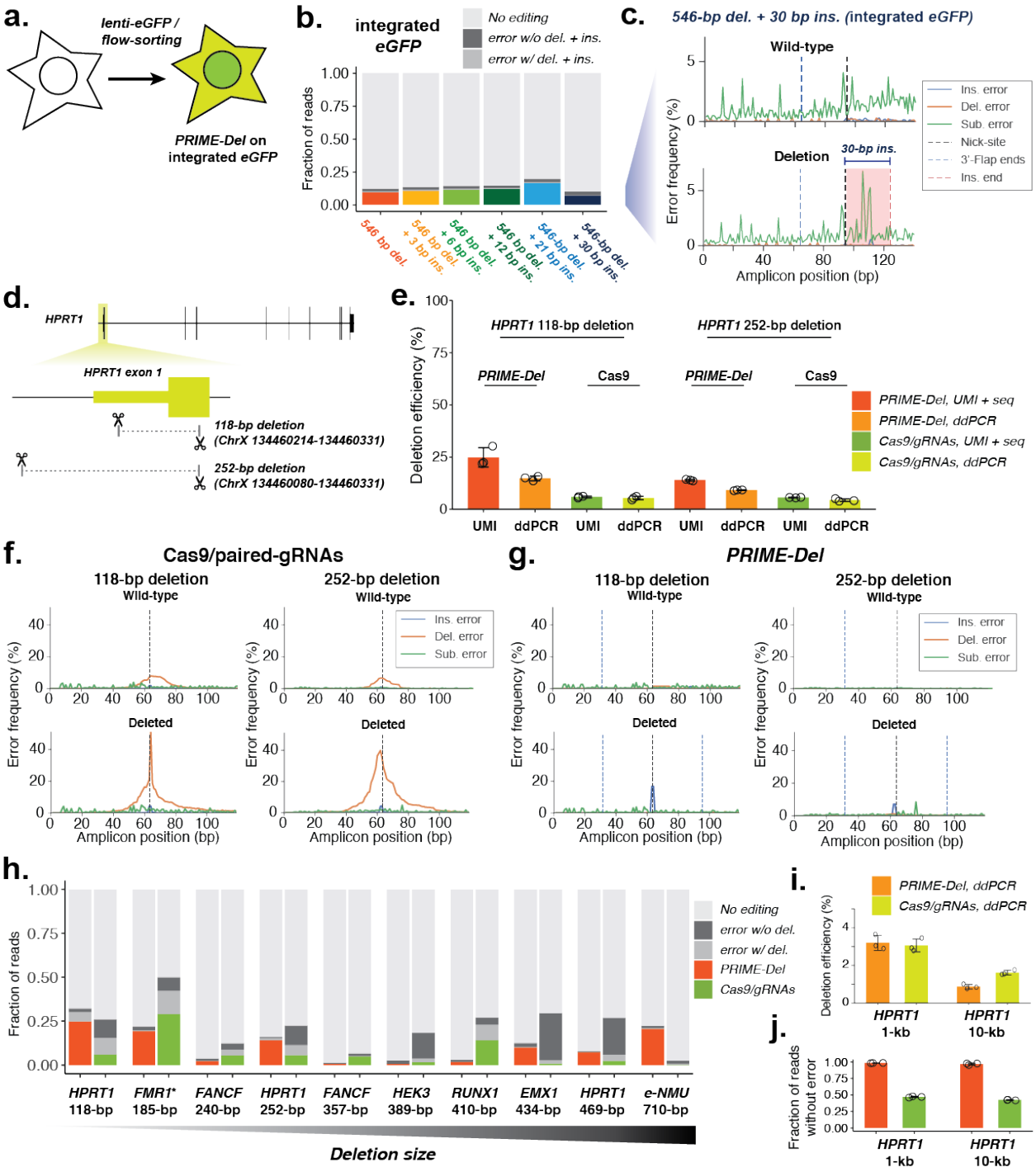
a.
lenti-eGFP / flow-sorting
PRIME-Del on integrated eGFP
b.
integrated eGFP
No editing
error w/o del. + ins.
error w/ del. + ins.
Fraction of reads
546 bp del.
546 bp del. + 3 bp ins.
546 bp del. + 6 bp ins.
546 bp del. + 12 bp ins.
546 bp del. + 21 bp ins.
546 bp del. + 30 bp ins.
c.
546-bp del. + 30 bp ins. (integrated eGFP)
Wild-type
Deletion
30-bp ins.
Error frequency (%)
Amplicon position (bp)
Ins. error
Del. error
Sub. error
Nick-site
3'-Flap ends
Ins. end
d.
HPRT1
HPRT1 exon 1
118-bp deletion (ChrX 134460214-134460331)
252-bp deletion (ChrX 134460080-134460331)
e.
HPRT1 118-bp deletion
HPRT1 252-bp deletion
PRIME-Del
Cas9
Deletion efficiency (%)
PRIME-Del, UMI + seq
PRIME-Del, ddPCR
Cas9/gRNAs, UMI + seq
Cas9/gRNAs, ddPCR
UMI
ddPCR
f.
Cas9/paired-gRNAs
118-bp deletion
252-bp deletion
Wild-type
Deleted
Error frequency (%)
Amplicon position (bp)
g.
PRIME-Del
h.
Fraction of reads
No editing
error w/o del.
error w/ del.
PRIME-Del
Cas9/gRNAs
HPRT1 118-bp
FMR1* 185-bp
FANCF 240-bp
HPRT1 252-bp
FANCF 357-bp
HEK3 389-bp
RUNX1 410-bp
EMX1 434-bp
HPRT1 469-bp
e-NMU 710-bp
Deletion size
i.
PRIME-Del, ddPCR
Cas9/gRNAs, ddPCR
Deletion efficiency (%)
HPRT1 1-kb
HPRT1 10-kb
j.
Fraction of reads without error
HPRT1 1-kb
HPRT1 10-kb

Figure 3.3: Precise genomic deletions using *PRIME-Del*. **a.** Schematic of generation of the eGFP-integrated HEK293T cell line. **b.** Estimated deletion efficiencies and error frequencies in using *PRIME-Del* for concurrent deletion and insertion on genomically integrated eGFP in HEK293T cells (n = 3). **c.** Representative insertion, deletion and substitution error frequencies plotted across sequencing reads from concurrent 546-bp deletion and 30-bp insertion condition on genomically integrated eGFP. Plots are from single-end reads without UMI correction. **d.** Cartoon representation of deletions programmed within the *HPRT1* gene. **e.** Deletion efficiencies measured for the 118-bp and 252-bp deletion using either *PRIME-Del* or Cas9/paired-gRNA (abbreviated to Cas9) strategies in HEK293T cells, quantified using either the unique-molecular identifier-based sequencing assay (UMI) or the droplet-digital PCR (ddPCR) assay. Error bars represent standard deviation for three replicates. **f.** Representative insertion, deletion and substitution error frequencies plotted across sequencing reads from 118-bp deletion (left) and 252-bp deletion (right) at HPRT exon 1, using the Cas9/paired-gRNA strategy. Different error classes are colored the same as in **(c)**. **g.** Same as **(f)**, but for *PRIME-Del* strategy. **h.** Estimated deletion efficiencies and indel error frequencies for different deletions across the genome for *PRIME-Del* (left) and Cas9/paired-gRNA (right) methods (averaged over replicates; n = 3). UMI-based sequencing assay was used for quantification (except the GC-rich amplicon of *FMR1**, where added DMSO interfered with the UMI-addition reaction). **i.** Deletion efficiencies measured for 1-kb and 10-kb deletions at *HPRT1* using either *PRIME-Del* (left) or Cas9/paired-gRNA (right) with ddPCR-based assay in HEK293T cells. Error bars represent standard deviation for three replicates. **j.** Fraction of reads with precise deletion measured for the 1-kb and 10-kb deletion on *HPRT1* gene with either *PRIME-Del* (left) or Cas9/paired-gRNA (right) using sequencing of the deletion amplicons. Error bars represent standard deviation for three replicates.

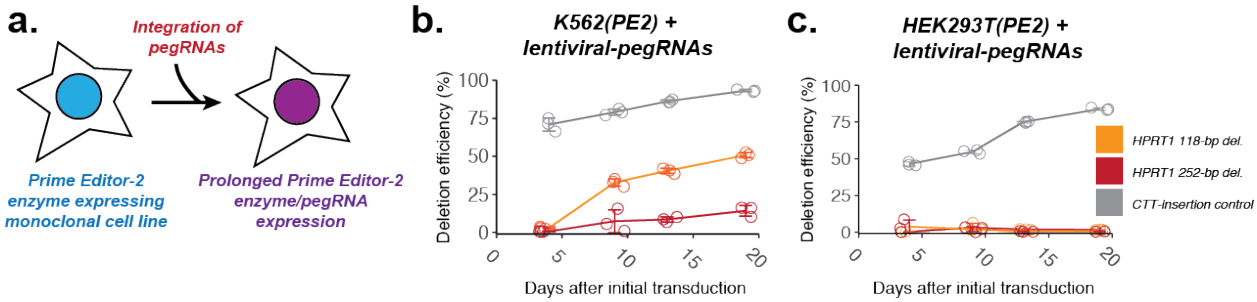

Figure 3.4: Extending the editing time window enhances prime editing and *PRIME-Del* efficiency. a. Schematic for stably expressing both Prime Editor-2 enzyme and pegRNAs via two-step genome integration. **b-c.** Editing efficiencies measured for the 118-bp and 252-bp deletions at genomic *HPRT1* exon 1 using *PRIME-Del* (paired-pegRNA construct) or CTT-insertion using prime editing (single-pegRNA construct) in K562(PE2) cells **(b)** or HEK293T(PE2) cells **(c)**, as a function of time after initial transduction of pegRNA(s). Error bars represent standard deviation for three replicates.

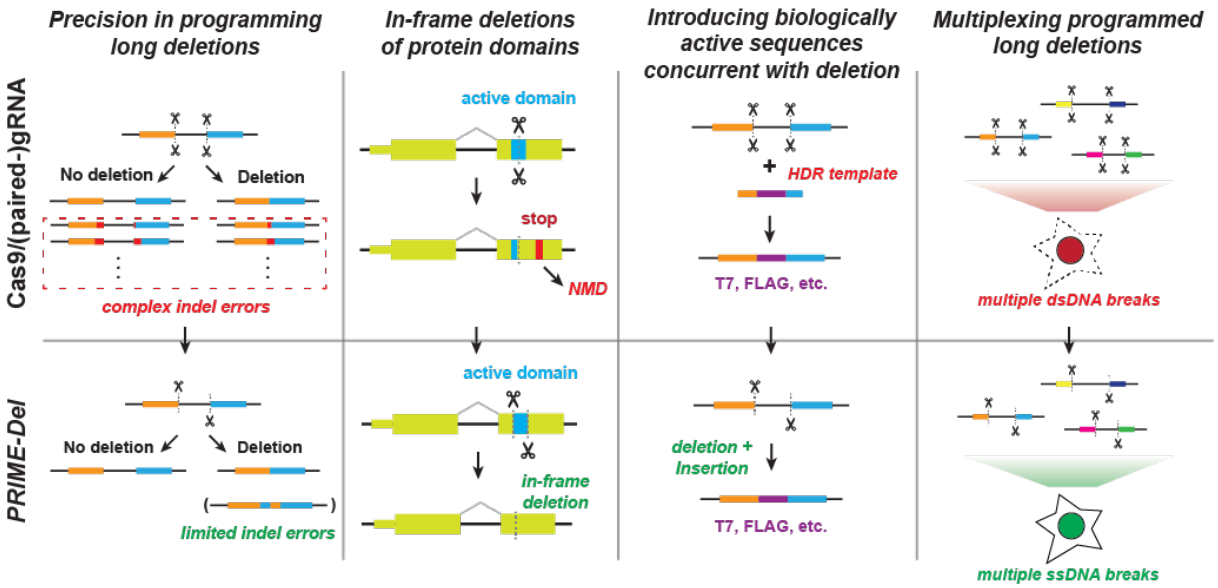

Figure 3.5: Potential advantages of using PRIME-Del in various genome editing applications. The *PRIME-Del* strategy can be used to program precise genomic deletions without generation of short indel errors at Cas9 target sequences. Precision deletion, combined with ability to insert a short arbitrary sequence at the deletion junction, may allow robust gene knockout of active protein domains without generating a premature in-frame stop codon, which can trigger the nonsense-mediated decay (NMD) pathway. *PRIME-Del* may also allow replacement of genomic regions up to 10 Kb base-pairs with arbitrary sequences such as epitope tags or RNA transcription start sites. Single-stranded breaks generated during *PRIME-Del* are likely to be less toxic to the cell, especially when multiple regions are edited in parallel, potentially facilitating its multiplexing.

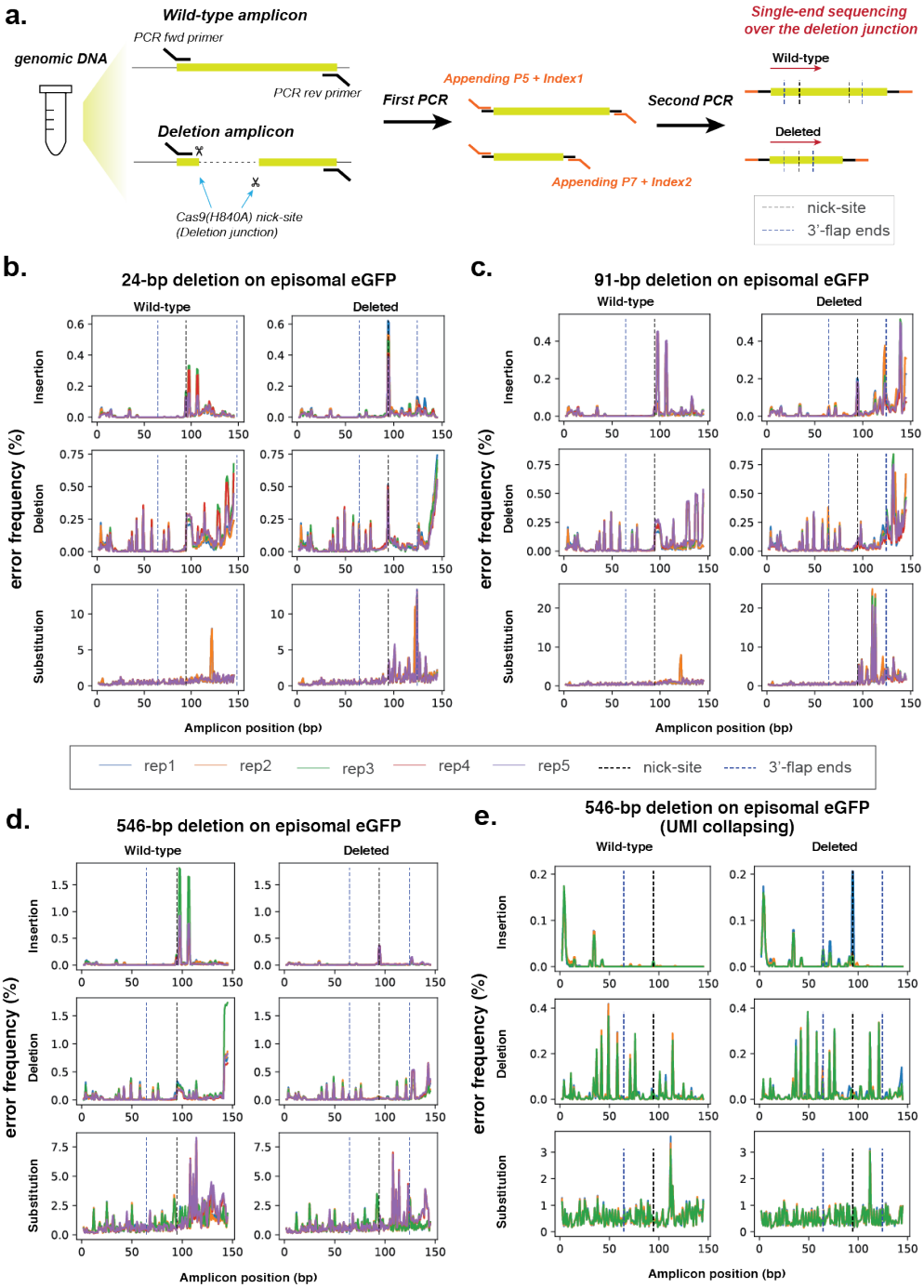
a.
Wild-type amplicon
PCR fwd primer
genomic DNA
PCR rev primer
Deletion amplicon
Cas9(H840A) nick-site
(Deletion junction)
First PCR
Appending P5 + Index1
Appending P7 + Index2
Second PCR
Single-end sequencing
over the deletion junction
Wild-type
Deleted
nick-site
3'-flap ends
b.
24-bp deletion on episomal eGFP
c.
91-bp deletion on episomal eGFP
Wild-type
Deleted
Insertion
Deletion
Substitution
error frequency (%)
Amplicon position (bp)
rep1
rep2
rep3
rep4
rep5
nick-site
3'-flap ends
d.
546-bp deletion on episomal eGFP
e.
546-bp deletion on episomal eGFP
(UMI collapsing)

Figure 3.6: Error profiles with PRIME-Del deletions targeting episomally encoded eGFP. a. Sample preparation schematic for amplicon sequencing. Region around the segment targeted for deletion is amplified from the genomic DNA using two-step PCR amplification that appends sequencing adaptors in the second step. **b-d.** Insertion, deletion and substitution error frequencies across sequencing reads for 24-bp deletion **(b)**, 91-bp deletion **(c)**, and 546-bp deletion **(d)**. These are based on single-end sequencing, with five replicates per experiment, all sequenced on one run, overlaid. Note that except for 24-bp deletion, only one of the two 3′-DNA-flaps is covered by the sequencing read in amplicons lacking the deletion (labeled as wild-type). Y-axis scaling is different for each plot. e. Error frequencies across 546-bp deletion after repeating amplification to allow unique molecular identifier (UMI) correction. PCR duplicates identified by UMIs were collapsed into a single read by taking the most frequent sequence sharing the same UMI. These are based on single-end sequencing, with three replicates per experiment, all sequenced on one run, overlaid. Y-axis scaling is different for each plot.

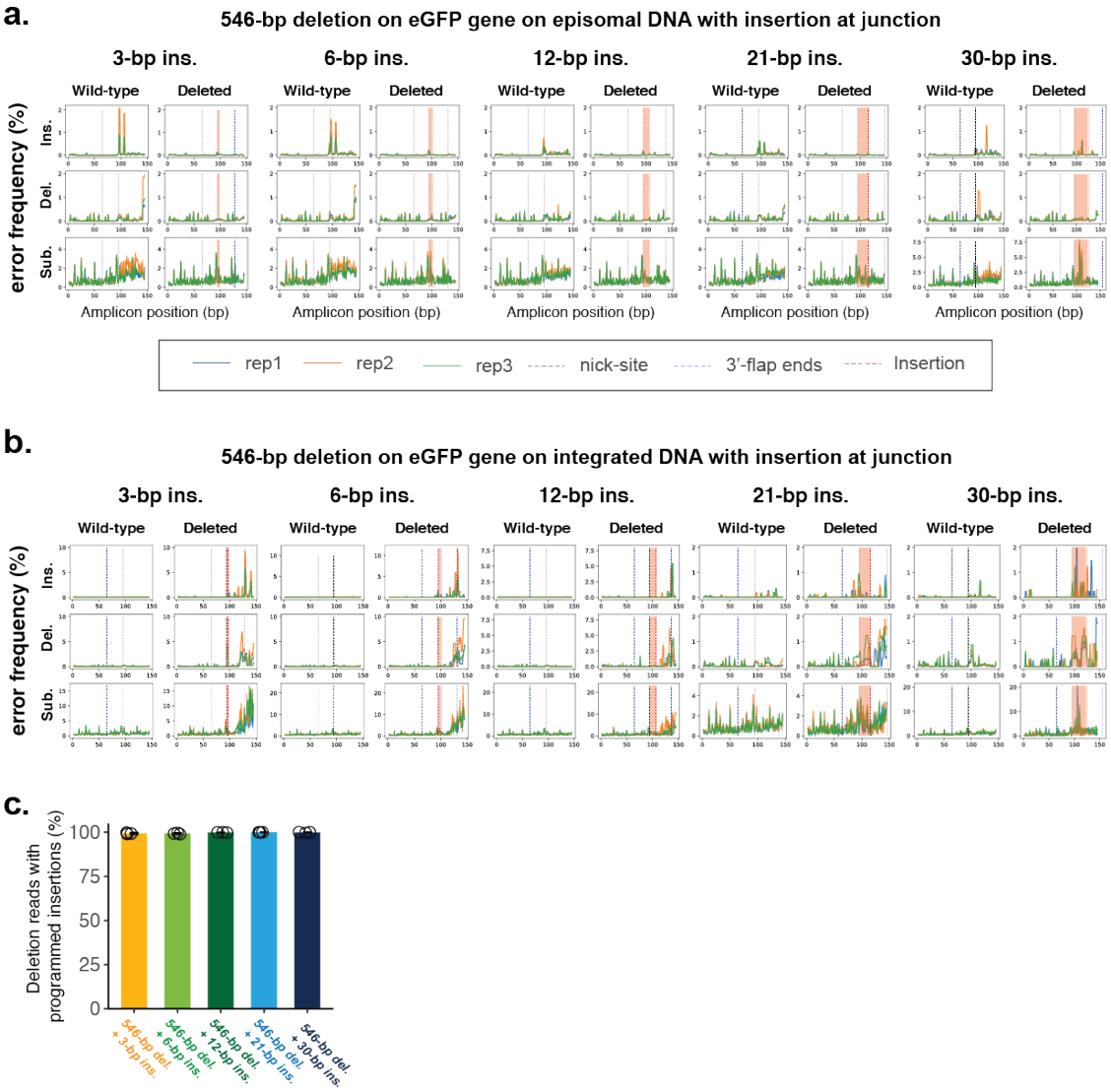

Figure 3.7: Error profiles with concurrent deletion and insertion at episomally or genomically encoded eGFP. a. Insertion, deletion and substitution error frequencies plotted across sequencing reads from concurrent 546-bp deletion and various insertion conditions, targeting episomally encoded eGFP. These are based on single-end sequencing, with three replicates per experiment, all sequenced on one run, overlaid. Note that only one of the two 3′-DNA-flaps is covered by the sequencing read in amplicons lacking the deletion (labeled as wild-type). Locations within read corresponding to insertions at deletion junction are highlighted between the nick-site (black dotted line) and end of insertion (red dotted line). Y-axis scaling is different for each plot. **b.** Same as **(a)**, but for experiments targeting a genomically integrated copy of eGFP. **c.** The percentage of reads containing the programmed deletion that also contain the programmed insertion. Similar to **Figure 3.2f**, but for experiments targeting a genomically integrated copy of eGFP. Error bars represent standard deviation for at least three replicates.

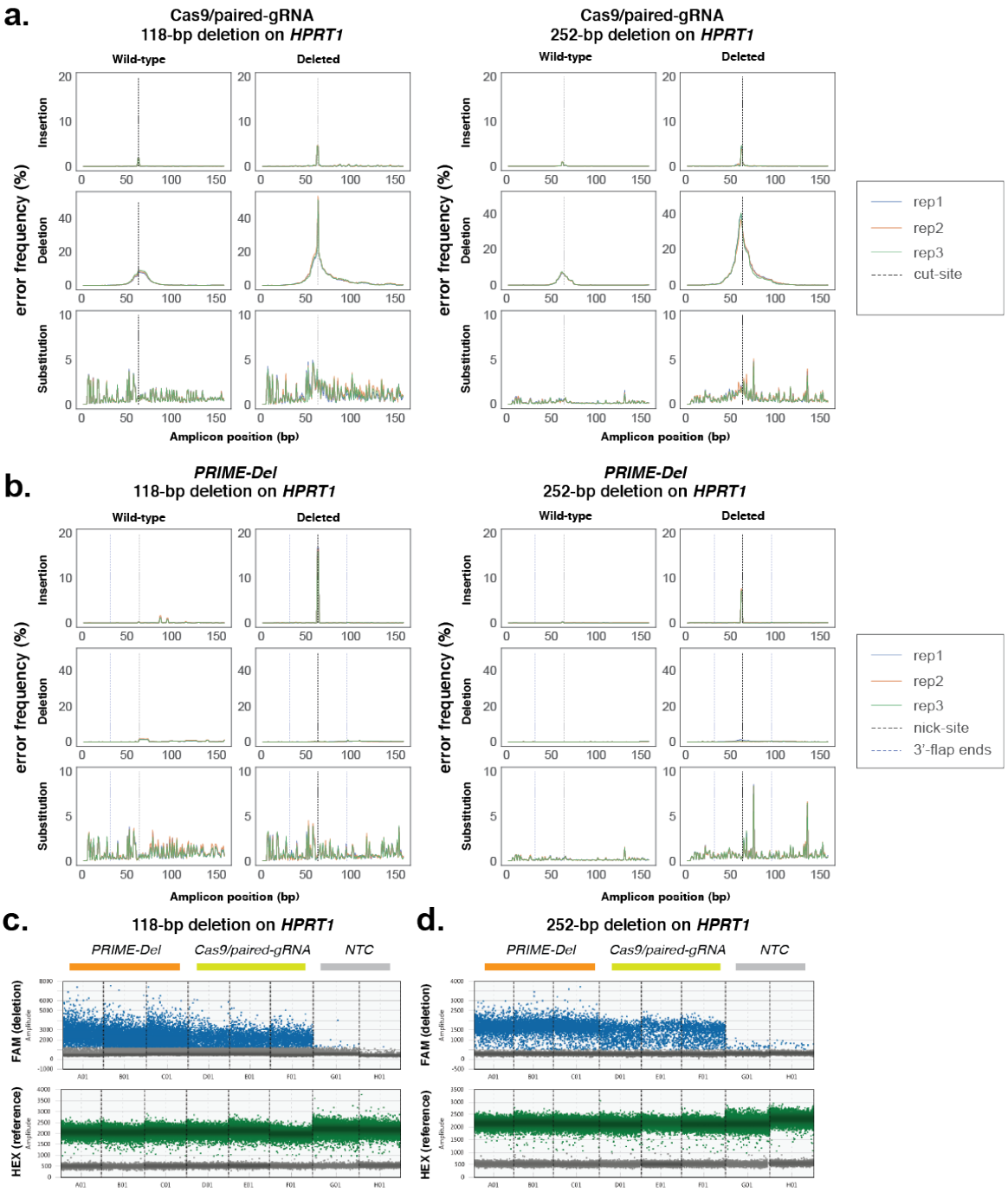
a.
Cas9/paired-gRNA
118-bp deletion on HPRT1
Cas9/paired-gRNA
252-bp deletion on HPRT1
Wild-type
Deleted
Insertion
Deletion
Substitution
error frequency (%)
Amplicon position (bp)
rep1
rep2
rep3
cut-site
b.
PRIME-Del
118-bp deletion on HPRT1
PRIME-Del
252-bp deletion on HPRT1
nick-site
3'-flap ends
c.
118-bp deletion on HPRT1
d.
252-bp deletion on HPRT1
PRIME-Del
Cas9/paired-gRNA
NTC
FAM (deletion)
HEX (reference)

Figure 3.8: Quantifying deletion efficiency and error frequency on native *HPRT1* gene. a,b. Insertion, deletion and substitution error frequencies plotted across sequencing reads from: **(a)** 118-bp or 252-bp deletion on *HPRT1* using the Cas9/paired-gRNA strategy and **(b)** 118-bp or 252-bp deletion on *HPRT1* using the *PRIME-Del* strategy. Sequencing reads aligning to the deletion reference for *HPRT1* condition are based on paired-end sequencing, while all the other conditions are based on the single-end sequencing. Each experiment has three replicates sequenced on one run, overlaid. Note that only one of the two 3′-DNA-flaps is covered by the sequencing read in amplicons lacking the deletion (labeled as wild-type) and that y-axis scaling is different for each insertion, deletion and substitution plots. **c,d**. Droplet fluorescence level in Droplet digital PCR (ddPCR) assay for: **(c)** 118-bp deletion and **(d)** 252-bp deletion. Ratio of FAM-positive droplets (detecting precise-deletion; upper panels) to HEX-positive droplets (detecting genomic DNA concentration; bottom panels) was used for measuring deletion efficiencies with *PRIME-Del* (left three wells) and Cas9/paired-gRNA (middle three wells) methods. For each probe set, negative control (NTC) was performed to ensure specific signal from precise deletion. We note that the separation is less clear (with more substantial raining patterns between negative and positive levels) in the FAM channel compared to HEX channel, possibly due to inefficient PCR amplification within the droplet. This phenomenon is more pronounced in Cas9/paired-gRNA samples, possibly due to annealing of FAM-probe to deletion junction with short (1 bp) mismatches as described previously[90]

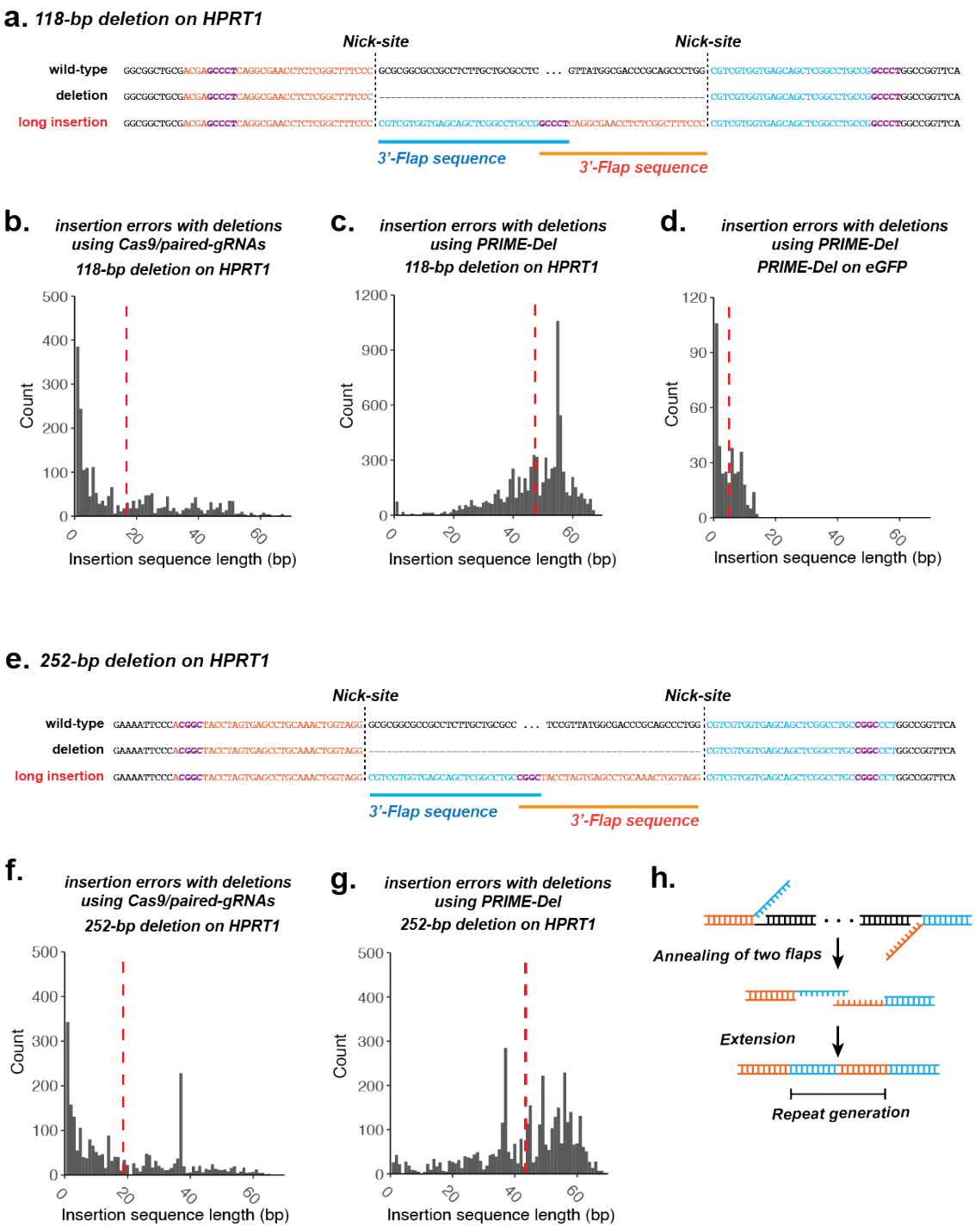
a. 118-bp deletion on HPRT1
Nick-site
Nick-site
wild-type
deletion
long insertion
3'-Flap sequence
3'-Flap sequence
b. insertion errors with deletions using Cas9/paired-gRNAs
118-bp deletion on HPRT1
c. insertion errors with deletions using PRIME-Del
118-bp deletion on HPRT1
d. insertion errors with deletions using PRIME-Del
PRIME-Del on eGFP
Count
Insertion sequence length (bp)
e. 252-bp deletion on HPRT1
Nick-site
Nick-site
wild-type
deletion
long insertion
3'-Flap sequence
3'-Flap sequence
f. insertion errors with deletions using Cas9/paired-gRNAs
252-bp deletion on HPRT1
g. insertion errors with deletions using PRIME-Del
252-bp deletion on HPRT1
h.
Annealing of two flaps
Extension
Repeat generation

Figure 3.9: Rare long insertions upon *PRIME-Del* editing of the *HPRT1* exon 1. a. We performed paired-end sequencing of amplicons derived from the *PRIME-Del*-edited *HPRT1* locus to bidirectionally cover the deletion junction and facilitate removal of PCR duplicates using 15-bp UMI sequences. This revealed recurrent long insertions that upon inspection appear to be chimeras of the two 3′ flap sequences, with overlap at their GC-rich ends (highlighted in purple). Shown here is a representative insertion from the 118-bp deletion condition. **b-d.** Histograms of insertion sequence lengths for *HPRT1* 118-bp deletion with Cas9/paired-gRNA **(b)**, *HPRT1* 118-bp deletion with *PRIME-Del* **(c)**, or eGFP 546-bp deletion with *PRIME-Del* **(d).** Red vertical lines denote the mean insertion lengths. e. Same as **(a)**, but representative insertion from the 252-bp deletion condition, also a chimera of the two 3′ flap sequences, with overlap at their GC-rich ends. **f-g.** Histogram of insertion sequence lengths for *HPRT1* 252-bp deletion with *PRIME-Del* **(f)** or Cas9/paired-gRNA **(g)**. **h.** Potential mechanism of long insertions with *PRIME-Del*. GC-rich ends of 3′-flaps of paired pegRNAs (*GCCCT* in case of 118-bp deletion and *CGGC* in case of 252-bp deletion) anneal to one another, or to another GC-rich stretch, resulting in insertion upon repair.

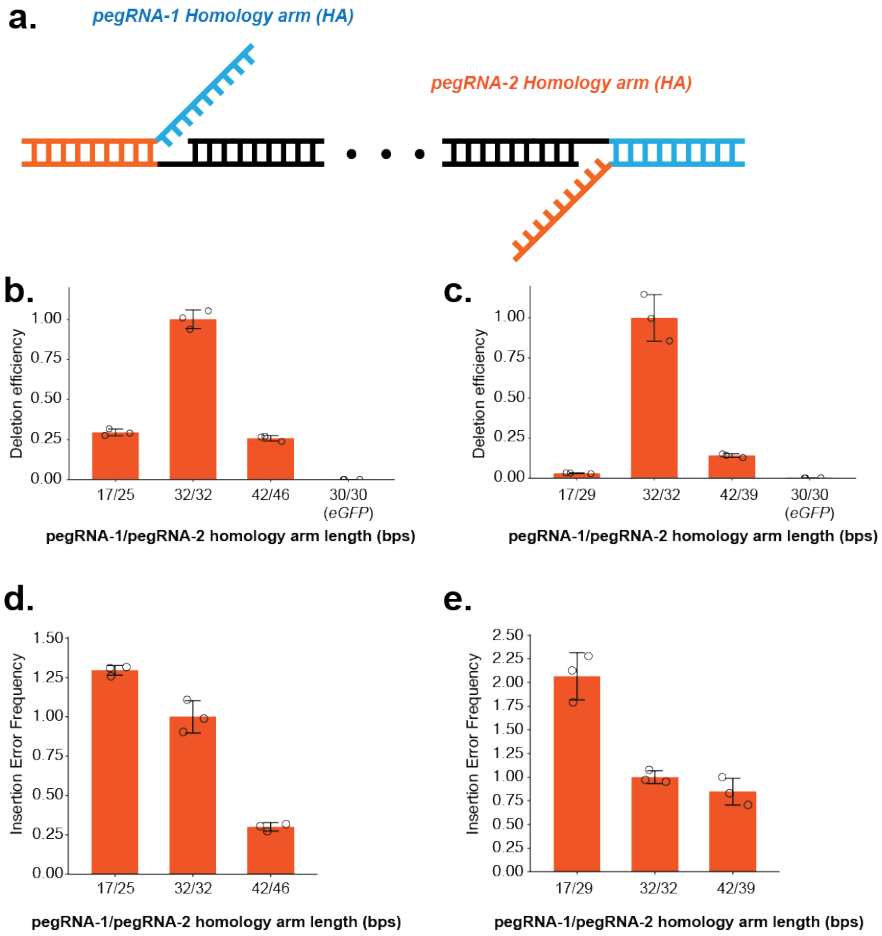

Figure 3.10: *PRIME-Del* efficiency and accuracy depends on homology arm lengths. a. Paired pegRNAs can be designed with different RT-template lengths, which effectively alters the homology arm lengths to guide the editing in *PRIME-Del*. **b-c**. Deletion efficiencies from using different homology arm lengths for **(b)** 118-bp and **(c)** 252-bp deletions of *HPRT1* exon1, normalized to the standard designs (32-bps RT templates; used in **Figure 3.3**). Using a non-homologous RT template sequence from making 546-bp deletion on eGFP (used in **Figure 3.1, 3.2**; denoted as 30/30 eGFP) does not result in deletion. **d-e.** Long-insertion frequency in *PRIME-Del* from using different homology arm lengths for **(d)** 118-bp and **(e)** 252-bp deletions of *HPRT1* exon1, normalized to the standard designs.

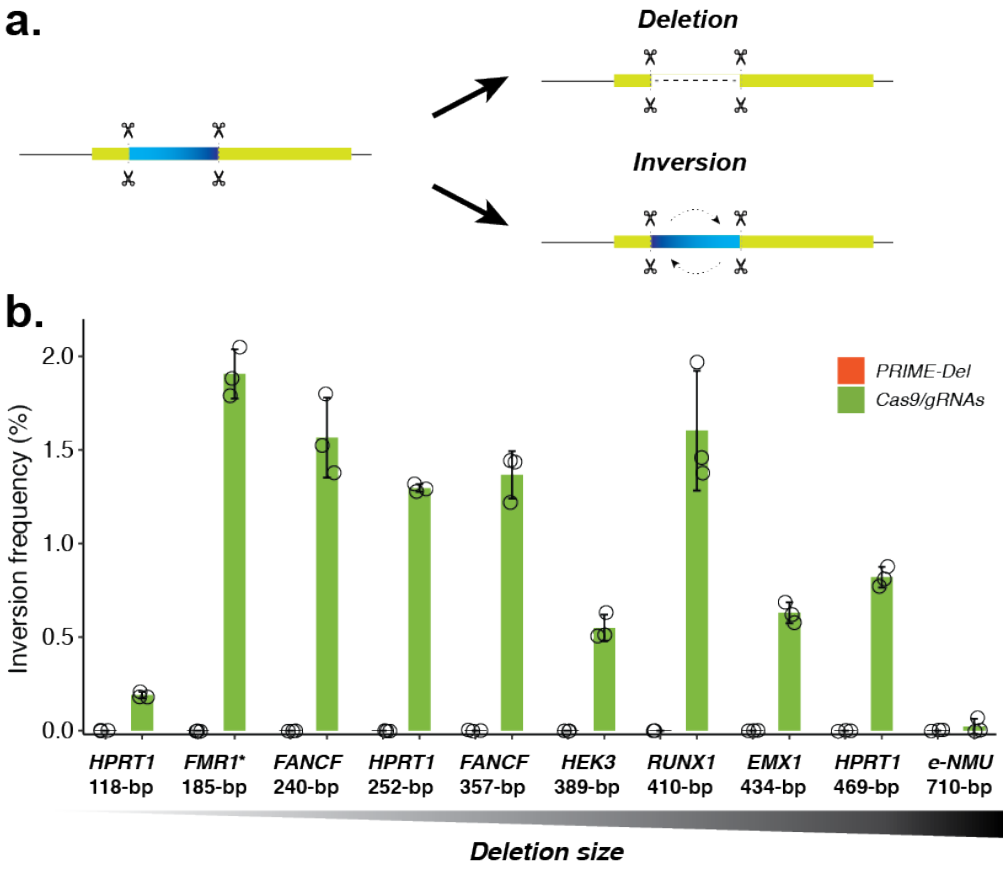

Figure 3.11: Quantifying inversion frequency on native genomic loci a. Schematic of a sequence inversion event, which is a known error mode in Cas9/paired-gRNA-mediated deletion. **b.** Estimated inversion frequencies for different deletions across the genome for *PRIME-Del* (left) and Cas9/paired-gRNA (right) methods (n = 3). UMI-based sequencing assay was used for quantification (except the GC-rich amplicon of *FMR1**, where added DMSO interfered with the UMI-addition reaction). Note that whereas they are observed for all but one of the Cas9/paired-gRNA-mediated deletions at an appreciable frequency, virtually no inversions are observed for any of these ten deletions using *PRIME-Del*.

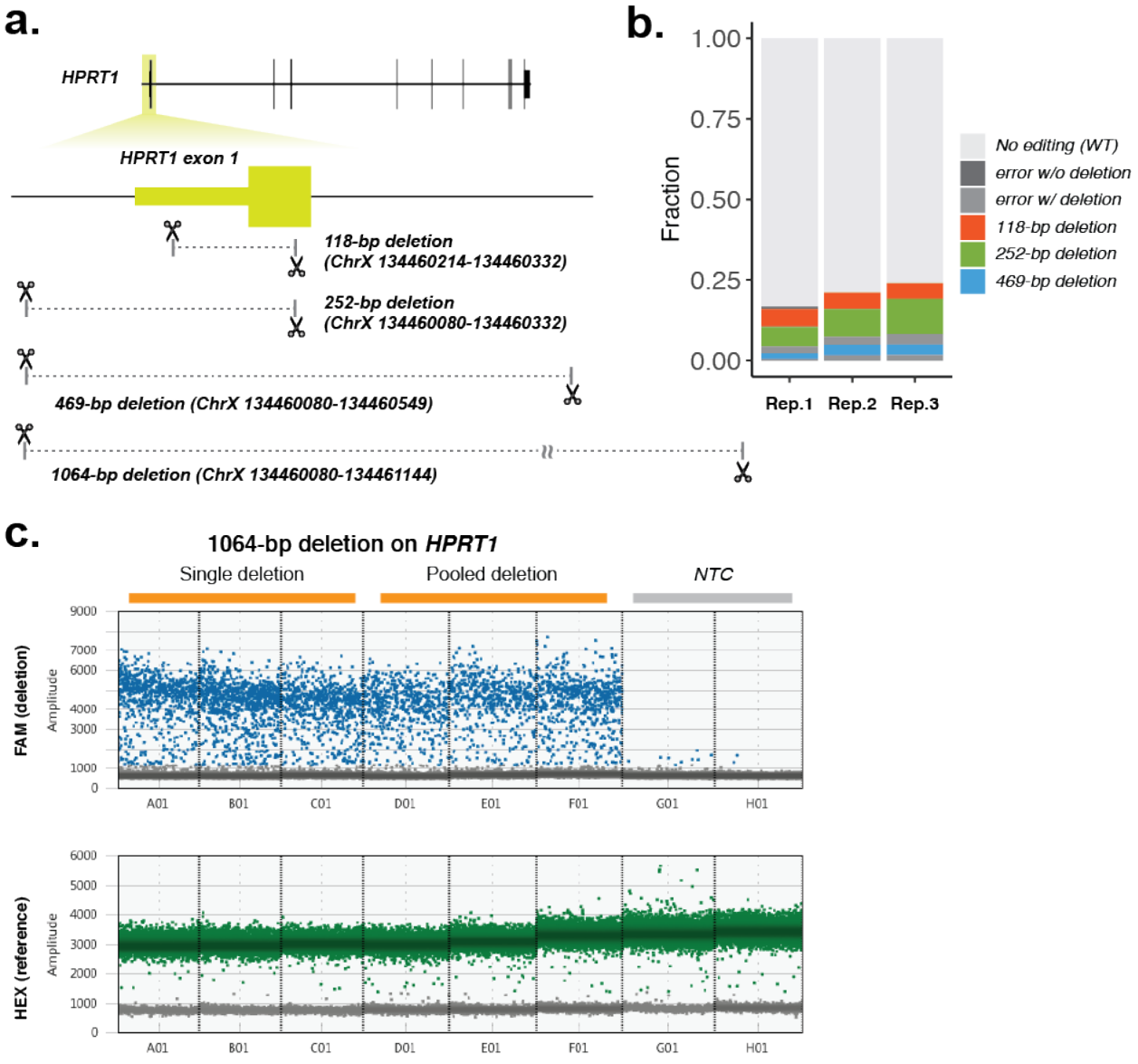

Figure 3.12: Pooled deletion using PRIME-Del. a. Cartoon representation of four deletions programmed within the *HPRT1* gene, pooled together for transfection. **b.** Deletion efficiencies and error frequencies for 3 overlapping-deletions (118, 252 and 469 bps) on *HPRT1* gene using PRIME-Del in HEK293T cells. Three transfection replicates are plotted separately. **c.** 1064-bp deletion efficiencies compared between single-deletion (left three wells) and pooled PRIME-Del (middle three wells). Estimated editing efficiencies for 1064-bp deletion in pooled PRIME-Del are 1.7%, 1.9% and 2.0% for three transfection replicates.

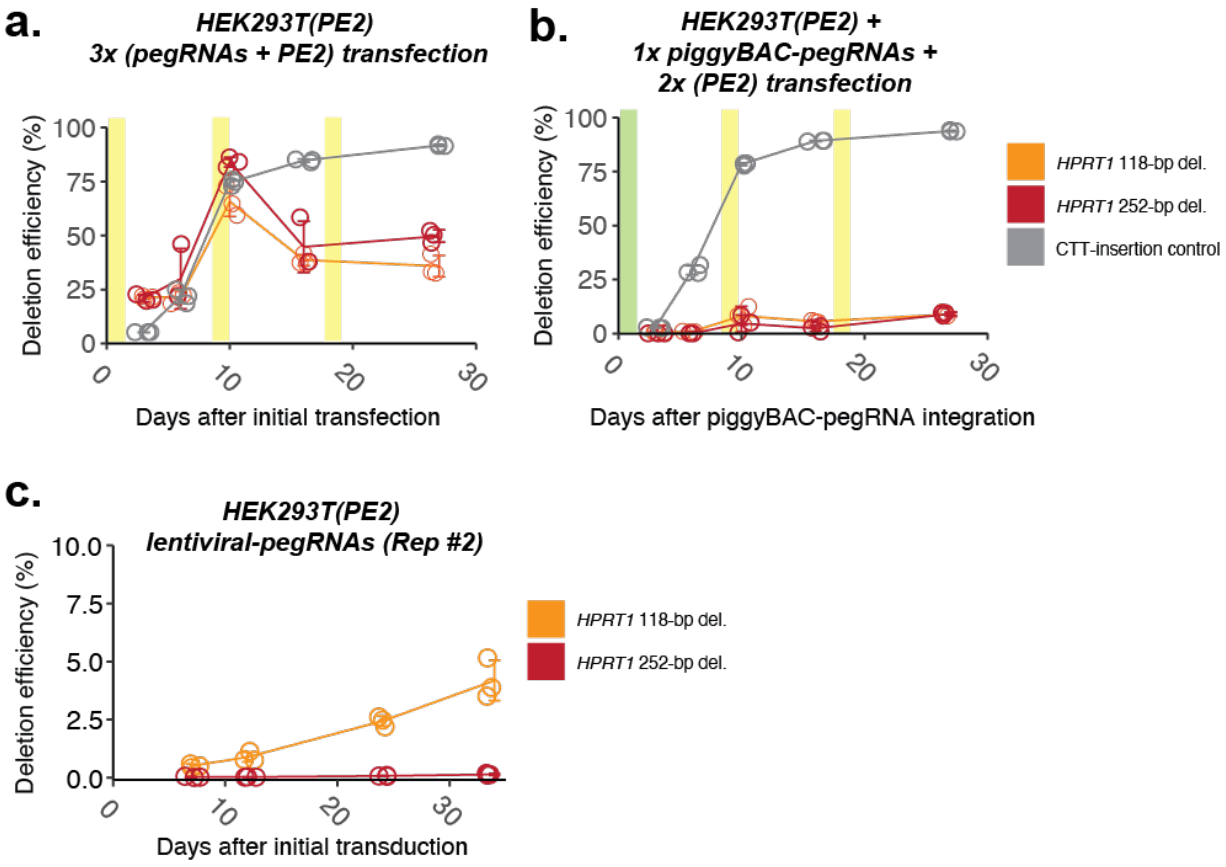

Figure 3.13: Multiple transfections enhance *PRIME-Del* efficiency in monoclonal HEK293T (PE2) cells. a. Editing efficiencies measured for the 118-bp and 252-bp deletions at genomic *HPRT1* exon 1 using *PRIME-Del* (paired-pegRNA construct) or CTT-insertion using prime-editing (single-pegRNA construct), as a function of time after initial transduction of pegRNA(s). Plasmids bearing paired-pegRNAs and Prime Editor-2 enzyme were transfected 3 times (days 0, 9, 18; highlighted in yellow) into Prime Editor-2 enzyme-expressing HEK293T cells. Error bars represent standard deviation for three replicates. **b.** Same as **(a)**, but first with integration of pegRNAs to PE2-expressing HEK293T via piggyBAC transposon system on Day 0 (highlighted in green), followed by two additional transfections of plasmid bearing Prime Editor-2 enzyme only on Day 9 and 18 (highlighted in yellow). Error bars represent standard deviation for three replicates. **c.** Second replicate for experiment shown in **Figure 3.4c**, where deletion efficiencies are measured for the 118-bp and 252-bp deletions at *HPRT1* exon 1 using *PRIME-Del* as a function of time after initial transduction of pegRNA(s).

Chapter 4

ENGRAM: MULTIPLEX MOLECULAR RECORDING OF BIOLOGICAL SIGNALS AND EVENTS

This Chapter is adopted from manuscript under review with minimum changes

Chen, W.*, Choi, J.*, Nathans, J.F., Agarwal, V., Martin, B., Nichols, E., Leith, A., Lee, C. and Shendure, J., 2021. Multiplex genomic recording of enhancer and signal transduction activity in mammalian cells. *bioRxiv*.

Author contribution: Wei Chen designed and performed the experiments with the help from Junhong Choi, Xiaoyi Li, Wei Yang and Jenny Nathans. Wei Chen analyzed the data. Wei Chen and Jay Shendure wrote the manuscript.

A back story about ENGRAM: Recording molecular events and signals in cells have long been my dream since my first day in the Shendure lab. We first started with CRISPR/Cas9 as the writing unit. It works but only records binary information of if certain events happened before, no quantitative information or order of events were able to be captured by Cas9. We waited for years until the prime editor came out in 2019. With its precision and programmable feature, we quickly realized that the prime editor is the ideal molecular recorder for our purpose. I and Junhong immediately started to work on information encoding using the prime editor. Many ideas are from our late-night discussions and we worked closely on almost everything. I mainly focused on ENGRAM while he mainly focused on DNA TICKER TAPE (DTT) and a side journey to Prime-del. We decided to share authorships on two of these papers to celebrate our friendship and co-efforts in developing these tools.

4.1 Abstract

Measurements of gene expression and signal transduction activity are conventionally performed with methods that require either the destruction or live imaging of a biological sample within the timeframe of interest. Here we demonstrate an alternative paradigm, termed ENGRAM (ENhancer-driven Genomic Recording of transcriptional Activity in Multiplex), in which the activity and dynamics of multiple transcriptional reporters are stably recorded to DNA. ENGRAM is based on the prime editing-mediated insertion of signal- or enhancer-specific barcodes to a genomically encoded recording unit. We show how this strategy can be used to concurrently record the relative activity of at least hundreds of enhancers with high fidelity, sensitivity and reproducibility. Leveraging synthetic enhancers that are responsive to specific signal transduction pathways, we further demonstrate time- and concentration-dependent genomic recording of Wnt, NF-κB, and Tet-On activity. Finally, by coupling ENGRAM to sequential genome editing, we show how serially occurring molecular events can potentially be ordered. Looking forward, we envision that multiplex, ENGRAM-based recording of the strength, duration and order of enhancer and signal transduction activities has broad potential for application in functional genomics, developmental biology and neuroscience.

4.2 Introduction

During development, a modest number of core signaling pathways and gene regulatory modules are leveraged to program a precise spatiotemporal unfolding of programs of cell differentiation, proliferation, morphogenesis and tissue patterning[107]. Across species, differences in how these conserved pathways and modules are used underlies an incredible diversity of organismal form and function. Within species, genetic differences and environmental effects are presumed to influence these core modules in specific developmental or homeostatic contexts, giving rise to both natural phenotypic variation as well as myriad disease states.

How do we capture the activity of these pathways and modules? Measurements of gene expression and signal transduction activity are conventionally performed with methods that require either the destruction or live imaging of a biological sample. These include RNA sequencing (RNA-seq), which measures the global transcriptional state of a system; massively parallel reporter assays (MPRAs), which use sequencing to measure the relative ability of members of a library of DNA fragments to act as enhancers of transcriptional activity in a controlled context[108]; and fluorescent probes and reporters, which track the dynamics of specific signaling pathways in living systems[109].

These classes of methods are remarkably useful and yet limited in key ways. For example, with RNA-seq, individual samples provide only static snapshots of cell state, such that the temporal dynamics of gene expression must be pieced together by inference with a resolution that is limited by sampling density. Sequencing-based reporter assays are also destructive and static. Although time-series MPRAs can successfully define the temporal dynamics of enhancer activity[110], such studies are similarly limited by inference and sampling density. Fluorescent probes and reporters are better positioned to capture temporal dynamics, but require that the biological system be physically transparent, at least for live imaging, and are limited in terms of multiplexibility. Overall, there remains a need for a means of capturing signaling and gene regulatory activity that is at once quantitative, reproducible, non-destructive, multiplexable, applicable to physically opaque biological systems and capable of integrating large numbers of signals.

DNA is the natural medium for biological information storage, and is easily read through sequencing. To date, a variety of enzymatic systems have been used to alter primary DNA sequence in a signal-responsive manner, most prominently site-specific recombinases (SSRs) and CRISPR genome editing[16]. For example, the Cre SSR is typically expressed under the control of a signal-specific promoter or enhancer. Cre-mediated recombination at a target locus excises sequences between pairs of lox sites, resulting in expression of one or multiple fluorescent reporters[7, 8, 9]. The recording event is irreversible, *i.e.* the descendants of those cells in which the DNA rearrangement occurred continue to express the reporter regardless

of whether the gating regulatory element remains active or not.

CRISPR systems have been used for signal-specific recording in several different ways. A first class of methods repurpose the CRISPR-Cas spacer acquisition system to log events or content in prokaryotic systems, *e.g.* DNA, RNA or metabolites[19, 20, 111, 112, 113]. A second class of systems, including the CAMERA and DOMINO[17, 18] methods, link the activity of CRISPR base editors to the presence of specific small molecules or signaling pathway activity, such that observed base edits serve to irreversibly record those signals.

Although pioneering, these systems are fundamentally limited with respect to multiplexibility – that is, the number of independent signals that can be recorded at once. In examples that have been demonstrated to date, enhancers are used to selectively drive the enzyme that mediates an alteration in DNA sequence or limited transcription repression elements are used to drive the gRNA expression in response to signals. In this framing, each signal requires its own enzyme or repression element, and it is difficult to imagine how more than a handful of independent signals could be concurrently recorded within the same cell or population of cells, let alone how extensive, concurrent recording of large numbers of biological signals could be achieved throughout the development of a multicellular organism.

Another challenge for the current recording systems is reading out. In current Cas9/Base editor based recording systems, single guide RNAs (sgRNAs) program the location of editing but not the edit itself. As such, each sgRNA would require its own target, making it difficult to read out all targets at once. This challenge has been partially solved with paired sgRNA-target[24, 114, 23, 115] or homing gRNA (hgRNA) or self-targeting gRNA (stgRNA)[116, 117], but still has limited compatibility with recent development of RNA-seq. (See summary in **Table 4.1**). An ideal recorder should be able to simultaneously record multiple signals and read them out with either DNA amplicon or RNA sequencing.

Here we describe a new framework for multiplex transcriptional recording, which we term ENGRAM (ENhancer-driven Genomic Recording of transcriptional Activity in Multiplex). In brief, ENGRAM relies on enzymatic release[118, 119, 120, 121] of prime editing guide RNAs (pegRNAs)[122] from synthetic transcripts driven by *cis*-regulatory-element (CRE)-

coupled Pol-II promoters. Each pegRNA programs the insertion of a specific barcode to a genomically-encoded recording locus (DNA Tape). Because each CRE is coupled to a distinct pegRNA-encoded insertion, multiple ENGRAM recorders can operate in parallel, all relying on the same prime editing enzyme and all competing to write to the same DNA Tape, which can be read out at either the DNA or RNA level. Of note, ENGRAM is the hypothetical memory storage unit in the brain. We would like to use this as the memory storage in cells too.

4.3 Results

4.3.1 Development and evaluation of ENGRAM

An ideal DNA-based transcriptional recorder would log the production of specific transcripts, *cis*-regulatory activities and/or signal transduction pathways, via specific changes to the primary sequence of a genomic recorder locus. In seeking to develop a DNA-based recorder for mammalian systems, we were inspired by reporter assays, an established approach wherein a *cis*-regulatory element (CRE) of interest is positioned upstream of a minimal promoter (minP) and reporter gene (*e.g.* luciferase). Reporter assays are amenable to extensive multiplexing, as the reporter can include a transcribed barcode that is linked to the CRE, resulting in the MPRA[123]. However, as noted above, MPRAs depend on targeted RNA-seq of the barcodes, which is destructive and static. Nonetheless, we reasoned that the basic MPRA architecture, *i.e.* a library of synthetic or natural enhancers positioned upstream of a minimal promoter, might be coupled to the expression of a library of writing units, in the form of pegRNAs (**Figure 4.1a**). Specifically in ENGRAM, each CRE is linked to a pegRNA encoding a specific insertion to a common DNA TAPE.

One challenge to this scheme is that in order to be appropriately processed, transcripts for most translated genes, including CRE-minP-driven reporter transcripts, are made by RNA polymerase II (Pol-2), whereas small untranslated RNAs, including guide RNAs, are made by RNA polymerase III (Pol-3). To address this, we leveraged the CRISPR endori-

bonuclease Csy4 (also known as Cas6f), which recognizes and cuts at the 3′ end of 17-bp RNA hairpins (*csy4*)[118, 119, 120, 121]. Expression of Csy4, together with CRE-activity-dependent expression of *csy4*-pegRNA-*csy4*, should result in a liberated functional pegRNA (**Figure 4.1a**).

We first developed ENGRAM 1.0, in which *csy4*-pegRNA-*csy4* is embedded within the 3′ untranslated region (UTR) of a GFP transcript and the Csy4 is constitutively expressed (**Figure 4.5a**). To benchmark the activity of pegRNAs released from Pol-2 transcripts, we compared an ENGRAM 1.0 recorder driven by a constitutive Pol-2 promoter (PGK) to a conventional, U6-driven pegRNA. In both cases, the pegRNAs target the endogenous HEK293 target 3 (*HEK3*) locus and are designed to insert three nucleotides (CTT)[122]. These constructs were separately transiently transfected to monoclonal HEK293T cells constitutively expressing Prime-Editor-2 (PE2) and Csy4. Five days after transfection, we harvested genomic DNA, and then PCR amplified and sequenced the *HEK3* locus. We observed comparable, reproducible efficiencies of CTT insertion between the ENGRAM 1.0 and U6 recorders (mean 5.9% and 5.3% across three replicates, respectively; **Figure 4.5b**). Next, we replaced the constitutive PGK promoter with a CRE-minP architecture, in which thirteen 170-bp sequences with known enhancer activity in K562 cells were selected[123] We compared the editing efficiency of the pool of enhancer-driven recorders vs. a pool of negative controls (minP with no upstream enhancer) via their transient transfection to K562 cells constitutively expressing both PE2 and Csy4. We successfully recorded enhancer-activated barcode insertions with a collective efficiency of 3.9%, 1.93-fold higher than the editing efficiency of pegRNAs driven by minP alone (**Figure 4.5c**). Overall, these results suggest that ENGRAM-based recording can work. However, the signal-to-noise ratio was considerably more modest than we had hoped for. We speculated that this was due in part to the accumulation of background edits due to constitutive expression of Csy4.

To reduce the background accumulation of edits to the DNA Tape, we designed a new ENGRAM architecture in which the GFP ORF is replaced by Csy4 ORF, and Csy4 is no longer constitutively expressed (**Figure 4.1b, Figure 4.5d**). In this recorder design,

termed ENGRAM 2.0, the expression of Csy4 and the pegRNA are both dependent on enhancer activity. To evaluate whether these modifications reduce background recording, we tested ENGRAM 1.0 vs. 2.0 in the absence of any enhancer, *i.e.* minP alone driving peg5N. Transiently co-transfecting these constructs into HEK293T cells with either PE2-Csy4 plasmid (for ENGRAM 1.0) or PE2 plasmid (for ENGRAM 2.0) in triplicate, we indeed observed a 2.8-fold reduction in background recording with ENGRAM 2.0 relative to ENGRAM 1.0 (mean 1.4% for ENGRAM 1.0 $\rightarrow$ 0.5% for ENGRAM 2.0, 3 days post-transfection) (**Figure 4.5e**).

Towards further reducing background, we designed two additional recorders: 5′ ENGRAM 2.0, in which the *csy4* hairpin-flanked pegRNA is embedded within the 5′ (rather than 3′) UTR of the Csy4 transcript; and 3′-FT ENGRAM 2.0, which contains an additional *csy4* hairpin in its 5′ UTR to create auto-regulatory negative feedback loop on Csy4 levels (**Figure 4.2b**). We first measured the background recording activity by integrating them into HEK293T cells expressing PE2 (PE2(+) HEK293T) cells via piggyBAC. The 5′ ENGRAM 2.0 and 3′-FT ENGRAM 2.0 recorders respectively exhibited 12-fold and ¿100-fold reductions in background activity, relative to 3′ ENGRAM 2.0 (10 days post-transfection; **Figure 4.1c**). Of note, for all three of these integrated ENGRAM 2.0 recorders, the level of background recording plateaued after several days (**Figure 4.1c**). This suggested to us that the accumulation of background recording events mostly occurs shortly after transfection, potentially due to ORI-driven, plasmid-mediated transcription28,29, rather than minP-driven transcription from integrated recorders. However, some degree of accumulation persisted with the 3′ ENGRAM 2.0 recorder, suggesting an additional component of genomically driven background activity. We then measured their responsiveness to enhancer activation by placing a NF-κB responsive element (activated by TNFα) in the upstream of the minP. All three recorders with NF-KB responsive element are integrated into PE2(+) HEK293T cells via piggyBAC. We measured their recording activity in the absence or presence of the ligand TNFα. We observed 1.4, 13.3, 23.8 fold activation for 3′, 5′ and 3′-FT recorders, respectively (**Figure 4.1d**). Although the 3′-FT design exhibited the lowest back-

ground activity, and highest activation response to enhancer activation, we moved forward with the 5′ ENGRAM 2.0 design because its organization facilitates straightforward pairing of CREs and pegRNA-mediated insertions during cloning. Unless specified, ENGRAM in the paper specifically refers to ENGRAM 2.0 5′ architecture.

From above recording data, we observed different efficiency for 5N barcodes. To systematically analyze the editing efficiency bias, we cloned an ENGRAM recorder with pegRNA targeting *HEK3* locus to install 5N degenerate insertion driven by a PGK promoter (**Figure 4.1e**). We transiently transfected PE2(+) HEK293T cells and measured recording efficiency at 3 days post transfection. Overall, we observed 1,023 of 1,024 all possible 5-bp insertions at the *HEK3* locus with highly reproducible frequencies (**Figure 4.1f, Figure 4.5a-c**). After normalizing for their abundance in the plasmid pool and removing under-represented barcodes, we observed 948 5-mers with balanced insertional efficiencies, with 91% falling within a 4-fold range (**Figure 4.1g**). We suspected that heterogeneity in insertional efficiencies might be a consequence of the influence of the 5-mer on pegRNA secondary structure. Consistent with this, the least efficient 5-mer is predicted to pair with the spacer sequence to form a more stable secondary structure, while the most efficient 5-mer insertion does not (**Figure 4.5d, e**). To ask whether we could predict insertional bias, we performed linear lasso regression with 84 binary sequence features and 1 secondary structural feature (minimum free energy (MFE), Methods). The resulting model was reasonably accurate, with MFE emerging as the most predictive feature (**Figure 4.1h, Figure 4.5f, g**). For subsequent experiments shown in this paper, we rigorously controlled for this bias by picking barcodes with more balanced insertion efficiency.

During the development of ENGRAM, two studies showed that engineered pegRNA (epegRNA, with tevoPreQ1 hairpin)[124] and new prime editor architecture (PEmax)[125] can improve the editing efficiency. To improve ENGRAM recording efficiency, we tested both epegRNA and PEmax in the context of 5′-ENGRAM. We transiently transfected PE2(+) HEK293T cells with pegRNA and epegRNA encoding a 5N insertion, both driven by PGK promoter, and measured their recording efficiency at 3 days post transfection. Surprisingly,

we observed a slightly lower efficiency in epegRNA than pegRNA (16.6% vs 22.2% in epegRNA and pegRNA, respectively, ∼30% lower. **Figure 4.7a**). We reasoned that the *csy4* hairpin might serve a similar role as tevoPreQ1 to protect pegRNA from degradation, additional hairpin to *csy4* might disrupt RNA folding. We co-transfected PE2 or PEmax with PGK-5N and measured their editing efficiency at 3 days post-transfection. We observed a 1.7 fold increase in editing efficiency with PEmax (**Figure 4.7b**). We would recommend using PEmax for all future ENGRAM recording experiments. With 5′ ENGRAM, we also tested if tRNA[126] can be an alternative pegRNA processing architecture for ENGRAM. We replaced *csy4* hairpin with tRNA and measured their recording activity. However, we don't see any edits with tRNA-ENGRAM.

4.3.2 Multiplex recording of enhancer activity with ENGRAM

With sensitive and robust 5′-ENGRAM, we next sought to test if ENGRAM can work as traditional MPRA. We cloned enhancer libraries to the upstream of minP in the 5′-ENGRAM construct and integrated them into PE2+ K562 cells. The pegRNA is targeting the *HEK3* locus and encoding a 5-bp or 6-bp short insertion. Thus, enhancer activity can be recorded on either endogenous DNA TAPE (genomic *HEK3* locus, 2 copies) or synthetic DNA TAPE (piggyBac integrated *HEK3* locus, 10-30 copies). The abundance of barcodes in DNA TAPE is compared to the barcode abundance in pegRNA (**Figure 4.2a**). We first cloned a pair of 170-bp sequences previously shown to have either high vs. minimal enhancer activity in K562 cells[123] upstream of minP, together with minP-only and promoter-less constructs (**Figure 4.2b**). Each of these four constructs drove pegRNAs encoding two distinct 5-bp insertions. An equimolar mixture of these 8 recorder plasmids was introduced via piggyBAC integration into PE2+ K562 cells in triplicate. At five days post-transfection, 3.14% of endogenous *HEK3* target sites were edited, but ∼90% of inserted barcodes were associated with the active enhancer (**Figure 4.2b**). Of note, the 17.3-fold difference in recorded insertional frequency between the active and inactive enhancer roughly matched the 15-fold difference between them measured by MPRA[123].

To more generally evaluate whether the enhancer activities recorded by ENGRAM are quantitatively comparable to corresponding measurements made by MPRA, we cloned 300 enhancer fragments[123] to the 5′ ENGRAM construct, each driving a pegRNA encoding a unique 6-bp insertion(**Figure 4.2g**). Five days after introducing these recorders via piggyBAC to PE2-expressing K562 in triplicate, we separately recovered, amplified and sequenced the *HEK3* locus (from DNA, both endogenous locus and synthetic locus) or the transcribed barcode itself (from RNA). From DNA, we observed an overall editing efficiency of 3.08% and 1.76% for endogenous and synthetic *HEK3* locus, respectively (**Figure 4.8a**), and recovered 292 of 300 barcodes. We sampled various depths (6,000, 12,000, 24,000, 48,000, 96,000 cells) on both endogenous and synthetic *HEK3* locus and compared their recording efficiency and sensitivity. Overall, the enhancer activity recorded on endogenous and synthetic DNA TAPE are highly correlated (**Figure 4.8b**). With 15 copies of synthetic DNA TAPE in the genome, we are able to record 300 enhancer activity with as little as 12,000 cells with reasonable capture efficiency and reproducibility (**Figure 4.8c, d**). We recommend having at least 100 cells/enhancer for robust enhancer activity recording. We reasoned that with improved recording efficiency, this number can be lower. Both RNA and DNA-based measurements were highly consistent between transfection replicates (**Figure 4.6e, f**). Furthermore, we observed a strong correlation between the recorded activities (ENGRAM; DNA) and the directly measured activities (MPRA; RNA), indicating that the relative transcriptional activities of enhancer reporters can be quantitatively recorded to genomic DNA (**Figure 4.2c**).

4.3.3 Quantitative recording of signaling pathway activation or small molecule exposure with ENGRAM

We next sought to ask whether ENGRAM could be used to record the intensity or duration of signaling pathway activation or small molecule exposure. For this, we selected several signal-responsive regulatory elements: the Tet Response Element (TRE; activated by doxycycline)[127], a NF-κB responsive element (activated by TNFα)[128], and a TCF-LEF

responsive element (Wnt signaling pathway; activated by CHIR99021)[129], each previously used to drive fluorescent reporters in a signal-responsive manner (**Table 4.2**). These signal-responsive sequences were cloned upstream of minP within 5′ ENGRAM 2.0 recorders, with each driving expression of a pegRNA encoding one or two specific insertions to the endogenous *HEK3* locus (**Figure 4.3a**). The three recorders were separately integrated into the genomes of PE2(+) HEK293T cells via piggyBAC in triplicate (for the doxycycline recorder, constitutively expressed reverse tetracycline-controlled transactivator (rtTA) was integrated separately). A 2-fold dilution series of doxycycline, TNFα or CHIR99021 (for CHIR99021 we tested more concentration around 1-4 μM) was added to the media of the cell lines into which the relevant recorder had been integrated, and genomic DNA was harvested 48 hours after the onset of exposure.

For all three signal-responsive ENGRAM recorders, editing rates at the *HEK3* locus exhibited a strikingly sigmoidal dependence on the log-transformed concentration of the corresponding stimulant (**Figure 4.3b-d**). This was particularly the case for the Wnt signaling, wherein the corresponding recorder exhibited nearly switch-like behavior across an approximately four-fold range of CHIR99021 concentration (**Figure 4.3d**). As with previous experiments, each ENGRAM recorder exhibited minimum non-accumulating, basal recording even in the absence of signal exposure (0.1-0.2%; **Figure 4.9a**), potentially due to ORI-driven, plasmid-mediated transcription[130, 123] shortly after transfection, as discussed above. We observed a dynamic range in editing efficiency between background vs. maximal stimulation of 11.5-fold, 19.0-fold and 22.6-fold for the Tet, NF-κB and Wnt recorders, respectively (**Figure 4.3e**).

To explore the dependence of ENGRAM on not only the intensity of signals but also their duration, we performed a matrix experiment on the NF-κB and Wnt recorders, varying stimulant concentration as previously but also varying the duration of exposure from 6 to 48 hours (2 recorders x 8 concentrations x 8 durations x 3 replicates = 384 conditions; **Figure 4.3f-g**). In this experiment, each batch of cells was harvested 24 hours after the removal of stimulants from the media. In the resulting levels of editing, the dependency

of the NF-κB and Wnt recorders on both the intensity and duration of stimulation was immediately evident (**Figure 4.3f-g**). For both recorders, even 6 hours of stimulation was sufficient to observe signal in excess of background. However, the NF-κB recorder appeared to exhibit faster kinetics than the Wnt recorder (**Figure 4.9 b-c**).

4.3.4 Multiplex recording of signaling pathway activity with ENGRAM

We next sought to introduce multiple ENGRAM recorders for different signaling pathways into a single population of cells, to evaluate whether they could be used together, *i.e.* competing to write to a shared DNA Tape (**Figure 4.3h**). In brief, constructs corresponding to the TetON, NF-κB and Wnt recorders were mixed at an equimolar ratio and co-integrated to PE2(+) HEK293T cells. Each recorder drives pegRNA(s) encoding the insertion of one or two distinct, signal-specific barcodes (**Table 4.2**). These cells were exposed to a high concentration of all possible combinations of 0 to 3 stimuli, in triplicate (8 on/off stimulus combinations x 3 replicates = 24 conditions). Harvesting cells after 48 hours of stimulation, we performed PCR amplification and sequencing of the shared DNA tape. As predicted, the abundances of signal-specific barcodes were highly dependent on the precise combination of stimuli applied (**Figure 4.3i**). Put another way, we observed minimal cross-talk, consistent with the orthogonality of these signaling pathways to one another (**Figure 4.9d**). To push this system further, we performed a separate experiment in which populations of cells bearing all three recorders were exposed to all possible combinations of low, medium or high concentrations of each stimulus (3 concentrations $^{3stimuli}\times$ 3 replicates = 81 conditions). Once again harvesting cells after 48 hours and reading the DNA Tape, we observe that signal-specific barcodes are introduced at rates correlated with the concentration of the corresponding stimulus (**Figure 4.3j, Figure 4.9**), further supporting the conclusion that these recorders are able to capture quantitative information on separate channels despite writing to a shared DNA Tape.

4.3.5 Capturing the order in which ENGRAM recorders are active

In the context of a multiplex signal recorder, it is obviously of interest to capture not only the intensity and duration of individual signals, but also the order in which they are active relative to one another. To this end, we devised ENGRAM 2.0 recorders that each comprise an operon of multiple, *csy4* hairpin-flanked pegRNAs, each designed to program insertional edits but in a manner that depends on whether other edits had (or had not) already occurred. For example, in the simplest version of this scheme, we might want to map the order of two signaling events, A and B (**Figure 4.4d**). For this goal, an A-responsive recorder would encode a first pegRNA that that wrote an A-specific barcode to blank DNA Tape **a.**, but also a second pegRNA that only targeted an already B-edited DNA Tape with a different barcode (A'). Meanwhile, a B-responsive recorder would encode a first pegRNA that that wrote an B-specific barcode to blank DNA Tape **b.**, but also a second pegRNA that only targeted an already A-edited DNA Tape with a different barcode (B').

To test this concept, we cloned ENGRAM 2.0 recorders encoding AA' or BB' pegRNA operons, each driven by the constitutive PGK promoter. We then performed a series of transfection programs in which either both A & B were introduced simultaneously (1 program), only A or B was introduced (2 programs), or the recorders were serially transfected (A→B or B→A) with the recovery time between transfections varying between 8 and 72 hours (8 programs) (**Figure 4.4e**). These experiments were performed in triplicate in PE2(+) HEK293T cells, with harvesting, amplification and sequencing of the DNA Tape at five days after the first transfection (11 programs × 3 transfection replicates = 33 conditions) (**Figure 4.8c**). As predicted, and provided there were 24+ hours of recovery between transfections, we observed grossly different ratios of AB'/BA' edits for (A→B) vs. (B→A) programs (**Figure 4.4f, Figure 4.8d**), indicating that the general scheme is compatible with the recovery of information about the order in which ENGRAM 2.0 recorders are active.

4.4 Discussion

Here we describe ENGRAM, a new strategy for multiplex, DNA-based signal recording, wherein each biological signal of interest is coupled to the Pol-2-mediated transcription of a specific guide RNA, whose expression then programs the insertion of a signal-specific barcode to a genomically encoded DNA Tape. As DNA is stable, recorded signals can be read out at any subsequent point in time, *e.g.* by DNA sequencing or, potentially, even by DNA FISH. A key strength of ENGRAM is its multiplexibility. For example, with the 5-bp or 6-bp insertions used here, thousands of distinct biological signals can potentially be recorded within the same cell, all competing to write to a shared DNA Tape. We demonstrate this multiplexibility by showing that analogous to an MPRA, ENGRAM can be applied to concurrently and quantitatively capture the activity of hundreds of enhancers. However, unlike an MPRA, these activities are recorded in the relative abundances of the corresponding insertional barcodes in DNA, rather than being measured from active transcription.

In metazoans, a modest number of core signaling pathways are leveraged to give rise to developmental and functional complexity. To demonstrate how ENGRAM can be applied to record the activity of core signaling pathways, we used Wnt and NF-κB-responsive regulatory elements to drive pegRNAs that write to DNA Tape in a quantitative, specific, signal-responsive manner. We further showed how both the intensity and duration of pathway stimulation contribute to observed levels of recording. We also built and characterized a recorder for Tet-On, highlighting the potential of ENGRAM to be used in conjunction with heterologous signal transduction systems. In a multiplex implementation of these three recorders, there was minimal cross-talk, consistent with the expected orthogonality of these signaling pathways to one another.

Finally, we demonstrated a variant of the ENGRAM method in which the recorder comprises an operon of multiple pegRNAs, which are designed to either program or restrict successive edits to the DNA Tape. The resulting pattern of insertional edits allows us to infer the temporal order in which the recorders were activated. Of note, in parallel to this

work, we developed a different strategy for pseudo-processive genome editing called DNA Ticker Tape[131]. In principle, ENGRAM and DNA Ticker Tape are compatible. For the goal of multiplex, temporally resolved recording of core signaling pathway activity over extended periods of time, the combination of ENGRAM and DNA Ticker Tape may be more powerful than the ENGRAM variant described here.

A number of limitations remain to be addressed. First, as our initial attempts to implement ENGRAM exhibited poor signal-to-noise, we sought to reduce background recording by various means, including by expressing Csy4 as part of the recorder, by introducing auto-regulatory negative feedback on Csy4 levels, by integrating recorder constructs to the genome, and by blocking editing with non-targeting pegRNAs during the post-transfection, pre-integration window. Although these strategies were successful, we imagine that further improvements to signal-to-noise might derive from: 1) integrating these and other background reduction strategies; 2) general improvements to prime editing efficiency[125]; and 3) optimization of the specificity of signal-responsive CREs.

Second, here we have only demonstrated ENGRAM recorders for a few hundred enhancer fragments, along with two core (Wnt, NF-κB) and one heterologous (Tet-On) signaling pathway. Looking forward, we envision that many additional such signal-specific recorders can be constructed, validated and optimized. In addition to core signaling pathways, one can also imagine ENGRAM recorders for specific electrical and chemical signals. For pathways for which a signal-responsive, synthetic enhancer is unknown or may not exist, heterologous signal conversion machinery can potentially be constructed and introduced along with the recorder, as we did for Tet-On. Finally, we envision that a set of cell type-specific recorders based on developmental enhancers can be constructed to facilitate recording of the identity of cells' ancestors.

Third, ENGRAM recorders write to a shared DNA Tape (or DNA Ticker Tape), but each unique recorder is presently $\sim$1.3 Kb. As such, although transient transfection of a pool of hundreds to thousands of ENGRAM recorders is straightforward, it is more difficult to imagine how dozens or hundreds of recorders can be concurrently integrated to the genome.

However, we envision that as additional signal-specific recorders are designed and validated, they can be consolidated to a single recorder locus, which can then serve as a common reagent for the multiplex recording of dozens of biological, chemical and electrical signals of interest.

Finally, the deconvolution of ENGRAM signals, especially when recorded to DNA Ticker Tape, will undoubtedly pose some interesting algorithmic challenges. For example, here we show how both the duration and intensity of a signal can contribute to the overall editing rate of a given ENGRAM recorder. If we now imagine recording multiple, fluctuating signals in the context of a dividing and differentiating population of cells, how can we effectively and accurately deconvolve their dynamics?

In summary, ENGRAM is a method for recording specific biological signals to the genome. It is general – any signal that can be converted to Pol-2 mediated transcription can be used to construct an ENGRAM recorder. It is multiplexable – by coupling specific signals to specific insertions, the number of signals that can be encoded grows exponentially with the insertion length. It is quantitative – the strength or duration of signals, and potentially both, can be recorded and recovered. Particularly if combined with DNA Ticker Tape, we envision that ENGRAM can be applied as a means of enriching DNA-based recordings of cellular histories, across state, space and time.

4.5 Materials and Methods

4.5.1 Cell culture, transient transfections and piggyBAC integrations

HEK293T cells (CRL-11268) and K562 cells (CCL-243) were purchased from ATCC. HEK293T cells and K562 cells were cultured in DMEM High glucose (GIBCO) and RPMI 1640 medium (GIBCO), respectively, supplemented with 10% Fetal Bovine Serum (Rocky Mountain Biologicals) and 1% penicillin-streptomycin (GIBCO). Cells were grown with 5% CO2 at $37°C$.

For transient transfections, 1 x 10^5 cells were seeded on a 24-well plate a day before transfection and were transfected with 500 ng plasmid using Lipofectamine 3000 (ThermoFisher

L3000015) following the manufacturer's protocol.

For integrations mediated by the piggyBAC transposon, 1 x 10^5 cells were seeded on a 24-well plate a day before transfection and then transfected with 500 ng cargo plasmid and 200 ng Super piggyBAC transposase expression vector (SBI) using Lipofectamine 3000 following the manufacturer's protocol. Monoclonal lines expressing PE2 were constructed by sorting single cells into 96 wells and selected based on prime editing efficiency.

Most ENGRAM recorders tested in this study were integrated into monoclonal PE2(+) HEK293T cell line via the piggyBac transposon method described above. Of note, for doxycycline recorders, an extra integration was performed to introduce reverse tetracycline-controlled transactivator (rtTA), which is activated by doxycycline and binds to the tetracycline response element to activate downstream recorder expression. For recorders co-transfected with blocking pegRNA plasmid, 200 ng plasmid was added to the 500 ng cargo plasmid and 200 ng piggyBac transposase plasmid.

For ligand recording experiments, 1 x 10^5 cells were seeded on a 48-well plate 6h prior to treatment. 1 ml medium with ligand or negative control was added to each well. For the time series experiment, cells were washed with warm medium and were harvested 24 hours after ligand removal. Doxycycline hyclate (Dox; Sigma-Aldrich D9891) was reconstituted in 1X Phosphate Buffer Solution (PBS) to the final concentration of 10 mg/mL. TNFα (R&D systems, 210-TA-020/CF) was reconstituted in 1 ml PBS to make a 20 μg/ml stock. CHIR-99021 (Selleck, S2924) was purchased as 10 mM stock (1 ml in DMSO). All ligands were stored at $-20°C$. Ligands were thawed immediately before experiments, and diluted with the appropriate culturing medium. The same volume of DMSO or PBS was added to the medium as a negative control.

4.5.2 Library Cloning

The pegRNA-5N recorder (including ENGRAM 1.0, and all three variants of ENGRAM 2.0) was cloned with two steps. First, gene fragment containing CTT pegRNA (Addgene #132778) was PCR amplified using primer sets adding 5-bp degenerate barcode and flanking

BsmBI site for the downstream cloning steps. A carrier plasmid containing two BsmBI sites and two *csy4* hairpins was ordered from Twist. Carrier plasmid and the PCR product from the last step were digested with BsmBI (NEB, buffer 3.1) at $55°C$ for 1h and were purified for ligation. The complete pegRNA with 5N degenerate barcode and *csy4* hairpins was PCR amplified from the ligation product. ENGRAM plasmid and PCR product from above were digested with BsmBI (NEB, buffer 3.1) at $55°C$ for 1h and purified for ligation. Ligation products were purified and resuspended with 5ul H_2O for electroporation. Electroporation was performed using NEB 10-beta Electrocompetent E. coli (C3020) with manufacturer's protocol. Transformed cells were cultured at $30°C$ overnight.

The libraries of 300 enhancers or plasmids bearing signal-responsive elements were cloned in two steps. First, oligos containing enhancer/CRE, two BsmBI restriction site, barcode, 3′ end of pegRNA and *csy4* hairpin were ordered as oPools from IDT. 5′-ENGRAM 2.0 recorder was digested with Xbal and Ncol (NEB, CutSmart buffer) at $37°C$ for 1h and purified. Oligos were cloned into the 5′-ENGRAM2.0 recorder using Gibson assembly. Second, a gene fragment containing minP, *csy4* hairpin, *HEK3* spacer sequence and pegRNA backbone flanking with two BsmBI sites were ordered as gBlock from IDT. gBlock and construct from step1 were digested with BsmBI (NEB, buffer 3.1) at $55°C$ for 1h to generate compatible sticky ends and were purified for ligation. Ligation products were transformed into Stable Competent E.coli (NEB C3040). Transformed cells were cultured at $30°C$ overnight.

All PCR and digestion purification were purified with AMPure XP beads (0.6x for plasmids and 1.2x for fragments with size 200-300 bp) using manufacturer's protocol unless specified. All ligation reactions were using Quick ligase (NEB) with vector:insert ratio 1:6 unless specified. All Gibson reactions were using NEBuilder (NEB) with vector:insert ratio 1:6 unless specified. All plasmid DNA was prepared using a ZymoPURE II Plasmid Kit.

4.5.3 Sequencing Library Generation

Genomic DNA was extracted using protocol as follows: Wash harvested cells with PBS, add 200 μL of freshly prepared lysis buffer (10 mM Tris-HCl, pH 7.5; 0.05% SDS; 25 μg/ml

protease (ThermoFisher)) per 0.5-1M cells directly into each well of the tissue culture plate. The genomic DNA mixture was incubated at $50°C$ for 1 h, followed by an $80°C$ enzyme inactivation step for 30 min.

For each reaction we used 2 μL of cell lysate, 0.25 μL 100mM forward and reverse primer sets, 22.5 μL H_2O and 25 μL Robust HotStart ReadyMix 2x (KAPA Biosystems). PCR reactions were performed as follows: $95°C$ x 3 mins, 22 cycles of ($98°C$ x 20 seconds, $65°C$ x 15 seconds and $72°C$ x 40 seconds). The resulting PCR product was then size-selected using a dual size-selection cleanup of 0.5x and 1x AMPure XP beads (Beckman Coulter) to remove genomic DNA and small fragments (¡200 bp) respectively. This size-selected product was subsequently re-amplified to add flow-cell adapter and sample index for 5 cycles. The final PCR product was cleaned with 0.9x AMPure XP beads (Beckman Coulter). The library was sequenced on an Illumina NextSeq 500 sequencer, an Illumina Miseq sequencer, or an Illumina NextSeq 2000 sequencer following manufacturer's protocol.

4.5.4 Sequence processing pipeline

Sequences were first aligned to *HEK3* target reference using Burrows-Wheeler Aligner software (bwa) with default settings. Aligned reads were then parsed and analyzed for insertion editing efficiencies using pattern-matching functions. For the pool of hexamer barcodes used for enhancer recording, as well as the pentamer barcodes used for signal responsive recording, barcode sequences were chosen to have a Hamming Distance of greater than 2 from all other members of the same set. After extracting barcode sequences from the aligned reads, unexpected barcodes within 1 Hamming Distance from the expected sequences were corrected for insertion counts.

4.5.5 RNA structure prediction and editing score prediction

RNA structure and minimal free energy prediction was performed using the NUPACK python package[132] with default settings. Linear lasso regression model to predict editing score of 5bp barcodes was trained using scikit-learn python package. We defined 85 features

to characterize the 5-bp sequence for which the insertional efficiency is being predicted. These were: 1) Sequence features: 84 binary features corresponding to one-hot encoded sequence, including 20 for single nucleotide content (4 nucleotides * 5 positions) and 64 for dinucleotide content (16 dinucleotides * 4 positions); 2) Structure feature: rescaled minimum free energy within range (0,1). Samples were split with 724 barcodes in a training set and 300 barcodes in a test set. The model was trained with 10-fold cross validation on the training set, and then used to predict the test set.

4.6 Figures and Tables

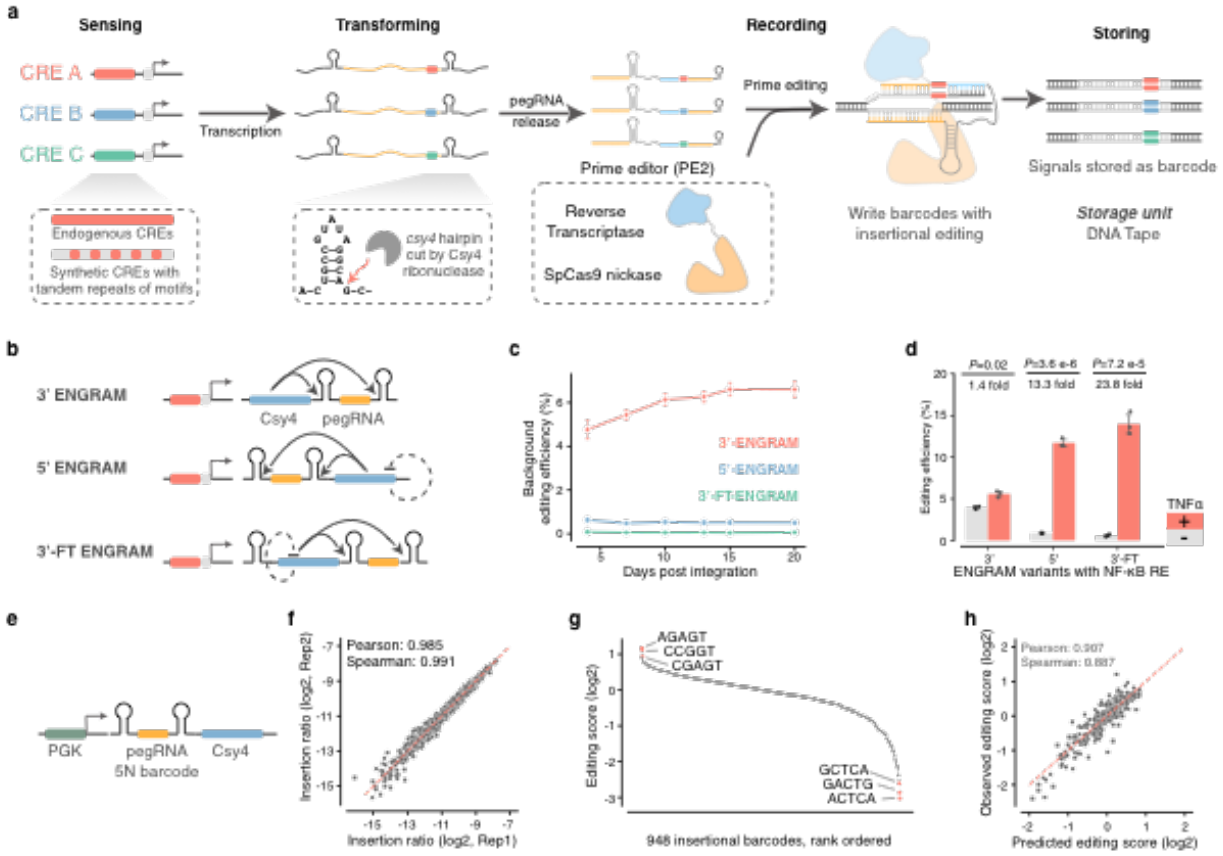

Figure 4.1: ENhancer-driven Genomic Recording of transcriptional Activity in Multiplex (ENGRAM). a. Schematic of ENGRAM. Endogenous or synthetic *cis*-regulatory elements (CREs) drive activity-dependent transcription of a prime editing guide RNA (pegRNA) encoding a CRE-specific insertion. pegRNA is flanked by two 17bp *csy4* hairpin and can be released from pol-2 transcript by Csy4 ribonuclease. Endogenous CREs are sequences with enhancer activity measured by MPRA. Synthetic CREs are tandem repeats of TF motifs. The insertion is written to a natural or synthetic recording site within genomic DNA (DNA Tape). Thus the signal is stored as barcode in the DNA Tape for further readout. **b.** Three versions of ENGRAM 2.0 with the *csy4* hairpin-flanked pegRNA embedded in the 5′ or 3′ UTR of a transcript encoding Csy4, or 3′ ENGRAM 2.0 with an additional *csy4* hairpin in the 5′ UTR in order to impose auto-regulatory negative feedback on Csy4 levels. **c.** All three ENGRAM 2.0 recorders were integrated via piggyBAC into PE2-expressing cells in triplicate, each driving 1,024 5N barcodes with minP. The background editing efficiency was periodically checked over 20 days. Error bars correspond to standard deviations across 3 transfection replicates. **d.** NF-κB response element is cloned to upstream of minP in all three ENGRAM 2.0 recorders. NF-κB responsive ENGRAM recorders were integrated via piggyBAC into PE2-expressing cells. Recording activity was measured in the absence or presence of 10ng/ml of TNFα in triplicate. Both 5′-ENGRAM and 3′-FT ENGRAM showed low background activity and strong activation in response to NF-κB activation, while 3′-ENGRAM showed high background and limited activation. Error bars correspond to standard deviations across 3 replicates. P-values were obtained using the two-tailed Student's t-test. **e.** Schematic of 5N barcode recording. pegRNA encoding degenerate 5N is cloned into 5′-ENGRAM architecture and driven by a PGK promoter. **f.** Log-scaled insertion proportions (calculated as the proportion of edited *HEK3* sites with a given insertion) are highly correlated between transfection replicates. **g.** Range of editing scores (ES) for 5N insertions. ES are calculated as (genomic reads with specific insertion/total edited *HEK3* reads)/(plasmid reads with specific insertion/total plasmid reads), plotted here in rank order on a log2-scale. A total of 948 of 1024 all potential 5N barcodes were recovered after removing under represented barcodes. A few of the highest and lowest ranked insertions are highlighted (sequences shown are those observed in DNA Tape, which are the reverse complement of sequences in pegRNAs). **h.** A linear lasso regression model trained on these data with one-hot encoded single and dinucleotide content of the 5-mer and MFE of secondary structure as features predicts insertional efficiencies with reasonably high accuracy. Samples were split with 680 barcodes in a training set and 268 barcodes in a test set. The model was trained with 10-fold cross validation on the training set, and then used to predict the test set.

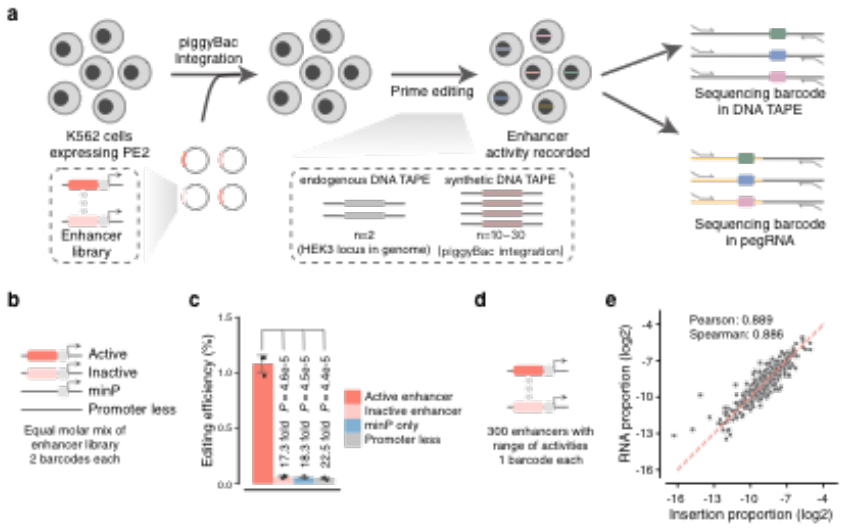

Figure 4.2: Recording enhancer activity with 5′ ENGRAM recorders. a. Schematic of enhancer recording. Enhancer library is cloned to upstream of a minP in 5′-ENGRAM recorders and integrated into PE2+ K562 cells using piggyBac. Enhancer activity can be recorded to endogenous DNA TAPE (genomic *HEK3* locus, n=2) or synthetic DNA TAPE (*HEK3* locus integrated into the genome via piggyBac, n=10-30). **b.** Benchmarking of ENGRAM with enhancers with known activities in a reporter assay. 5′-ENGRAM recorders with active and inactive enhancers upstream of a minP, together with minP-only and promoter-less constructs, were cloned, each driving expression of distinct pegRNA-encoded barcodes. **c.** Barcodes corresponding to the active enhancer showed 17.3, 18.3, and 22.5 fold more abundance than inactive enhancer, minP and promoter less control, respectively. Error bars correspond to standard deviations from 3 transfection replicates. Error bars correspond to standard deviations across 3 transfection replicates. P-values were obtained using the two-tailed Student's t-test. **d, e**. Further benchmarking of ENGRAM 2.0 with 300 enhancers known to have a range of activities in a reporter assay. **d.** This library was designed such that each enhancer drove expression of a distinct pegRNA-encoded 6-mer insertional barcode. **e.** Values correspond to the proportion of each barcode read out from the *HEK3* genomic locus (ENGRAM) or from the pegRNAs (MPRA), out of the total. The log-scaled proportions of ENGRAM events recorded to DNA were highly correlated with log-scaled proportions of barcodes measured directly from RNA.

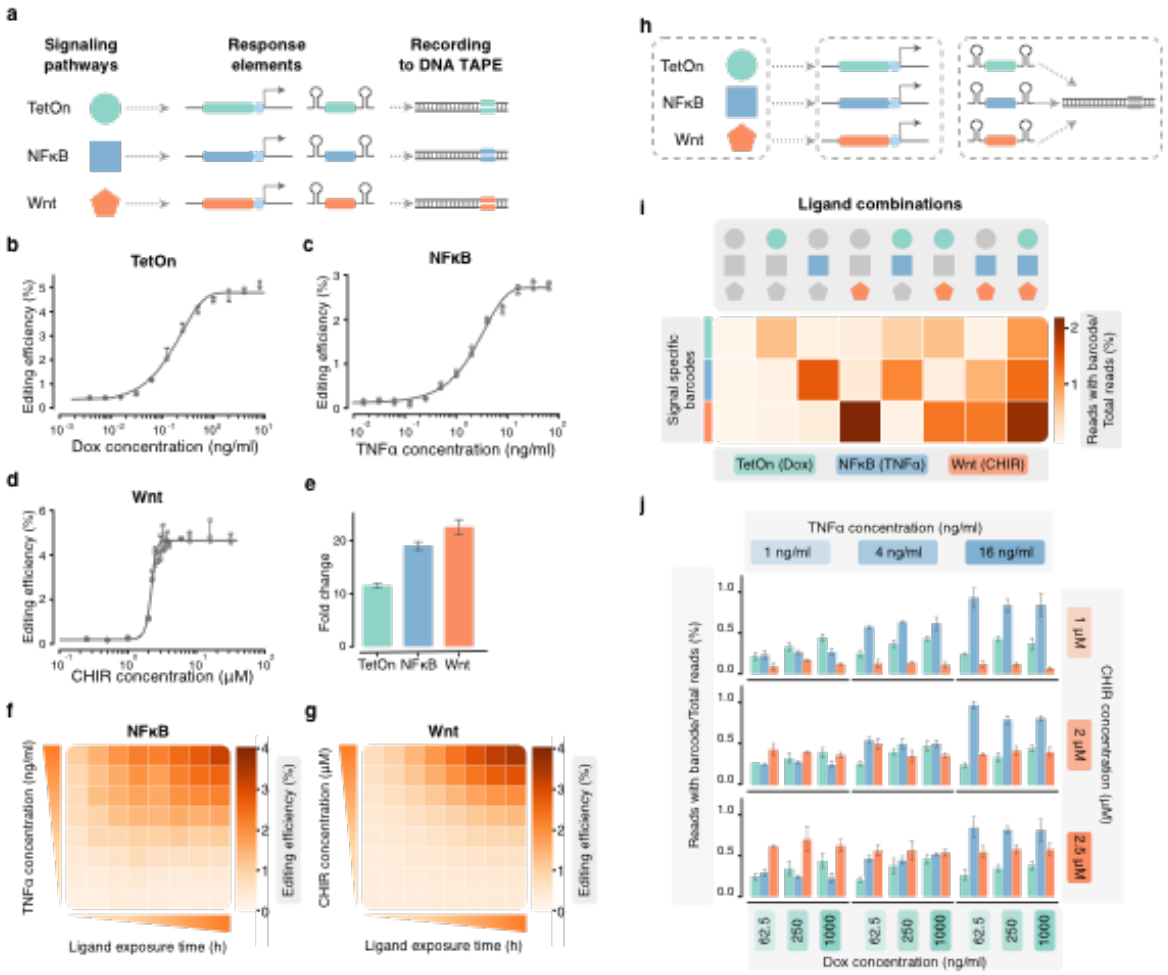
a
Signaling pathways
Response elements
Recording to DNA TAPE
TetOn
NFκB
Wnt
b
TetOn
Editing efficiency (%)
Dox concentration (ng/ml)
c
NFκB
Editing efficiency (%)
TNFα concentration (ng/ml)
d
Wnt
Editing efficiency (%)
CHIR concentration (μM)
e
Fold change
TetOn NFκB Wnt
f
NFκB
TNFα concentration (ng/ml)
Ligand exposure time (h)
Editing efficiency (%)
g
Wnt
CHIR concentration (μM)
Ligand exposure time (h)
Editing efficiency (%)
h
TetOn
NFκB
Wnt
i
Ligand combinations
Signal specific barcodes
Reads with barcode/ Total reads (%)
TetOn (Dox)
NFκB (TNFα)
Wnt (CHIR)
j
TNFα concentration (ng/ml)
1 ng/ml
4 ng/ml
16 ng/ml
Reads with barcode/Total reads (%)
CHIR concentration (μM)
1 μM
2 μM
2.5 μM
Dox concentration (ng/ml)

Figure 4.3: Recording the intensity and duration of signaling pathway activation or small molecule exposure. a. Signal-responsive regulatory elements were used to construct ENGRAM 2.0 recorders for activation by doxycycline (TetON; Tet Response Element), TNFα (a NF-κB responsive element) and CHIR99021 (a TCF-LEF responsive element, responsive to Wnt signaling). (b-d) Upon 48 hours of stimulation with the corresponding stimulant, the TetON **b.**, NF-κB **c.**, and Wnt **d.** recorders exhibited dose-dependent levels of recording. These experiments were conducted on separate, polyclonal cell lines, each of which had one recorder integrated via piggyBAC. Cells were exposed to a serial two-fold dilution series of doxycycline **b.**, TNFα **c.** or CHIR99021 **d.**, with starting concentrations of 8 ng/ml, 64 ng/ml and 32 μM, respectively. For CHIR99021, more concentrations were sampled between 1 to 4 μM **e.** Dynamic range observed in signal recording experiments. Recorders show an 11.5-fold, 19.0-fold and 22.6-fold between activation and background for the Tet, NF-κB and Wnt recorders, respectively. Colors as in panel a. Error bars correspond to standard deviations from 3 stimulus replicates. (f,g) Heatmap showing editing efficiencies resulting from matrix experiment on the NF-κB **f.** and Wnt **g.** recorders, in which both stimulant concentrations and durations of exposure were varied (2 recorders x 8 concentrations x 8 durations x 3 replicates = 384 conditions), illustrating the joint dependence of recording levels on the dose and duration of stimulation. **h.** Schematic of multiplex recording of signaling pathways. Similar to a except that all three recorders are integrated within a single population of cells and are writing to a shared DNA Tape. **i.** Cells bearing multiple recorders were exposed to all possible on/off combinations of three stimuli for 48 hrs, followed by harvesting and sequencing-based quantification of the levels of signal-specific barcodes. Colored shapes as in panel a. Concentrations used were 500 ng/ml, 10 ng/ml and 3 μM for doxycycline, TNFα and CHIR99021, respectively. **j.** Cells bearing multiple recorders were exposed to all possible combinations of high, medium or low concentrations of three stimuli for 48 hrs, followed by harvesting and sequencing-based quantification of the levels of signal-specific barcodes. For Dox, 62.5, 250 or 1000 ng/ml were used; for TNFα, 1, 4 or 16 ng/ml; and for CHIR99021, 1, 2 or 2.5 μM.

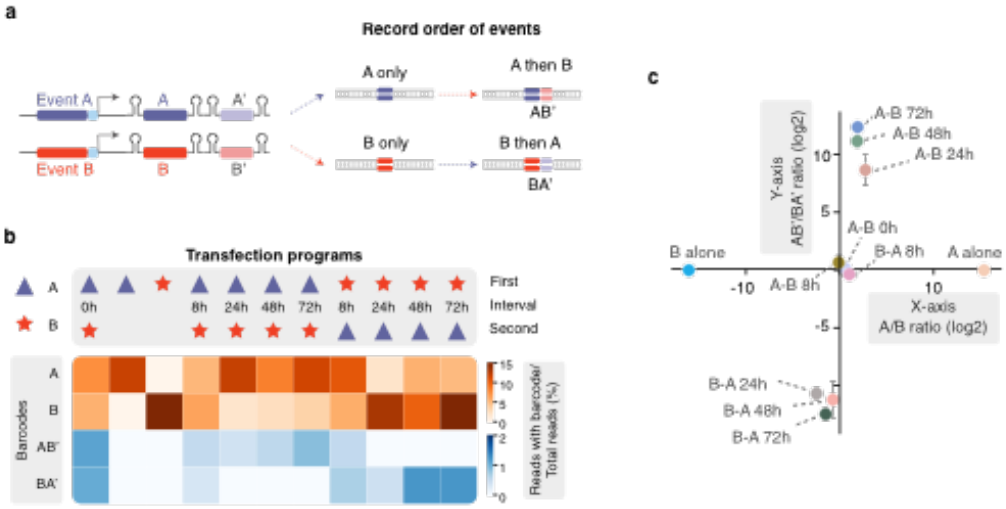

Figure 4.4: Multiplex recording of signaling pathways or the order of signaling events with ENGRAM. a. Strategy for ENGRAM-based recording of the order of events A & B. In brief, each signal-responsive recorder programs the expression of two pegRNAs, one of which targets blank DNA Tape, and the other of which targets DNA Tape that has already been edited in response to the other signal. **b.** We quantified the editing outcomes (A only, B only, A-B' and B-A') associated with 11 transfection programs in which either both A & B were introduced simultaneously (1 program), only A or B was introduced (2 programs), or the recorders were serially transfected with varying recovery periods (A *rightarrow*B or B *rightarrow*A; 8 programs). **c.** The different classes of transfection programs can be distinguished by the ratios of A-B'/B-A' (y-axis) and A/B editing (x-axis) outcomes. Provided at least 24 hrs of recovery between transfections, A *rightarrow*B programs are readily distinguished from B→A programs. Error bars correspond to standard deviations across 3 transfection replicates.

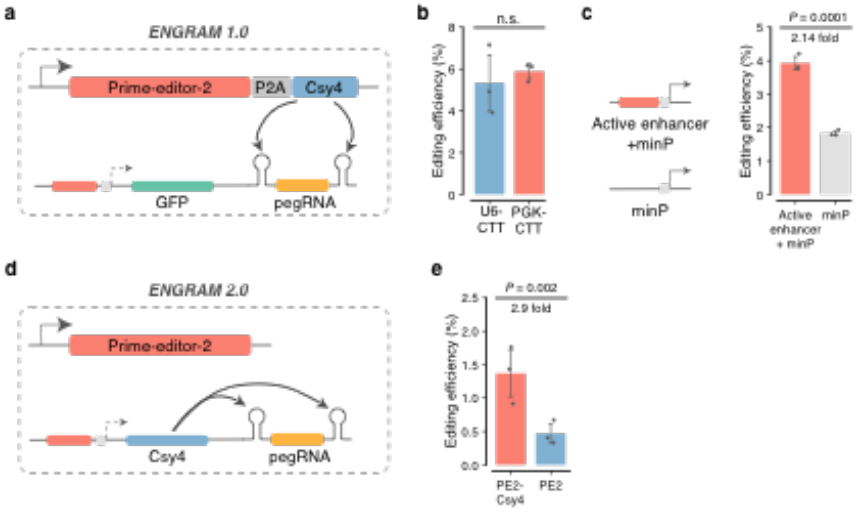

Figure 4.5: The architecture and performance of ENGRAM recorders. a. Schematic of the ENGRAM 1.0 recorder. A pegRNA writing unit is flanked by *csy4* hairpins and embedded within the 3′ UTR of a Pol-2-driven GFP mRNA. PE2 and Csy4 are constitutively expressed from a separate locus. Csy4 cleaves at the *csy4* hairpins and releases the active pegRNA. **b.** Across three transfection replicates, the ENGRAM 1.0 recorder driven by a constitutive Pol-2 PGK promoter (PGK-CTT) exhibited comparable efficiency for inserting CTT at the *HEK3* locus to a U6-driven CTT-pegRNA (U6-CTT). In the K562 cell line in which this experiment was performed, PE2 and Csy4 were constitutively expressed. **c.** A schematic of the constructs used for the two pools of ENGRAM 1.0 recorders is shown on the left, and the observed editing efficiency for each pool on the right. Briefly, a pool of 13 enhancers known to be active in this cell line, cloned upstream of minP and driving a pool of pegRNAs encoding insertion of a 5N degenerate sequence to *HEK3*, was 2.14-fold more active than a control construct bearing minP alone. Error bars correspond to standard deviations across 3 transfection replicates. P-values were obtained using the two-tailed Student's t-test. **d.** Schematic of the ENGRAM 2.0 recorder. A pegRNA writing unit is flanked by *csy4* hairpins and embedded within the 3′ or 5′ UTR of a Pol-2-driven Csy4 mRNA. PE2 is constitutively expressed from a separate locus. Csy4 cleaves at the *csy4* hairpins and releases the active pegRNA. **e.** ENGRAM 2.0 exhibits lower levels of background recording than ENGRAM 1.0. Measurements are for minP alone driving pegRNAs programming a degenerate 5N insertion to the *HEK3* locus in triplicate, 3 days post-transfection. Error bars correspond to standard deviations across 3 transfection replicates. P-values were obtained using the two-tailed Student's t-test.

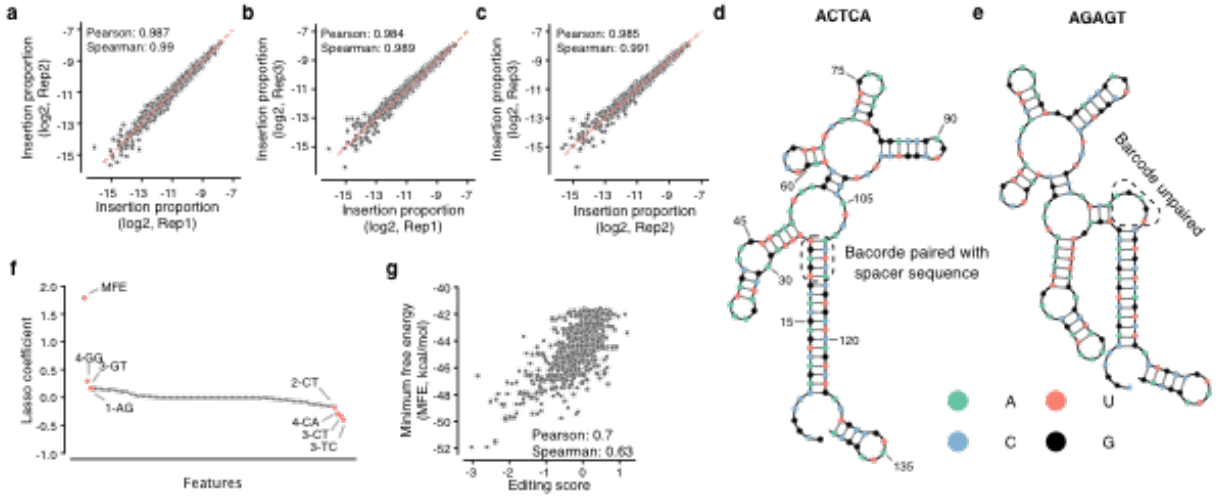

Figure 4.6: The ENGRAM recorder installs barcodes with reasonable efficiency and reproducibility a, c. Reproducibility of the relative proportions of 1023 5N barcodes installed by ENGRAM driven by the constitutive Pol-2 PGK promoter. Log-scaled insertion proportions (calculated as the proportion of edited *HEK3* sites with a given insertion) were well correlated between pairs of transfection replicates. **d, e**. Predicted secondary structures for pegRNAs with the lowest (left) and highest (right) insertional efficiencies. Sequences shown above are those observed in DNA Tape, which are the reverse complement of sequences in pegRNAs. **f.** The rank-ordered coefficients of the linear lasso regression. Positional information of single nucleotides and dinucleotides and minimum free energy (MFE) of secondary structure were used as input features for training. In addition to MFE, which received the highest coefficient, the top 4 and bottom 4 coefficients for sequence features are annotated (*e.g.* 1-A and 3-TC mean A at first nucleotide or TC dinucleotide starting at position 3, respectively). **g.** MEF alone can explain 70% of the variance of the model.

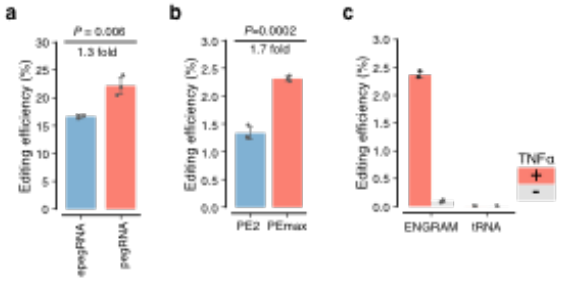

Figure 4.7: ENGRAM recording with new pegRNA and prime editor architecture. **a.** Comparison of recording efficiency between epegRNA and pegRNA. pegRNA/epegRNA encoding 5N degenerate barcode is cloned into 5′-ENGRAM architecture and is driven by a PGK promoter. These two libraries were transiently transfected into PE2+ HEK293T cells separately in triplicate. Genomic DNA was harvested at three days post transfection. Unexpectedly, pegRNA showed 30% higher recording efficiency than epegRNA. We reasoned that *csy4* hairpin might serve a similar role as tevoPreQ1 hairpin to protect pegRNA from degradation, additional hairpin might affect RNA folding. Error bars correspond to standard deviations across 3 transfection replicates. P-values were obtained using the two-tailed Student's t-test. **b.** Comparison of recording efficiency between PE2 and PEmax. PE2/PEmax and PGK-5N-ENGRAM were co-transfected into K562 cells in triplicate. Genomic DNA was harvested at three days post transfection.We observed that PEmax showed 1.7 fold more efficient than PE2. We recommend using PEmax for all future recording assays. In this paper, we used PE2. Error bars correspond to standard deviations across 3 transfection replicates. P-values were obtained using the two-tailed Student's t-test. **c.** tRNA processing for pegRNA release doesn't work in ENGRAM architecture. We replaced *csy4* hairpin with tRNA to see if tRNA can provide an alternative approach for pegRNA releasing. Both ENGRAM pegRNA and tRNA flanked pegRNA encoding 5N degenerate insertion were driven by NF-κB response element. Recorders were integrated into cells via piggyBac. Recording activities were measured in the absence or presence of 10ng/ml TNFα in triplicate. However, tRNA flanked pegRNA failed to show recording activity in both conditions.

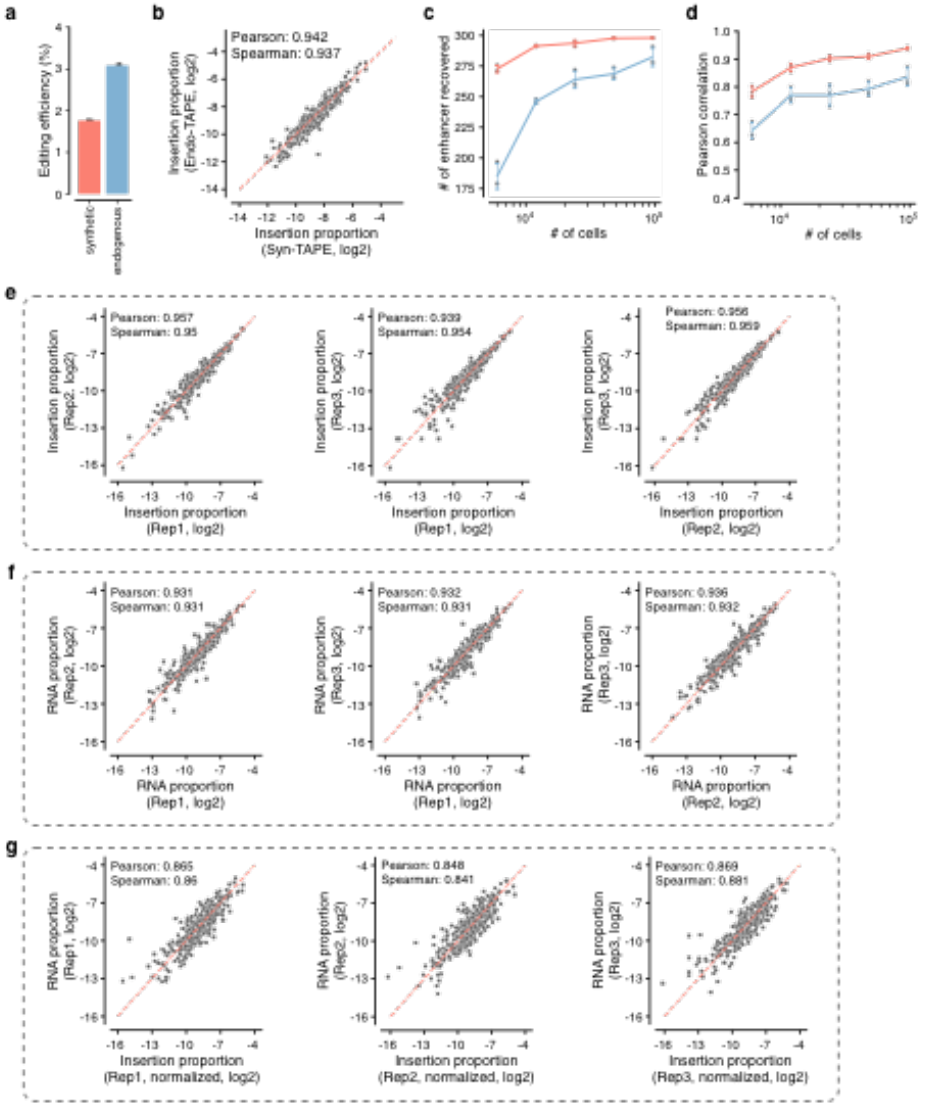

Figure 4.8: Benchmarking of ENGRAM 2.0 recorders. a. Recording efficiency on synthetic *HEK3* locus and endogenous *HEK3* locus. In the same pool of cells, endogenous and synthetic *HEK3* locus show 3.08% and 1.76% overall recording efficiency, respectively. Of note, ∼15 copies of synthetic *HEK3* are integrated. **b.** Log-transformed insertion proportions for 300 6-mer barcodes were highly correlated between synthetic and endogenous *HEK3* locus.**c, d.** Different cell numbers were sampled (6,000, 12,000, 24,000, 48,000, 96,000 cells) on both endogenous and synthetic *HEK3* locus to compare their recording efficiency and sensitivity. Overall, with 12,000 cells, most enhancers can be captured with reasonable reproducibility. **e.** Log-transformed insertion proportions for 300 6-mer barcodes were highly reproducible across transfection replicates. Each value corresponds to the proportion of barcodes read out at the DNA level from the *HEK3* locus. **f.** Log-transformed RNA proportions for 300 6-mer barcodes were highly reproducible across transfection replicates. Each value corresponds to the proportion of barcodes read out at the RNA level from transcribed pegRNAs. **g.** The log-scaled proportions of ENGRAM events recorded to DNA were highly correlated with log-scaled proportions of barcodes measured directly from RNA.

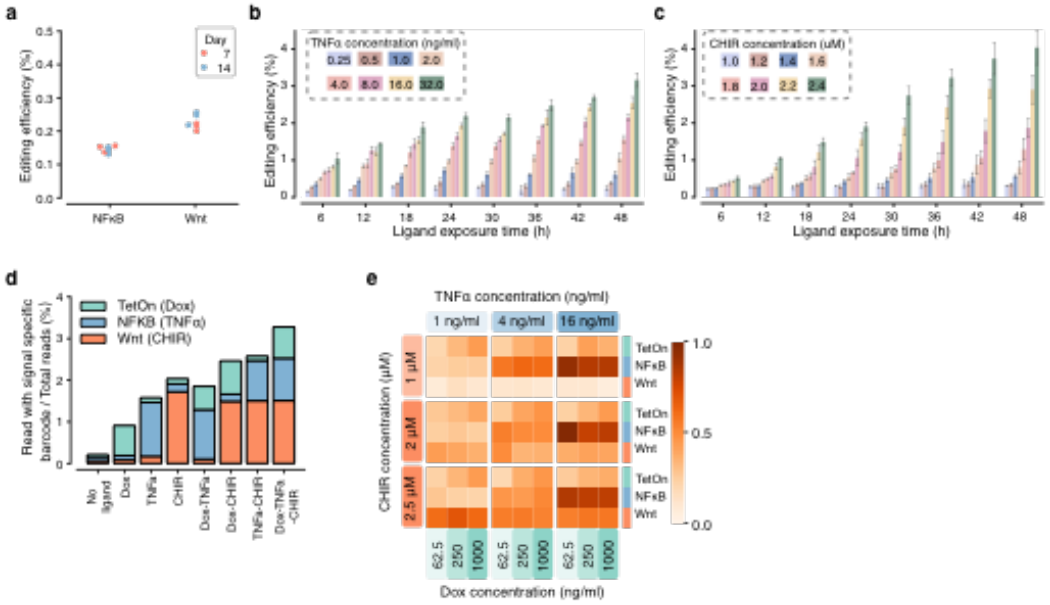

Figure 4.9: Multiplex recording of signaling pathway activation or small molecule exposure with ENGRAM. a. We observed minimal level of background recording in the absence of stimulus with the signal-responsive ENGRAM recorders. This background did not accumulate over time, consistent with the hypothesis that it primarily accumulates shortly after transfection, potentially due to ORI-driven, plasmid-mediated transcription. Plotted points correspond to three transfection replicates. (b-c) Histograms showing editing efficiencies resulting from matrix experiment on the NF-κB **b.** and Wnt **c.** recorders, in which both stimulant concentrations and durations of exposure were varied (2 recorders x 8 concentrations x 8 durations x 3 replicates = 384 conditions). Error bars correspond to standard deviations from 3 stimulus replicates. **d.** Barcode composition of DNA Tape from cells treated with different combinations of stimuli. The recorders exhibit minimal crosstalk between signaling pathways (*e.g.* stimulating with CHIR does not lead to appreciable recording by the NF-κB recorder). **e.** Heatmap visualization of the data shown in **Figure 4.3j**. Levels of recording are informative of each stimulant's concentration, even in the context of concurrent recording of three signals to a shared DNA Tape.

		CRISPR cut	Base editing	Prime editing
Writting	DNA editing modality	Cas9 nuclease	Base editors (dead Cas9 or Cas9 nickase with deaminase)	Prime editor (Cas9 nickase with reservse transcriptase
	Other writing components	gRNA + target (n+n) homing gRNA/stgRNA	gRNA + target (n+n) homing gRNA/stgRNA	pegRNA + target (n+1)
	Information content	pseudorandom ∼10ˆn n=number of target	pseudorandom 2^n n=number of target	short insertions 4^n n=insertion length
	Quantitative Writing	No	No	Yes
Reading	Recovery efficiency	Low (inter target deletion, hard to recover multiple targets)	High	High (single target capture, compatible with scRNA-seq)
	In vitro readout	No	Hard	Yes*
	Read out with RNA-seq	Hard	Hard	Yes
References		McKenna et al. 2016; Chan et al. 2019; Bowling et al. 2020	Tang and Liu 2018; Farzadfard et al. 2019	This paper

Table 4.1: Comparison of different recording syste.

Signaling Pathway	Motif	Repeats	Barcode(s) in pegRNA	Barcode(s) in genomic DNA
TetON	TCCCTATCAGTGATAGAGA	7	AGAAG	CTTCT
NF-κB	GGGACTTTCC	6	TCCTA AATAA	TAGGA TTATT
TCF/LEF (Wnt)	TAAGATCAAAGGG	7	GCAAC	GTTGC

Table 4.2: Responsive elements and their motifs.

Chapter 5

DISCUSSION

A journey of a thousand miles begins with a single step. -Laozi

In this concluding chapter, I will attempt to discuss the applications of recorders in a bigger picture. Acknowledging ENGRAM is only a small step forward toward Everything, everywhere, recorded all at once, I will sketch different potential usage of recorders and how recorders can bring new insights to our understanding of biological functions. In addition to using the recorder as a memory device, due to its feature of altering DNA sequences permanently in a programmable manner, we can also repurpose the recorder as a reprogrammer where pre-programmed cellular circuits can be turned on and off in response to signals by the recorder.

5.1 ENGRAM to study gene regulation and gene expression

5.1.1 Whole genome enhancer mapping with ENGRAM

As ENGRAM adopted the MPRA architecture, its straightforward to use ENGRAM to record more enhancer activities. The reporter assays have long been a golden rule for identifying enhancers. The first enhancer was identified by Banerji *et al.* in 1981[133] by cloning a short sequence from SV40 to the upstream of reporter transcripts to measure its enhanceability. The first generation of reporter assays relies on one designing an RNA probe to target the reporter transcript, which limits its scalability. Exactly 40 years ago in 1982, the reporter gene was replaced by chloramphenicol acetyltransferase (CAT)[134], in which the gene expression abundance is translated into the level of a specific enzyme. Following that, more enzyme-based reporters were developed to further scale up the throughput with luciferase or GFP[135, 136].

5.1.2 Recording gene expression in native loci

The architecture of reporter assays takes the DNA sequence out of its context and it has been debated that positional effect is important for enhancer activity[137]. Can we study the gene expression and enhancer activation in their native locus? This is impossible with traditional reporter assays but ENGRAM has brought new opportunities, allowing us to record and study gene regulation dynamics in native locus. A pegRNA can be integrated into the genome either randomly or with targeted integration. The location of randomly integrated pegRNA can be mapped using our recently developed T7-based genomic mapping assay. A pegRNA is transcribed and released from endogenous gene transcripts. The activity of gene expression could be recorded similarly to other enhancer recording assays. In addition to activities, the order of sequential activation of enhancers can be recorded as well.

5.1.3 Improve ENGRAM for high throughput gene expression study

One limitation of ENGRAM is its efficiency. With only a reasonable 3% overall recording efficiency demonstrated in our enhancer recording assay, its super challenging to test more enhancers or even whole-genome mapping mentioned above. Here, I propose a strategy that can enrich edit cells, allowing enrichment of active enhancers. A frame-shifted selection marker (antibiotic markers, such as Puromycin or Neomycin) can be added to the DNA TAPE, in which the marker is inactive without enhancer activation. With enhancer activation, pegRNA will be transcribed and install a short insertion to the DNA TAPE, restoring the expression of the selection marker. Thus, we are able to enrich cells with only active enhancers.

5.2 ENGRAM to program cell functions

It has been 22 years since the first genetic circuit was developed[1], marking the foundation of the field of synthetic biology. Early works focused on developing genetic circuits in bacteria cells. Only limited work of mammalian synthetic biology has been shown due

to the lack of genetic building blocks. Thanks to the mining of new tools as well as the development of *de novo* protein design in the past few years, more opportunities open up in the mammalian synthetic biology field. For example, orthogonal proteases have been used to program protein-based circuits[138]; Synthetic gene circuits can generate stable memory representing multiple cell states in mammalian cells[139]; Synthetic TF has been developed to target specific sequences to activate gene expression[140]; De novo design of protein binders and switches allows more designs of protein-based circuits [141, 142]; In addition to these amazing work, ENGRAM serves as a versatile tool that can be combined with either protein circuits to activate or deactivate certain proteins or synthetic TF system to record upstream signaling pathways.

5.3 Endmark

ENGRAM provides a unique platform to study fundamental biological questions. Limited by my creativity and imagination, many other interesting topics might be missed. We envision ENGRAM can be used to study cancer metastasis, cell signaling, and more therapeutic fields as sentinel cells to record cellular events. The next decade of recording systems will greatly change our way of doing biological research.

VITA

Wei Chen was born in a small town, JinXiang, in southeast China. Its a unique town with rich culture and own language. 100,000 population and 1 km^2 in size. Grew up with mountains, waters, and stars, Wei has long been interested in science. Wei received his undergrad degree in Biology from Shandong University and his master's degree in Applied and Engineering Physic from Cornell University before starting his Ph.D. in Molecular Engineering & Science at the University of Washington. Outside of research, he is an amateur photographer focusing on landscape and architecture photography. He is an avid reader who reads a lot of history, notifications, and biography. He enjoys both outdoor activities (running, hiking, camping, and sailing) and city life (coffee, food, and museums).

BIBLIOGRAPHY

[1] Michael B Elowitz and Stanislas Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767):335–338, 2000.

[2] Wen Xiong and James E Ferrell, Jr. A positive-feedback-based bistable 'memory module' that governs a cell fate decision. *Nature*, 426(6965):460–465, November 2003.

[3] Joe H Levine, Yihan Lin, and Michael B Elowitz. Functional roles of pulsing in genetic circuits. *Science*, 342(6163):1193–1200, December 2013.

[4] Minhee Park, Nikit Patel, Albert J Keung, and Ahmad S Khalil. Engineering epigenetic regulation using synthetic Read-Write modules. *Cell*, 176(1-2):227–238.e20, January 2019.

[5] James K Nuñez, Jin Chen, Greg C Pommier, J Zachery Cogan, Joseph M Replogle, Carmen Adriaens, Gokul N Ramadoss, Quanming Shi, King L Hung, Avi J Samelson, Angela N Pogson, James Y S Kim, Amanda Chung, Manuel D Leonetti, Howard Y Chang, Martin Kampmann, Bradley E Bernstein, Volker Hovestadt, Luke A Gilbert, and Jonathan S Weissman. Genome-wide programmable transcriptional memory by CRISPR-based epigenome editing. *Cell*, 184(9):2503–2519.e17, April 2021.

[6] Christine A Merrick, Jia Zhao, and Susan J Rosser. Serine integrases: advancing synthetic biology. *ACS synthetic biology*, 7(2):299–310, 2018.

[7] Kai Kretzschmar and Fiona M Watt. Lineage tracing. *Cell*, 148(1-2):33–45, 2012.

[8] Jean Livet, Tamily A Weissman, Hyuno Kang, Ryan W Draft, Ju Lu, Robyn A Bennis, Joshua R Sanes, and Jeff W Lichtman. Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature*, 450(7166):56–62, November 2007.

[9] M Lakso, B Sauer, B Mosinger, Jr, E J Lee, R W Manning, S H Yu, K L Mulder, and H Westphal. Targeted oncogene activation by site-specific recombination in transgenic mice. *Proc. Natl. Acad. Sci. U. S. A.*, 89(14):6232–6236, July 1992.

[10] Lei Yang, Alec AK Nielsen, Jesus Fernandez-Rodriguez, Conor J McClune, Michael T Laub, Timothy K Lu, and Christopher A Voigt. Permanent genetic memory with¿ 1-byte capacity. *Nature methods*, 11(12):1261–1266, 2014.

[11] Ke-Huan K Chow, Mark W Budde, Alejandro A Granados, Maria Cabrera, Shinae Yoon, Soomin Cho, Ting-hao Huang, Noushin Koulena, Kirsten L Frieda, Long Cai, et al. Imaging cell lineage with a synthetic digital recording system. *Science*, 372(6538):eabb3099, 2021.

[12] Nathaniel Roquet, Ava P Soleimany, Alyssa C Ferris, Scott Aaronson, and Timothy K Lu. Synthetic recombinase-based state machines in living cells. *Science*, 353(6297):aad8559, 2016.

[13] Aaron McKenna, Gregory M Findlay, James A Gagnon, Marshall S Horwitz, Alexander F Schier, and Jay Shendure. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science*, 353(6298):aaf7907, July 2016.

[14] Kirsten L Frieda, James M Linton, Sahand Hormoz, Joonhyuk Choi, Ke-Huan K Chow, Zakary S Singer, Mark W Budde, Michael B Elowitz, and Long Cai. Synthetic recording and in situ readout of lineage information in single cells. *Nature*, 541(7635):107–111, January 2017.

[15] Reza Kalhor, Kian Kalhor, Leo Mejia, Kathleen Leeper, Amanda Graveline, Prashant Mali, and George M Church. Developmental barcoding of whole mouse via homing CRISPR. *Science*, 361(6405), August 2018.

[16] Ravi U Sheth and Harris H Wang. DNA-based memory devices for recording cellular events. *Nat. Rev. Genet.*, 19(11):718–732, November 2018.

[17] Weixin Tang and David R Liu. Rewritable multi-event analog recording in bacterial and mammalian cells. *Science*, 360(6385), April 2018.

[18] Fahim Farzadfard, Nava Gharaei, Yasutomi Higashikuni, Giyoung Jung, Jicong Cao, and Timothy K Lu. Single-nucleotide-resolution computing and memory in living cells. *Molecular cell*, 75(4):769–780, 2019.

[19] Seth L Shipman, Jeff Nivala, Jeffrey D Macklis, and George M Church. Molecular recordings by directed CRISPR spacer acquisition. *Science*, 353(6298):aaf1175, July 2016.

[20] Seth L Shipman, Jeff Nivala, Jeffrey D Macklis, and George M Church. CRISPR-Cas encoding of a digital movie into the genomes of a population of living bacteria. *Nature*, 547(7663):345–349, July 2017.

[21] Megan van Overbeek, Daniel Capurso, Matthew M Carter, Matthew S Thompson, Elizabeth Frias, Carsten Russ, John S Reece-Hoyes, Christopher Nye, Scott Gradia,

Bastien Vidal, Jiashun Zheng, Gregory R Hoffman, Christopher K Fuller, and Andrew P May. DNA repair profiling reveals nonrandom outcomes at Cas9-Mediated breaks. *Mol. Cell*, 63(4):633–646, August 2016.

[22] Max W Shen, Mandana Arbab, Jonathan Y Hsu, Daniel Worstell, Sannie J Culbertson, Olga Krabbe, Christopher A Cassa, David R Liu, David K Gifford, and Richard I Sherwood. Predictable and precise template-free CRISPR editing of pathogenic variants. *Nature*, 563(7733):646–651, November 2018.

[23] Felicity Allen, Luca Crepaldi, Clara Alsinet, Alexander J Strong, Vitalii Kleshchevnikov, Pietro De Angeli, Petra Páleníková, Anton Khodak, Vladimir Kiselev, Michael Kosicki, Andrew R Bassett, Heather Harding, Yaron Galanty, Francisco Muñoz-Martínez, Emmanouil Metzakopian, Stephen P Jackson, and Leopold Parts. Predicting the mutations generated by repair of cas9-induced double-strand breaks. *Nat. Biotechnol.*, November 2018.

[24] Massively parallel profiling and predictive modeling of the outcomes of CRISPR/Cas9-mediated double-strand break repair.

[25] Patrick D Hsu, Eric S Lander, and Feng Zhang. Development and applications of CRISPR-Cas9 for genome engineering. *Cell*, 157(6):1262–1278, 2014.

[26] Michael R Lieber. The mechanism of Double-Strand DNA break repair by the nonhomologous DNA End-Joining pathway. *Annu. Rev. Biochem.*, 79(1):181–211, 2010.

[27] Mireille Bétermier, Pascale Bertrand, and Bernard S Lopez. Is non-homologous end-joining really an inherently error-prone process? *PLoS Genet.*, 10(1):e1004086, January 2014.

[28] Agnel Sfeir and Lorraine S Symington. Microhomology-Mediated end joining: A back-up survival mechanism or dedicated pathway? *Trends Biochem. Sci.*, 40(11):701–714, 2015.

[29] Prashant Mali, Luhan Yang, Kevin M Esvelt, John Aach, Marc Guell, James E DiCarlo, Julie E Norville, and George M Church. RNA-guided human genome engineering via cas9. *Science*, 339(6121):823–826, February 2013.

[30] Le Cong, F Ann Ran, David Cox, Shuailiang Lin, Robert Barretto, Naomi Habib, Patrick D Hsu, Xuebing Wu, Wenyan Jiang, Luciano A Marraffini, and Feng Zhang. Multiplex genome engineering using CRISPR/Cas systems. *Science*, 339(6121):819–823, February 2013.

[31] Martin Jinek, Alexandra East, Aaron Cheng, Steven Lin, Enbo Ma, and Jennifer Doudna. RNA-programmed genome editing in human cells. *Elife*, 2:e00471, January 2013.

[32] Tim Wang, Kıvanç Birsoy, Nicholas W Hughes, Kevin M Krupczak, Yorick Post, Jenny J Wei, Eric S Lander, and David M Sabatini. Identification and characterization of essential genes in the human genome. *Science*, 350(6264):1096–1101, November 2015.

[33] Molly Gasperini, Gregory M Findlay, Aaron McKenna, Jennifer H Milbank, Choli Lee, Melissa D Zhang, Darren A Cusanovich, and Jay Shendure. CRISPR/Cas9-Mediated scanning for regulatory elements required for HPRT1 expression via thousands of large, programmed genomic deletions. *Am. J. Hum. Genet.*, 101(2):192–205, August 2017.

[34] Jason C Klein, Wei Chen, Molly Gasperini, and Jay Shendure. Identifying novel enhancer elements with CRISPR-Based screens. *ACS Chem. Biol.*, 13(2):326–332, February 2018.

[35] Brandon J Aubrey, Gemma L Kelly, Andrew J Kueh, Margs S Brennan, Liam O'Connor, Liz Milla, Stephen Wilcox, Lin Tai, Andreas Strasser, and Marco J Herold. An inducible lentiviral guide RNA platform enables the identification of tumor-essential genes and tumor-promoting mutations in vivo. *Cell Rep.*, 10(8):1422–1432, March 2015.

[36] Howard H Y Chang, Go Watanabe, Christina A Gerodimos, Takashi Ochi, Tom L Blundell, Stephen P Jackson, and Michael R Lieber. Different DNA end configurations dictate which NHEJ components are most important for joining efficiency. *J. Biol. Chem.*, 291(47):24377–24389, November 2016.

[37] Anne Bothmer, Tanushree Phadke, Luis A Barrera, Carrie M Margulies, Christina S Lee, Frank Buquicchio, Sean Moss, Hayat S Abdulkerim, William Selleck, Hariharan Jayaram, Vic E Myer, and Cecilia Cotta-Ramusino. Characterization of the interplay between DNA repair and CRISPR/Cas9-induced DNA lesions at an endogenous locus. *Nat. Commun.*, 8:13905, January 2017.

[38] Maximilian Haeussler, Kai Schönig, Hélène Eckert, Alexis Eschstruth, Joffrey Mianné, Jean-Baptiste Renaud, Sylvie Schneider-Maunoury, Alena Shkumatava, Lydia Teboul, Jim Kent, Jean-Stephane Joly, and Jean-Paul Concordet. Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol.*, 17(1):148, July 2016.

[39] Laura Magill Sack, Teresa Davoli, Qikai Xu, Mamie Z Li, and Stephen J Elledge. Sources of error in mammalian genetic screens. *G3*, 6(9):2781–2790, September 2016.

[40] Andrew J Hill, José L McFaline-Figueroa, Lea M Starita, Molly J Gasperini, Kenneth A Matreyek, Jonathan Packer, Dana Jackson, Jay Shendure, and Cole Trapnell. On the design of CRISPR-based single-cell molecular screens. *Nat. Methods*, 15(4):271–274, April 2018.

[41] S B Needleman and C D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48(3):443–453, March 1970.

[42] Michael Kosicki, Kärt Tomberg, and Allan Bradley. Repair of double-strand breaks induced by CRISPR–Cas9 leads to large deletions and complex rearrangements. *Nat. Biotechnol.*, 2018.

[43] Brenda R Lemos, Adam C Kaplan, Ji Eun Bae, Alexander E Ferrazzoli, James Kuo, Ranjith P Anand, David P Waterman, and James E Haber. CRISPR/Cas9 cleavages in budding yeast reveal templated insertions and strand-specific insertion/deletion profiles. *Proc. Natl. Acad. Sci. U. S. A.*, 115(9):E2040–E2047, February 2018.

[44] Xuebing Wu and David P Bartel. kplogo: positional k-mer analysis reveals hidden specificity in biological sequences. *Nucleic Acids Res.*, 45(W1):W534–W538, July 2017.

[45] Martin Jinek, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A Doudna, and Emmanuelle Charpentier. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, 337(6096):816–821, August 2012.

[46] Anthony A Stephenson, Austin T Raper, and Zucai Suo. Bidirectional degradation of DNA cleavage products catalyzed by CRISPR/Cas9. *J. Am. Chem. Soc.*, 140(10):3743–3750, March 2018.

[47] Ludovic Deriano and David B Roth. Modernizing the nonhomologous end-joining repertoire: alternative and classical NHEJ share the stage. *Annu. Rev. Genet.*, 47:433–455, September 2013.

[48] Nicholas R Pannunzio, Sicong Li, Go Watanabe, and Michael R Lieber. Non-homologous end joining often uses microhomology: implications for alternative end joining. *DNA Repair*, 17:74–80, May 2014.

[49] Michael P Conlin, Dylan A Reid, George W Small, Howard H Chang, Go Watanabe, Michael R Lieber, Dale A Ramsden, and Eli Rothenberg. DNA ligase IV guides End-Processing choice during nonhomologous end joining. *Cell Rep.*, 20(12):2810–2819, September 2017.

[50] Fuguo Jiang, David W Taylor, Janice S Chen, Jack E Kornfeld, Kaihong Zhou, Aubri J Thompson, Eva Nogales, and Jennifer A Doudna. Structures of a CRISPR-Cas9 r-loop complex primed for DNA cleavage. *Science*, 351(6275):867–871, February 2016.

[51] Felicity Allen, Luca Crepaldi, Clara Alsinet, Alexander J Strong, Vitalii Kleshchevnikov, Pietro De Angeli, Petra Páleníková, Anton Khodak, Vladimir Kiselev, Michael Kosicki, Andrew R Bassett, Heather Harding, Yaron Galanty, Francisco Muñoz-Martínez, Emmanouil Metzakopian, Stephen P Jackson, and Leopold Parts. Predicting the mutations generated by repair of cas9-induced double-strand breaks. *Nat. Biotechnol.*, November 2018.

[52] Anob M Chakrabarti, Tristan Henser-Brownhill, Josep Monserrat, Anna R Poetsch, Nicholas M Luscombe, and Paola Scaffidi. Target-Specific precision of CRISPR-Mediated genome editing. *Molecular cell*, December 2018.

[53] Sangsu Bae, Jiyeon Kweon, Heon Seok Kim, and Jin-Soo Kim. Microhomology-based choice of cas9 nuclease target sites. *Nat. Methods*, 11(7):705–706, July 2014.

[54] Shang-Hsun Yang, Pei-Hsun Cheng, Robert T Sullivan, James W Thomas, and Anthony W S Chan. Lentiviral integration preferences in transgenic mice. *Genesis*, 46(12):711–718, December 2008.

[55] Duran Ustek, Sema Sirma, Ergun Gumus, Muzaffer Arikan, Aris Cakiris, Neslihan Abaci, Jaicy Mathew, Zeliha Emrence, Hulya Azakli, Fulya Cosan, Atilla Cakar, Mahmut Parlak, and Olcay Kursun. A genome-wide analysis of lentivector integration sites using targeted sequence capture and next generation sequencing technology. *Infect. Genet. Evol.*, 12(7):1349–1354, October 2012.

[56] Max A Horlbeck, Lea B Witkowsky, Benjamin Guglielmi, Joseph M Replogle, Luke A Gilbert, Jacqueline E Villalta, Sharon E Torigoe, Robert Tjian, and Jonathan S Weissman. Nucleosomes impede cas9 access to DNA in vivo and in vitro. *Elife*, 5, March 2016.

[57] Xiaoyu Chen, Marrit Rinsma, Josephine M Janssen, Jin Liu, Ignazio Maggio, and Manuel A F V Gonçalves. Probing the impact of chromatin conformation on genome editing tools. *Nucleic Acids Res.*, 44(13):6482–6492, July 2016.

[58] Eirini M Kallimasioti-Pazi, Keerthi Thelakkad Chathoth, Gillian C Taylor, Alison Meynert, Tracy Ballinger, Martijn JE Kelder, Sébastien Lalevée, Ildem Sanli, Robert Feil, and Andrew J Wood. Heterochromatin delays crispr-cas9 mutagenesis but does not influence the outcome of mutagenic dna repair. *PLoS Biology*, 16(12):e2005595, 2018.

[59] F Ann Ran, Patrick D Hsu, Chie-Yu Lin, Jonathan S Gootenberg, Silvana Konermann, Alexandro E Trevino, David A Scott, Azusa Inoue, Shogo Matoba, Yi Zhang, and Feng Zhang. Double nicking by RNA-guided CRISPR cas9 for enhanced genome editing specificity. *Cell*, 154(6):1380–1389, September 2013.

[60] Shengdar Q Tsai, Nicolas Wyvekens, Cyd Khayter, Jennifer A Foden, Vishal Thapar, Deepak Reyon, Mathew J Goodwin, Martin J Aryee, and J Keith Joung. Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nat. Biotechnol.*, 32(6):569–576, June 2014.

[61] John P Guilinger, David B Thompson, and David R Liu. Fusion of catalytically inactive cas9 to FokI nuclease improves the specificity of genome modification. *Nat. Biotechnol.*, 32(6):577–582, June 2014.

[62] Jason M Wolfs, Thomas A Hamilton, Jeremy T Lant, Marcon Laforet, Jenny Zhang, Louisa M Salemi, Gregory B Gloor, Caroline Schild-Poulter, and David R Edgell. Biasing genome-editing events toward precise length deletions with an RNA-guided TevCas9 dual nuclease. *Proc. Natl. Acad. Sci. U. S. A.*, 113(52):14988–14993, December 2016.

[63] Mehmet Fatih Bolukbasi, Pengpeng Liu, Kevin Luk, Samantha F Kwok, Ankit Gupta, Nadia Amrani, Erik J Sontheimer, Lihua Julie Zhu, and Scot A Wolfe. Orthogonal cas9–cas9 chimeras provide a versatile platform for genome editing. *Nature communications*, 9(1):1–12, 2018.

[64] Aaron McKenna and Jay Shendure. FlashFry: a fast and flexible tool for large-scale CRISPR target design. *BMC Biol.*, 16(1):74, July 2018.

[65] Patrick D Hsu, David A Scott, Joshua A Weinstein, F Ann Ran, Silvana Konermann, Vineeta Agarwala, Yinqing Li, Eli J Fine, Xuebing Wu, Ophir Shalem, Thomas J Cradick, Luciano A Marraffini, Gang Bao, and Feng Zhang. DNA targeting specificity of RNA-guided cas9 nucleases. *Nat. Biotechnol.*, 31(9):827–832, September 2013.

[66] Martin Šošić and Mile Šikić. Edlib: a C/C library for fast, exact sequence alignment using edit distance. *Bioinformatics*, 33(9):1394–1395, 2017.

[67] Jiajie Zhang, Kassian Kobert, Tomáš Flouri, and Alexandros Stamatakis. PEAR: a fast and accurate illumina Paired-End read merger. *Bioinformatics*, 30(5):614–620, March 2014.

[68] Tom Smith, Andreas Heger, and Ian Sudbery. UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res.*, 27(3):491–499, March 2017.

[69] Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 03 2009.

[70] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. December 2014.

[71] Francois Chollet et al. Keras, 2015.

[72] Rami Al-Rfou, Guillaume Alain, Amjad Almahairi, Christof Angermueller, Dzmitry Bahdanau, Nicolas Ballas, Frédéric Bastien, Justin Bayer, Anatoly Belikov, Alexander Belopolsky, Yoshua Bengio, Arnaud Bergeron, James Bergstra, Valentin Bisson, Josh Bleecher Snyder, Nicolas Bouchard, Nicolas Boulanger-Lewandowski, Xavier Bouthillier, Alexandre de Brébisson, Olivier Breuleux, Pierre-Luc Carrier, Kyunghyun Cho, Jan Chorowski, Paul Christiano, Tim Cooijmans, Marc-Alexandre Côté, Myriam Côté, Aaron Courville, Yann N. Dauphin, Olivier Delalleau, Julien Demouth, Guillaume Desjardins, Sander Dieleman, Laurent Dinh, Mélanie Ducoffe, Vincent Dumoulin, Samira Ebrahimi Kahou, Dumitru Erhan, Ziye Fan, Orhan Firat, Mathieu Germain, Xavier Glorot, Ian Goodfellow, Matt Graham, Caglar Gulcehre, Philippe Hamel, Iban Harlouchet, Jean-Philippe Heng, Balázs Hidasi, Sina Honari, Arjun Jain, Sébastien Jean, Kai Jia, Mikhail Korobov, Vivek Kulkarni, Alex Lamb, Pascal Lamblin, Eric Larsen, César Laurent, Sean Lee, Simon Lefrancois, Simon Lemieux, Nicholas Léonard, Zhouhan Lin, Jesse A. Livezey, Cory Lorenz, Jeremiah Lowin, Qianli Ma, Pierre-Antoine Manzagol, Olivier Mastropietro, Robert T. McGibbon, Roland Memisevic, Bart van Merriënboer, Vincent Michalski, Mehdi Mirza, Alberto Orlandi, Christopher Pal, Razvan Pascanu, Mohammad Pezeshki, Colin Raffel, Daniel Renshaw, Matthew Rocklin, Adriana Romero, Markus Roth, Peter Sadowski, John Salvatier, François Savard, Jan Schlüter, John Schulman, Gabriel Schwartz, Iulian Vlad Serban, Dmitriy Serdyuk, Samira Shabanian, Étienne Simon, Sigurd Spieckermann, S. Ramana Subramanyam, Jakub Sygnowski, Jérémie Tanguay, Gijs van Tulder, Joseph Turian, Sebastian Urban, Pascal Vincent, Francesco Visin, Harm de Vries, David Warde-Farley, Dustin J. Webb, Matthew Willson, Kelvin Xu, Lijun Xue, Li Yao, Saizheng Zhang, and Ying Zhang. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016.

[73] Gavin J Knott and Jennifer A Doudna. Crispr-cas guides the future of genetic engineering. *Science*, 361(6405):866–869, 2018.

[74] Le Cong, F Ann Ran, David Cox, Shuailiang Lin, Robert Barretto, Naomi Habib, Patrick D Hsu, Xuebing Wu, Wenyan Jiang, Luciano A Marraffini, and Feng Zhang.

Multiplex genome engineering using CRISPR/Cas systems. *Science*, 339(6121):819–823, February 2013.

[75] Matthew C Canver, Daniel E Bauer, Abhishek Dass, Yvette Y Yien, Jacky Chung, Takeshi Masuda, Takahiro Maeda, Barry H Paw, and Stuart H Orkin. Characterization of genomic deletion efficiency mediated by clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9 nuclease system in mammalian cells. *J. Biol. Chem.*, 289(31):21312–21324, August 2014.

[76] Susan M Byrne, Luis Ortiz, Prashant Mali, John Aach, and George M Church. Multi-kilobase homozygous targeted gene replacement in human induced pluripotent stem cells. *Nucleic Acids Res.*, 43(3):e21, February 2015.

[77] Molly Gasperini, Gregory M Findlay, Aaron McKenna, Jennifer H Milbank, Choli Lee, Melissa D Zhang, Darren A Cusanovich, and Jay Shendure. CRISPR/Cas9-Mediated scanning for regulatory elements required for HPRT1 expression via thousands of large, programmed genomic deletions. *Am. J. Hum. Genet.*, 101(2):192–205, August 2017.

[78] Molly Gasperini, Andrew J Hill, José L McFaline-Figueroa, Beth Martin, Seungsoo Kim, Melissa D Zhang, Dana Jackson, Anh Leith, Jacob Schreiber, William S Noble, Cole Trapnell, Nadav Ahituv, and Jay Shendure. A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell*, 176(6):1516, March 2019.

[79] Michael Kosicki, Kärt Tomberg, and Allan Bradley. Repair of double-strand breaks induced by CRISPR-Cas9 leads to large deletions and complex rearrangements. *Nat. Biotechnol.*, 36(8):765–771, September 2018.

[80] Michael V Zuccaro, Jia Xu, Carl Mitchell, Diego Marin, Raymond Zimmerman, Bhavini Rana, Everett Weinstein, Rebeca T King, Katherine L Palmerola, Morgan E Smith, Stephen H Tsang, Robin Goland, Maria Jasin, Rogerio Lobo, Nathan Treff, and Dieter Egli. Allele-Specific chromosome removal after cas9 cleavage in human embryos. *Cell*, October 2020.

[81] Anuja Mehta and James E Haber. Sources of DNA double-strand breaks and models of recombinational DNA repair. *Cold Spring Harb. Perspect. Biol.*, 6(9):a016428, August 2014.

[82] Yarui Diao, Rongxin Fang, Bin Li, Zhipeng Meng, Juntao Yu, Yunjiang Qiu, Kimberly C Lin, Hui Huang, Tristin Liu, Ryan J Marina, Inkyung Jung, Yin Shen, Kun-Liang Guan, and Bing Ren. A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nat. Methods*, 14(6):629–635, June 2017.

[83] Shiyou Zhu, Wei Li, Jingze Liu, Chen-Hao Chen, Qi Liao, Ping Xu, Han Xu, Tengfei Xiao, Zhongzheng Cao, Jingyu Peng, Pengfei Yuan, Myles Brown, Xiaole Shirley Liu, and Wensheng Wei. Genome-scale deletion screening of human long non-coding RNAs using a paired-guide RNA CRISPR-Cas9 library. *Nat. Biotechnol.*, 34(12):1279–1286, December 2016.

[84] Mohammad Ali Khosravi, Maryam Abbasalipour, Jean-Paul Concordet, Johannes Vom Berg, Sirous Zeinali, Arash Arashkia, Kayhan Azadmanesh, Thorsten Buch, and Morteza Karimipoor. Targeted deletion of BCL11A gene by CRISPR-Cas9 system for fetal hemoglobin reactivation: A promising approach for gene therapy of beta thalassemia disease. *Eur. J. Pharmacol.*, 854:398–405, July 2019.

[85] Sumitava Dastidar, Simon Ardui, Kshitiz Singh, Debanjana Majumdar, Nisha Nair, Yanfang Fu, Deepak Reyon, Ermira Samara, Mattia F M Gerli, Arnaud F Klein, Wito De Schrijver, Jaitip Tipanee, Sara Seneca, Warut Tulalamba, Hui Wang, Yoke Chin Chai, Peter In't Veld, Denis Furling, Francesco Saverio Tedesco, Joris R Vermeesch, J Keith Joung, Marinee K Chuah, and Thierry VandenDriessche. Efficient CRISPR/Cas9-mediated editing of trinucleotide repeat expansion in myotonic dystrophy patient-derived iPS and myogenic cells. *Nucleic Acids Res.*, 46(16):8275–8298, September 2018.

[86] Andrew V Anzalone, Peyton B Randolph, Jessie R Davis, Alexander A Sousa, Luke W Koblan, Jonathan M Levy, Peter J Chen, Christopher Wilson, Gregory A Newby, Aditya Raguram, and David R Liu. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature*, 576(7785):149–157, December 2019.

[87] Qiupeng Lin, Yuan Zong, Chenxiao Xue, Shengxing Wang, Shuai Jin, Zixu Zhu, Yanpeng Wang, Andrew V Anzalone, Aditya Raguram, Jordan L Doman, David R Liu, and Caixia Gao. Prime genome editing in rice and wheat. *Nat. Biotechnol.*, 38(5):582–585, May 2020.

[88] Teemu Kivioja, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson, and Jussi Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods*, 9(1):72–74, November 2011.

[89] Dan Dominissini, Sharon Moshitch-Moshkovitz, Schraga Schwartz, Mali Salmon-Divon, Lior Ungar, Sivan Osenberg, Karen Cesarkas, Jasmine Jacob-Hirsch, Ninette Amariglio, Martin Kupiec, Rotem Sorek, and Gideon Rechavi. Topology of the human and mouse m6a RNA methylomes revealed by m6a-seq. *Nature*, 485(7397):201–206, April 2012.

[90] Hannah L Watry, Carissa M Feliciano, Ketrin Gjoni, Gou Takahashi, Yuichiro Miyaoka, Bruce R Conklin, and Luke M Judge. Rapid, precise quantification of large DNA excisions and inversions by ddPCR. *Sci. Rep.*, 10(1):14896, September 2020.

[91] Pankaj K Mandal, Leonardo M R Ferreira, Ryan Collins, Torsten B Meissner, Christian L Boutwell, Max Friesen, Vladimir Vrbanac, Brian S Garrison, Alexei Stortchevoi, David Bryder, Kiran Musunuru, Harrison Brand, Andrew M Tager, Todd M Allen, Michael E Talkowski, Derrick J Rossi, and Chad A Cowan. Efficient ablation of genes in human hematopoietic stem and effector cells using CRISPR/Cas9. *Cell Stem Cell*, 15(5):643–652, November 2014.

[92] Russell T Walton, Kathleen A Christie, Madelynn N Whittaker, and Benjamin P Kleinstiver. Unconstrained genome targeting with near-PAMless engineered CRISPR-Cas9 variants. *Science*, 368(6488):290–296, April 2020.

[93] Jiyeon Kweon, Jung-Ki Yoon, An-Hee Jang, Ha Rim Shin, Ji-Eun See, Gayoung Jang, Jong-Il Kim, and Yongsub Kim. Engineered prime editors with PAM flexibility. *Mol. Ther.*, February 2021.

[94] Mitchell L Leibowitz, Stamatis Papathanasiou, Phillip A Doerfler, Logan J Blaine, Lili Sun, Yu Yao, Cheng-Zhong Zhang, Mitchell J Weiss, and David Pellman. Chromothripsis as an on-target consequence of crispr–cas9 genome editing. *Nature genetics*, 53(6):895–905, 2021.

[95] Imre F Schene, Indi P Joore, Rurika Oka, Michal Mokry, Anke H M van Vugt, Ruben van Boxtel, Hubert P J van der Doef, Luc J W van der Laan, Monique M A Verstegen, Peter M van Hasselt, Edward E S Nieuwenhuis, and Sabine A Fuchs. Prime editing for functional repair in patient-derived disease models. *Nat. Commun.*, 11(1):5352, October 2020.

[96] Dominic D G Owens, Adam Caulder, Vincent Frontera, Joe R Harman, Alasdair J Allan, Akin Bucakci, Lucas Greder, Gemma F Codner, Philip Hublitz, Peter J McHugh, Lydia Teboul, and Marella F T R de Bruijn. Microhomologies are prevalent at cas9-induced larger deletions. *Nucleic Acids Res.*, 47(14):7402–7417, August 2019.

[97] Do Yon Kim, Su Bin Moon, Jeong-Heon Ko, Yong-Sam Kim, and Daesik Kim. Unbiased investigation of specificities of prime editing systems in human cells. *Nucleic acids research*, 48(18):10576–10589, 2020.

[98] Mohamed A El-Brolosy, Zacharias Kontarakis, Andrea Rossi, Carsten Kuenne, Stefan Günther, Nana Fukuda, Khrievono Kikhi, Giulia L M Boezio, Carter M Takacs, Shih-Lei Lai, Ryuichi Fukuda, Claudia Gerri, Antonio J Giraldez, and Didier Y R

Stainier. Genetic compensation triggered by mutant mRNA degradation. *Nature*, 568(7751):193–197, April 2019.

[99] Zhipeng Ma, Peipei Zhu, Hui Shi, Liwei Guo, Qinghe Zhang, Yanan Chen, Shuming Chen, Zhe Zhang, Jinrong Peng, and Jun Chen. PTC-bearing mRNA elicits a genetic compensation response via upf3a and COMPASS components. *Nature*, 568(7751):259–263, April 2019.

[100] Jean-Paul Concordet and Maximilian Haeussler. CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Res.*, 46(W1):W242–W245, July 2018.

[101] John G Doench, Nicolo Fusi, Meagan Sullender, Mudra Hegde, Emma W Vaimberg, Katherine F Donovan, Ian Smith, Zuzana Tothova, Craig Wilen, Robert Orchard, Herbert W Virgin, Jennifer Listgarten, and David E Root. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.*, 34(2):184–191, February 2016.

[102] Hui Kwon Kim, Goosang Yu, Jinman Park, Seonwoo Min, Sungtae Lee, Sungroh Yoon, and Hyongbum Henry Kim. Predicting the efficiency of prime editing guide RNAs in human cells. *Nat. Biotechnol.*, September 2020.

[103] Aaron McKenna and Jay Shendure. FlashFry: a fast and flexible tool for large-scale CRISPR target design. *BMC Biol.*, 16(1):74, July 2018.

[104] Patrick D Hsu, David A Scott, Joshua A Weinstein, F Ann Ran, Silvana Konermann, Vineeta Agarwala, Yinqing Li, Eli J Fine, Xuebing Wu, Ophir Shalem, Thomas J Cradick, Luciano A Marraffini, Gang Bao, and Feng Zhang. DNA targeting specificity of RNA-guided cas9 nucleases. *Nat. Biotechnol.*, 31(9):827–832, September 2013.

[105] Jiajie Zhang, Kassian Kobert, Tomáš Flouri, and Alexandros Stamatakis. PEAR: a fast and accurate illumina Paired-End read merger. *Bioinformatics*, 30(5):614–620, March 2014.

[106] Kendell Clement, Holly Rees, Matthew C Canver, Jason M Gehrke, Rick Farouni, Jonathan Y Hsu, Mitchel A Cole, David R Liu, J Keith Joung, Daniel E Bauer, and Luca Pinello. CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nat. Biotechnol.*, 37(3):224–226, March 2019.

[107] Pulin Li and Michael B Elowitz. Communication codes in developmental signaling pathways. *Development*, 146(12), June 2019.

[108] Rupali P Patwardhan, Choli Lee, Oren Litvin, David L Young, Dana Pe'er, and Jay Shendure. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.*, 27(12):1173–1175, December 2009.

[109] David P Doupé and Norbert Perrimon. Visualizing and manipulating temporal signaling dynamics with fluorescence-based tools. *Sci. Signal.*, 7(319):re1, April 2014.

[110] Fumitaka Inoue, Anat Kreimer, Tal Ashuach, Nadav Ahituv, and Nir Yosef. Identification and massively parallel characterization of regulatory elements driving neural induction. *Cell Stem Cell*, 25(5):713–727.e10, November 2019.

[111] Ravi U Sheth, Sung Sun Yim, Felix L Wu, and Harris H Wang. Multiplex recording of cellular events over time on CRISPR biological tape. *Science*, 358(6369):1457–1461, December 2017.

[112] Sung Sun Yim, Ross M McBee, Alan M Song, Yiming Huang, Ravi U Sheth, and Harris H Wang. Robust direct digital-to-biological data storage in living cells. *Nat. Chem. Biol.*, 17(3):246–253, March 2021.

[113] Santi Bhattarai-Kline, Elana Lockshin, Max G Schubert, Jeff Nivala, George Church, and Seth L Shipman. Reconstructing transcriptional histories by CRISPR acquisition of retron-based genetic barcodes. August 2021.

[114] Max W Shen, Mandana Arbab, Jonathan Y Hsu, Daniel Worstell, Sannie J Culbertson, Olga Krabbe, Christopher A Cassa, David R Liu, David K Gifford, and Richard I Sherwood. Predictable and precise template-free CRISPR editing of pathogenic variants. *Nature*, 563(7733):646–651, November 2018.

[115] Hui Kwon Kim, Goosang Yu, Jinman Park, Seonwoo Min, Sungtae Lee, Sungroh Yoon, and Hyongbum Henry Kim. Predicting the efficiency of prime editing guide RNAs in human cells. *Nat. Biotechnol.*, 39(2):198–206, February 2021.

[116] Samuel D Perli, Cheryl H Cui, and Timothy K Lu. Continuous genetic recording with self-targeting CRISPR-Cas in human cells. *Science*, 353(6304), September 2016.

[117] Reza Kalhor, Prashant Mali, and George M Church. Rapidly evolving homing CRISPR barcodes. *Nat. Methods*, 14(2):195–200, February 2017.

[118] Rachel E Haurwitz, Martin Jinek, Blake Wiedenheft, Kaihong Zhou, and Jennifer A Doudna. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science*, 329(5997):1355–1358, September 2010.

[119] Samuel H Sternberg, Rachel E Haurwitz, and Jennifer A Doudna. Mechanism of substrate selection by a highly specific CRISPR endoribonuclease. *RNA*, 18(4):661–672, April 2012.

[120] Rachel E Haurwitz, Samuel H Sternberg, and Jennifer A Doudna. Csy4 relies on an unusual catalytic dyad to position and cleave crispr rna. *The EMBO journal*, 31(12):2824–2832, 2012.

[121] Lior Nissim, Samuel D Perli, Alexandra Fridkin, Pablo Perez-Pinera, and Timothy K Lu. Multiplexed and programmable regulation of gene networks with an integrated RNA and CRISPR/Cas toolkit in human cells. *Mol. Cell*, 54(4):698–710, May 2014.

[122] Andrew V Anzalone, Peyton B Randolph, Jessie R Davis, Alexander A Sousa, Luke W Koblan, Jonathan M Levy, Peter J Chen, Christopher Wilson, Gregory A Newby, Aditya Raguram, and David R Liu. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature*, 576(7785):149–157, December 2019.

[123] Jason C Klein, Vikram Agarwal, Fumitaka Inoue, Aidan Keith, Beth Martin, Martin Kircher, Nadav Ahituv, and Jay Shendure. A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nat. Methods*, 17(11):1083–1091, November 2020.

[124] James W Nelson, Peyton B Randolph, Simon P Shen, Kelcee A Everette, Peter J Chen, Andrew V Anzalone, Meirui An, Gregory A Newby, Jonathan C Chen, Alvin Hsu, and David R Liu. Engineered pegRNAs improve prime editing efficiency. *Nat. Biotechnol.*, 40(3):402–410, March 2022.

[125] Peter J Chen, Jeffrey A Hussmann, Jun Yan, Friederike Knipping, Purnima Ravisankar, Pin-Fang Chen, Cidi Chen, James W Nelson, Gregory A Newby, Mustafa Sahin, Mark J Osborn, Jonathan S Weissman, Britt Adamson, and David R Liu. Enhanced prime editing systems by manipulating cellular determinants of editing outcomes. *Cell*, 184(22):5635–5652.e29, October 2021.

[126] David J H F Knapp, Yale S Michaels, Max Jamilly, Quentin R V Ferry, Hector Barbosa, Thomas A Milne, and Tudor A Fulga. Decoupling tRNA promoter and processing activities enables specific Pol-II cas9 guide RNA expression. *Nat. Commun.*, 10(1):1490, April 2019.

[127] Manfred Gossen, Sabine Freundlieb, Gabriele Bender, Gerhard Müller, Wolfgang Hillen, and Hermann Bujard. Transcriptional activation by tetracyclines in mammalian cells. *Science*, 268(5218):1766–1769, 1995.

[128] Ulrike Zabel, Ralf Schreck, and PA Baeuerle. Dna binding of purified transcription factor nf-kappa b. affinity, specificity, zn2+ dependence, and differential half-site recognition. *Journal of Biological Chemistry*, 266(1):252–260, 1991.

[129] pGL4.49[luc2P/TCF-LEF/Hygro] vector protocol. `https://www.promega.com/resources/protocols/product-information-sheets/a/pgl4-49-vector-protocol/`. Accessed: 2021-10-24.

[130] Felix Muerdter, Łukasz M Boryń, Ashley R Woodfin, Christoph Neumayr, Martina Rath, Muhammad A Zabidi, Michaela Pagani, Vanja Haberle, Tomáš Kazmar, Rui R Catarino, Katharina Schernhuber, Cosmas D Arnold, and Alexander Stark. Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat. Methods*, 15(2):141–149, February 2018.

[131] J Choi, W Chen, A Minkina, F M Chardon, C C Suiter, S G Regalado, S Domcke, N Hamazaki, C Lee, B Martin, R M Daza, and J Shendure. A temporally resolved, multiplex molecular recorder based on sequential genome editing. November 2021.

[132] Mark E Fornace, Nicholas J Porubsky, and Niles A Pierce. A unified dynamic programming framework for the analysis of interacting nucleic acid strands: Enhanced models, scalability, and speed. *ACS Synth. Biol.*, 9(10):2665–2678, October 2020.

[133] Julian Banerji, Sandro Rusconi, and Walter Schaffner. Expression of a β-globin gene is enhanced by remote sv40 dna sequences. *Cell*, 27(2):299–308, 1981.

[134] Cornelia M Gorman, LESLIE F Moffat, and BRUCE H Howard. Recombinant genomes which express chloramphenicol acetyltransferase in mammalian cells. *Molecular and cellular biology*, 2(9):1044–1051, 1982.

[135] Jeffrey R De Wet, KV Wood, Marlene DeLuca, Da R Helinski, and S Subramani. Firefly luciferase gene: structure and expression in mammalian cells. *Molecular and cellular biology*, 7(2):725–737, 1987.

[136] Martin Chalfie, Yuan Tu, Ghia Euskirchen, William W Ward, and Douglas C Prasher. Green fluorescent protein as a marker for gene expression. *Science*, 263(5148):802–805, 1994.

[137] Eileen EM Furlong and Michael Levine. Developmental enhancers and chromosome topology. *Science*, 361(6409):1341–1345, 2018.

[138] Xiaojing J Gao, Lucy S Chong, Matthew S Kim, and Michael B Elowitz. Programmable protein circuits in living cells. *Science*, 361(6408):1252–1258, 2018.

[139] Ronghui Zhu, Jesus M del Rio-Salgado, Jordi Garcia-Ojalvo, and Michael B Elowitz. Synthetic multistability in mammalian cells. *Science*, 375(6578):eabg9765, 2021.

[140] Divya V Israni, Hui-Shan Li, Keith A Gagnon, Jeffry D Sander, Kole T Roybal, J Keith Joung, Wilson W Wong, and Ahmad S Khalil. Clinically-driven design of synthetic gene regulatory programs in human cells. *bioRxiv*, 2021.

[141] Andrew H Ng, Taylor H Nguyen, Mariana Gómez-Schiavon, Galen Dods, Robert A Langan, Scott E Boyken, Jennifer A Samson, Lucas M Waldburger, John E Dueber, David Baker, et al. Modular and tunable biological feedback control using a de novo protein switch. *Nature*, 572(7768):265–269, 2019.

[142] Zibo Chen and Michael B Elowitz. Programmable protein circuit design. *Cell*, 184(9):2284–2301, 2021.