Understanding Genomes Through Engineered Structural Variation

Jonas Koeppel

Pembroke College

University of Cambridge

Wellcome Sanger Institute

This thesis is submitted for the degree of Doctor of Philosophy

January, 2024

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. It is not substantially the same as any work that has already been submitted before for any degree or other qualification except as declared in the preface and specified in the text. It does not exceed the prescribed word limit for the Biological Sciences Degree Committee.

Acknowledgments

I would first like to thank my supervisor, Leopold Parts, for supporting, inspiring, and mentoring me throughout the wonderful journey of the last four years. Thank you for trusting and encouraging me to pursue crazy ideas and then supporting me every step of the way. Thank you for caring about me as a person and always making time to hear my concerns and share my excitement. You are the best mentor I could have hoped for!

I would also like to thank the incredible people at Sanger who infused my time here with warmth. I would especially like to thank a few people without whom this thesis would not have been possible. Thanks to Juliane Weller for making me see things from fresh perspectives, inspiring me, and sharing my highest highs and lowest lows. To Elin Madli Peets for helping me out with countless experiments, for making our lab a nurturing environment, and for entertaining me with an endless supply of memes. To Thomas Vanderstichele for inspiring conversations about science and life. To Fabio Liberante for building the wonderful SynGen community and for mentorship. To Mélanie Gouley and Valentin Rebernig for venturing to a far-away country, trusting me with their thesis projects, and teaching me so much in the process. To my PhD cohort, especially Emma Dann, and the Wellcome Sanger PhD program for making the experience a truly special one.

I shared this journey with many fantastic collaborators who generously offered their help and expertise along the way. Thanks to Tom Ellis for incepting the idea of mammalian scramble and his sound advice. To Raphael Feirerra for pushing through with me on the genome scramble project. To Klaudia Ciurkot for teaching me nanopore sequencing, to Jannat Ijaz and Peter Campbell for teaching me how to analyze structural variation.

I would like to thank Mikolaj Slabicki who taught me everything about how to do science. I would have had no idea where to even begin if it wasn't for you! Thanks to Pouya Baniasadi for friendship and many interesting conversations. Thanks to Arne Scheu for a life-long friendship. After so many years, you are still among the most inspiring people I've ever met. I cannot wait to see where our futures will take us, and I'm sure it's going to be exciting!

Finally, I would like to thank my parents for their boundless love and support. As the world spins and my journey takes me through countries close and far, I can always count on the warmth of home. I would not have dared any of this if it wasn't for you.

Abstract

Sequencing of the human genome has provided us with a detailed map of its content. While enormous progress has been made towards understanding the 1% of the human genome that is protein coding, we are still mostly in the dark about the function and relevance of the remaining 99%. Progress has been difficult because the non-coding genome is vast, the individual nucleotides hold less information, and we have lacked the tools to engineer and probe it to the necessary extent. This is beginning to change with the advent of 'search and replace' genome engineering technologies such as CRISPR prime editing. I leveraged the ability of prime editors to insert recognition sequences for recombinases at high throughput to engineer genomes at an unprecedented scale. In the process, I made discoveries about the biology of genome engineering, structural variation, and gene regulation.

I first outlined the determinants of short sequence insertion using prime editing by systematically measuring the frequency of insertion for 3,604 short sequences in four target sites of three human cell lines with varying DNA repair contexts. I characterized how insertion sequence length and two cellular DNA processing pathways affected the incorporation rate. I reaffirmed that DNA mismatch repair suppressed the insertion of shorter sequences and made the discovery that 3' flap nucleases TREX1 and TREX2 suppressed the insertion of longer sequences. I further delineated the effects of nucleotide composition and secondary structure of the insertion sequence on editing rates.

Next, I targeted a prime editor to the high copy number LINE-1 retrotransposon to insert hundreds of recombinase sites into a single human genome. These engineered cell lines provided a latent substrate for large-scale genome randomization. After induction with Cre recombinase, I mapped thousands of deletions, inversions, extrachromosomal circular DNA, translocations, and fold-back inversions and tracked their abundance over time. Sequencing surviving variants and comparing them to early ones revealed strong selection pressures against creating non-segregable derivative chromosomes or deleting essential genes. However, it also demonstrated that haploid human cell lines could survive while losing megabases of DNA. I isolated 21 cell clones and linked variants to gene expression changes for three clones with multiple Cre-induced rearrangements.

Finally, I used prime editing to insert loxPsym sites into the regulatory region of the *OTX2* developmental transcription factor. Cre recombinase induced stochastic deletions and inversions

across the recombinase sites, and created diverse and novel enhancer arrangements. By endogenously fusing *OTX2* with a fluorophore and sorting, I could associate alternative regulatory architectures with *OTX2* expression and track changes in CpG methylation and chromatin accessibility. I discovered that three enhancers in a 20 kb cluster drove 50% of *OTX2* expression and that moving the cluster closer to the transcription start site while simultaneously deleting intermediate regulatory elements resulted in strong *OTX2* expression.

The strategies presented here to more efficiently insert short DNA sequences with prime editing, shuffle DNA, and rearrange regulatory regions give a fundamentally new approach to randomizing mammalian genomes which will open new avenues to go beyond the 1% of coding sequence and study the 99% of underexplored regions. The data garnered from molecular phenotyping of novel genome architectures after randomization will allow predictive models to learn parameters beyond the limited diversity of our DNA.

Table of Contents

List of Figures

List of Tables

Abbreviations

ATAC-seq	Assay for transposase-accessible chromatin using sequencing
attB	Attachment site of bacteria
attP	Attachment site of phage
BAF	B-Allele frequency
bp	Base pair
BWA	Burrows-Wheeler algorithm
Cas	CRISPR-associated
cDNA	Complementary DNA
CDS	Coding sequence
ChIP-seq	Chromatin immunoprecipitation sequencing
chr	Chromosome
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
crRNA	CRISPR RNA
CTCF	CCCTC-binding Factor
dCas9	dead Cas9
DNase-seq	DNase I hypersensitive sites sequencing
epegRNA	Engineered prime editor guide RNA
FACS	Fluorescence-activated cell sorting
gDNA	Genomic DNA
H3K27ac	Histone H3 protein Lysine (K) 27 acetylation
H3K27me3	Histone H3 protein Lysine (K) 27 tri-methylation
H3K4me1	Histone H3 protein Lysine (K) 4 methylation
H3K4me3	Histone H3 protein Lysine (K) 4 tri-methylation
H3K9me3	Histone H3 protein Lysine (K) 9 tri-methylation
H3K36me3	Histone H3 protein Lysine (K) 36 tri-methylation
HA	Homology arm
HDR	Homology directed repair
hg38	Genome Reference Consortium Human Build 38
indel	Insertion deletion

kb	Kilo base pair
LINE	Long interspersed elements
loxP	Locus of X-over P1
loxPsym	Symmetrical loxP
m6A	N^6-methyladenosine
mCpG	5-methylcytosine in the context of a cytosine-guanine dinucleotide
Mb	Mega base pair
MMLV	Murine leukemia virus
MMR	Mismatch repair
MPRA	Multiplexed reporter assay
nCas9	nicking Cas9
NGS	Next generation sequencing
nt	Nucleotide
ORF	Open reading frame
PAM	Protospacer adjacent motif
PBS	Primer binding site
PCR	Polymerase chain reaction
PE	Prime editor
pegRNA	Prime editor guide RNA
RNP	Ribonucleoprotein
RTT	Reverse transcriptase template (of the pegRNA)
SCRaMbLE	Synthetic chromosome rearrangement and modification by loxP-mediated evolution
sgRNA	Single guide RNA
TAD	Topologically associating domain
TE	Transposable element
tracrRNA	Tracer RNA
tRNA	Transfer RNA
TSS	Transcription start site
UTR	Untranslated region

1

Introduction

The assumption that our genome is like an instruction manual where sections neatly lay out how to build a cell and ultimately a whole organism has proven naive. While these instructions exist, they only make up a small fraction of a large book. The instructive paragraphs are out of order and interrupted by pages and pages of seemingly useless and repetitive information; pages which do not tell a story on how to build a human but instead vaguely document millions of years of encounters with selfish genetic elements. Some of these pages are so old that most of the information has eroded. The book is a mess, yet cells found a way to read it and use its content to robustly execute the intricate steps from a zygote to a complex organism.

We will begin by drawing a map of what's in this book and summarize decades of work that illuminated how a cell can know which pages of its chaotic manual it needs to read at a given time. Subsequently, we will delve into the fantastic molecular tools that were developed to erase words or even change individual letters in this book. These tools have dramatically advanced our understanding of the instructive sentences but we barely grasp the relevance of the vast sea of pages in-between. The majority of the work presented here will focus on developing methods to shuffle or cut out pages and create many new versions of the instruction manual with changed content and page order to observe which of these can be understood by cells and how they might subtly differ from the original copy.

1.1 What is in our genome?

Genomes are carriers of hereditary information and embedded within are the instructions for making proteins, non-coding RNAs, as well as clues on how much and when to make them. Much of this information is captured in genes, which are transcribed into RNA that in turn can be translated into proteins that carry out most functions inside cells. However, the makeup of genomes and their relationship with genes was mysterious. For example, a puzzling observation was that the number of genes in organisms only poorly correlated with their genome size and neither correlated well with organismal complexity (C-value paradox, (C. A. Thomas Jr 1971)). The human genome is 200 times larger than that of the baker's yeast (*S. cerevisiae*) while containing only 3.4 times more protein-coding genes. In contrast, the European Mistletoe genome (*Viscum album)* is 30 times larger than the human genome and contains 1.6 times more genes (https://onlinelibrary.wiley.com/doi/full/10.1111/tpj.15558). This paradox was largely explained through an understanding that genomes can contain vast amounts of repetitive sequences, especially in multicellular organisms (Craig 1994). The initial sequencing of the human genome (Lander et al. 2001; Venter et al. 2001) provided the first approximate map of a complex and repetitive genome and it would take another 20 years to complete the first gapless assemblies of human genomes resulting in the most comprehensive picture of its content yet (Nurk et al. 2022).

The picture that emerged was this: The human genome is a vast desert of non-coding sequences sprinkled with small islands of DNA that contain the instructions to make proteins (Figure 1.1). These protein-coding sequences only make up ~1% (3% with untranslated regions) of our genomes and are interrupted by an average of 8-9 long introns that need to be actively removed upon transcription (Roy and Gilbert 2006). Another ~8% of our genome consists of non-coding sequences that regulate the expression of coding ones, sometimes from far away (ENCODE candidate cis-regulatory elements (ENCODE Project Consortium et al. 2020)). The majority (54%) of the genome desert is repetitive (Hoyt et al. 2022). Some repeat sequences such as centromeres and telomeres have known structural functions, but the functional relevance, if there is any, of most other repetitive elements remains enigmatic. To the naive eyes of someone looking for order, the organization of our genome appears highly non-streamlined. While the first gapless genome assemblies have mapped the contents of the human genome in unprecedented detail, we are still far away from understanding how the genome functions and to what degree that function is pinned on its organizational principles.

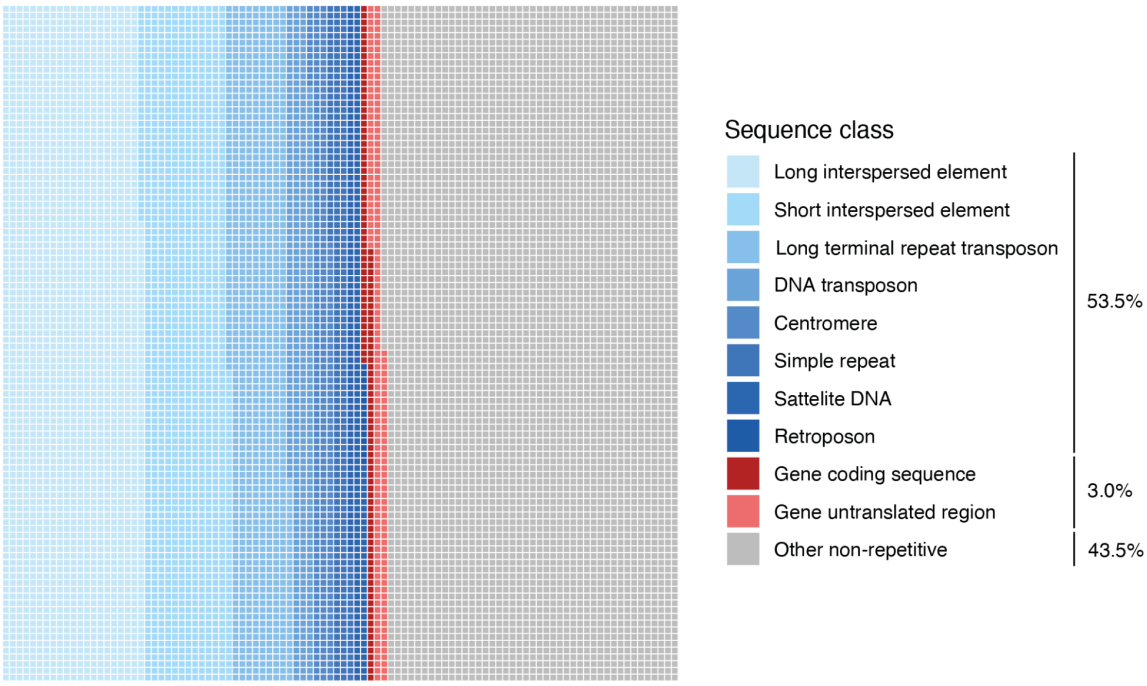

Figure 1.1. The composition of a human genome. Waffle plot of sequence classes in the human genome (CHM13v2). Repetitive elements are represented in shades of blue and sequences associated with protein-coding genes are in shades of red. Each box represents 0.01% of the genome. Excluding Y chromosomes. Data from (Hoyt et al. 2022; Nurk et al. 2022).

Early DNA renaturation experiments provided an estimate that about half of the genome is repetitive (Craig 1994; Waring and Britten 1966; Britten and Kohne 1968) which was confirmed by genome sequencing efforts (Lander et al. 2001; Venter et al. 2001). Human repetitive elements can be classified based on their arrangement and distribution throughout the genome which is either in tandem (satellites, centromeres, telomeres, simple repeats) or dispersed throughout (transposable elements and retrotransposed sequences).

1.1.1 Tandem repeats and centromeres

The presence of tandemly repeated DNA was first discovered in density gradient ultracentrifugation experiments where genomic DNA separated into two buoyant bands, a main fraction and a ‘satellite’ fraction (Kit 1961). The satellite fraction includes the sequences that underlie mammalian centromeres, which are formed from various types of satellite DNA repeated for millions of base pairs that together occupy 6.2% of the human genome (Altemose et al. 2022). During cell division mitotic spindles attach to centromeres (mediated by large protein complexes called kinetochores) to coordinate an intricate process leading to the faithful separation of sister chromatids into the new daughter cells (McKinley and Cheeseman 2016). Because centromeric sequences are highly repetitive, short-read technologies could not resolve them and it took until 2022 to generate the first complete map of human centromeres (Altemose et al. 2022) soon

followed by maps of more individuals and great apes (Logsdon et al. 2023). These efforts revealed that centromeres are fast-changing sequences where new classes of repeats emerge in the center and drive out old repeats toward the periphery. Centromeres display high levels of CpG DNA methylation but contain a dip in methylation at the position of kinetochore assembly. Intriguingly, the large and repetitive structure of centromeric sequence is not necessary for centromere function and cases exist where the location of kinetochore assembly is moved to ectopic, non-repetitive DNA forming neo-centromeres that are stably maintained through cell division (DeBose-Scarlett and Sullivan 2021).

1.1.2 Interspersed repeats and transposable elements

Interspersed repeats are mostly made up of transposable elements (TEs) which have spread throughout the genome over millions of years. Two important categories that transpose through an RNA intermediate are long interspersed elements (LINEs) and short interspersed elements (SINEs). LINEs make up ~21% of our genome (Hoyt et al. 2022) and encode two open reading frames over 6,000 bp which include the instruction for the reverse transcriptase. While the majority of LINE (>99.9%) have been rendered inactive through mutations, a small fraction of LINE-1 elements remain active and their mobilization dynamically shapes our genomes (Nam et al. 2023; Beck et al. 2011). LINE-1 encoded proteins can also occasionally integrate cellular mRNAs to new genomic locations and generate pseudogenes in the process. SINEs are much smaller (100-300bp) and do not encode proteins, but instead depend on LINE-encoded proteins for retrotransposition. They are among the most successful transposable elements in human genomes and account for about 14% of our DNA, sometimes occurring in extremely high copy numbers (well over 1 million copies for Alu sequences (Lander et al. 2001)).

TEs are fundamentally selfish and their mobilization has been associated with at least 65 genetic diseases (Belancio, Hedges, and Deininger 2008) as well as cancer (Tubio et al. 2014). In organisms with large effective population sizes and short doubling times, efficient selection is thought to keep the copies of TEs in check (Lynch and Conery 2003). In complex multicellular organisms with small population sizes, selection against the spread of TEs is ineffective and elements might accumulate over time until they become a burden to the host (Vinogradov 2003). However, the relationship between TEs and their hosts is complex, and mobile elements can act as substrates for evolutionary innovations. For example, a retroviral integrase is required for V(D)J recombination in immune cells (Roth and Craig 1998) underlying the ability to generate diverse B and T-cell receptors while the insertion of an Alu element is presumably responsible for tail-loss in humans (Xia et al. 2021). Beyond specific examples, it is unclear if the majority of

TEs in our genome have any function and an open question is how many TEs one could in principle remove from a human genome without disturbing its function.

1.1.3 Genes

The other half of our genomes consists of largely non-repeating sequences that include genes and the regulatory sequences necessary to instruct when and where genes are expressed. Genes are the basic unit of inheritance and physically consist of nucleic acids that typically encode the instructions to make RNA and proteins. However, the exact definition of a gene is blurry. Many genes do not encode proteins and may not have functional relevance; sometimes genes can overlap (common in viruses); pseudogenes physically resemble genes but cannot be made into protein, and are often the result of spurious reverse transcription from selfish genetic elements. According to the latest human gene annotation data set (Ensembl 110.38) (Howe et al. 2021; Birney et al. 2004), our genome comprises 19,831 protein-coding genes and 25,959 non-coding genes. A given cell type typically expresses 11-13,500 protein-coding genes, of which around 8,000 are ubiquitously expressed (Ramsköld et al. 2009). But what determines which genes are expressed and which ones are not in a given cell type?

1.2 Transcriptional regulation

The genome in all human cells is nearly identical, yet we are composed of tissues containing more than 400 specialized cell types with vastly different morphologies and functions (Tabula Sapiens Consortium et al. 2022). In addition, the exact protein and RNA makeup of each cell varies over time as it reacts to changing environments, most notably during development. This differing use of the same genome is orchestrated by complex interactions between proteins and regulatory DNA sequences which result in the reversible modification of chromatin and DNA bases. These modifications are sensed and maintained by additional proteins and determine where, when, and at what quantity gene products should be made. Additional regulatory layers control mRNA abundance, translation, and protein turnover. Here we will focus on the layer of transcriptional control which is shaped by three important categories of regulatory DNA sequences: (1) promoters (2) distal transcription factor binding sites (enhancer/silencers) and (3) binding sites for factors that shape chromatin topology (Figure 1.2).

1.2.1 Regulatory elements

While other promoters and polymerases exist in the human genome (reviewed in Loeb and Monnat 2008), we will focus on promoters of protein-coding genes that are transcribed by RNA polymerase II. To initiate gene expression, RNA polymerase needs to be recruited to the

transcriptional start site of a gene. Promoters are regions where this binding is facilitated. A core promoter contains binding sites for general transcription factors – the TATA box, the initiator element, and the TFIIB-recognition element (BRE) are well-known motifs (Goldberg 1979; Smale and Baltimore 1989). General transcription factors assembled on a core promoter, recruit RNA polymerase II and together form the pre-initiation complex (Roeder 1996; Orphanides, Lagrange, and Reinberg 1996). Promoters differ in which sequence motifs they contain and ~25% do not contain any of the classical binding motifs for general transcription factors (Gershenzon and Ioshikhes 2005). Often additional transcription factors bind in the proximity of the core promoter (proximal promoter elements) and recruit co-activators/co-repressors (e.g. CBP-P300, mediator) that together modulate loading of RNA polymerase II.

Transcription from a promoter alone is often weak without input from distal transcription factor binding events at enhancer sequences (Banerji, Rusconi, and Schaffner 1981). Enhancers are generally a few hundred nucleotides in size and contain binding motifs of transcriptional regulators and epigenetic modifiers (Reményi, Schöler, and Wilmanns 2004). In a working model, the transcriptional regulators assembled on the enhancers interact with the promoter through DNA looping and thereby influence the presence and composition of transcription initiation complexes loaded onto transcription start sites. Chromosome looping is often necessary as enhancers can be located far away from genes whose expression they enhance (Schoenfelder and Fraser 2019). Promoters and enhancers are loosely classified by their distance from the transcription start site but share many similarities. For example, promoters can have enhancer activities and enhancers can drive local transcription initiation (Andersson and Sandelin 2020).

Another type of distal regulatory element is silencers which are bound by repressive transcription factors and decrease the expression of interacting genes (Maston, Evans, and Green 2006). The integration of signals from enhancers and silencers might help fine-tune gene expression and some silencers can turn into active, tissue-specific enhancers during development (Ngan et al. 2020; Huang and Ovcharenko 2022). Silencers are less well studied due to increased experimental difficulties in finding sequences that repress rather than activate expression (it's much easier to spot a tall crop in a field full of wheat compared to a short crop). Three recent studies have attempted to systematically find silencers by screening for repression of reporter constructs or by measuring interaction with the polycomb repressive complex (Doni Jayavelu et al. 2020; Pang and Snyder 2020; Ngan et al. 2020). They find that silencers are highly cell type specific and there is generally little overlap between the sequences identified in these studies.

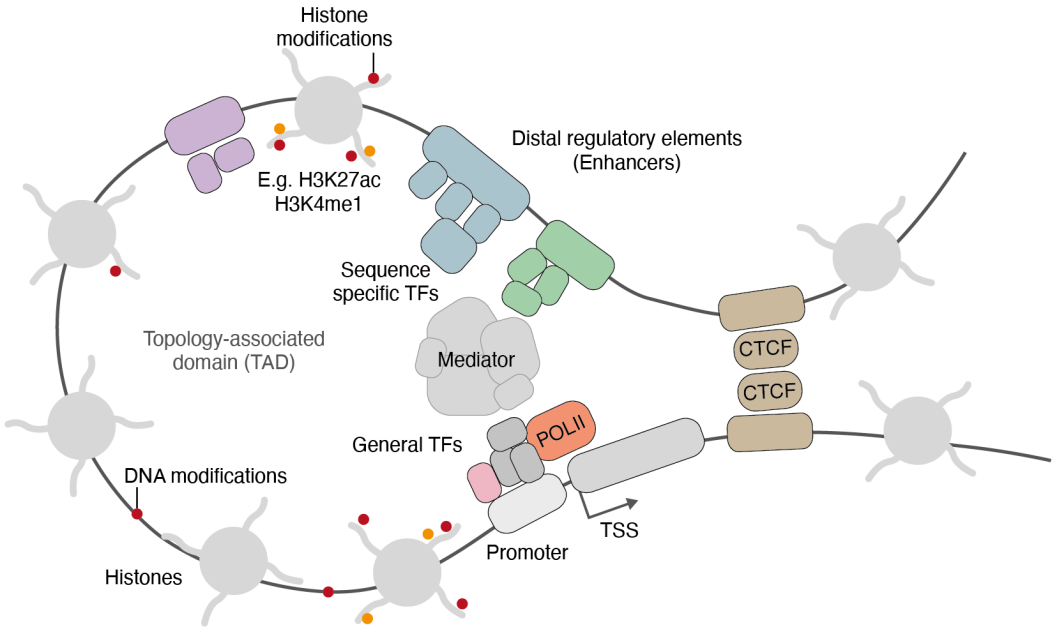

Figure 1.2. The complex orchestra of gene regulation. A schematic of a topology-associated domain is shown where distal regulatory elements interact with a promoter through chromatin looping and transcription factor interactions. This results in the recruitment and positioning of RNA polymerase II at the transcription start site. Regulatory elements are marked by DNA and histone modifications. The chromatin looping is maintained by cohesin (not shown) and CTCF binding. POLII: RNA Polymerase II. Rounded and outlined squares: proteins. Gray boxes: DNA regions. TF: Transcription factor. TSS: Transcription start site.

Distal regulatory elements can be located far away from their target genes (sometimes more than 1 Mb, (Bahr et al. 2018; Lettice et al. 2003; Long et al. 2020)) and an important question is how they find and regulate their cognate targets. A general understanding is that interacting enhancers and promoters occupy the same topologically associated domains (TADs) (Y. Shen et al. 2012), islands of a few hundred kilobases where sequences within interact more with each other than with neighboring sequences outside (Bonev and Cavalli 2016). TAD boundaries are usually enriched for highly transcribed genes or the presence of binding site clusters for CTCF (Bonev and Cavalli 2016; Ong and Corces 2014). Indeed, depletion of CTCF increases the frequency of interactions between sequences in adjacent TADs (Zuin et al. 2014). A prominent example of the importance of boundary elements is the *IGF2*/*H19* imprinted locus (Z. Zhao et al. 2006). An enhancer cluster either activates *IGF2* or H19 based on the methylation status of a tandem array of CTCF binding sites. In the paternal allele, the cluster is unmethylated and bound by CTCF, creating a boundary between the enhancers and *IGF2*, consequently activating *H19* but not *IGF2*. In the maternal allele, the cluster is methylated and not occupied by CTCF; no boundary is established and the enhancers drive *IGF2* expression but not *H19* expression. However, the importance of TAD boundaries on the maintenance of gene expression more generally is

challenged by data showing that disruption of TAD boundaries or depletion of CTCF does not lead to systematic gene expression changes (Ghavi-Helm et al. 2019; Despang et al. 2019; Nora et al. 2017).

Active regulatory elements can be mapped across developmental stages and cell types through the detection of characteristic chromatin modification and accessibility patterns. DNA is wrapped around histones which have a long and unstructured N-terminal tail that can be decorated with various modifications. Histone modifications influence higher-order chromatin structure and affect interactions with chromatin-binding proteins (Kouzarides 2007). Antibodies specific to histone modifications can be used to pull down DNA wrapped around modified histones which can subsequently be mapped using microarrays (ChIP-chip) (Bernstein et al. 2004; C. L. Liu et al. 2005) or sequencing (ChIP-seq) (Barski et al. 2007). Regulatory elements tend to be enriched for histone 3 lysine 27 acetylation (H3K27ac) and histone 3 lysine 4 methylation (H3K4me1). In addition, histones are evicted from the core of regulatory elements and promoters to provide space for the binding of transcription factors. These histone-devoid regions are more sensitive to digestion by DNase I and the resulting fragments can be mapped by sequencing to reveal accessibility footprints (DNase-seq) (Boyle et al. 2008). An alternative method to study accessibility is through an 'assay of transposase accessible chromatin' (ATAC-seq). This strategy leverages the bias of Tn5 transposase toward accessible chromatin to cleverly integrate sequencing adaptors (Buenrostro et al. 2013).

Concerted efforts have been made to functionally annotate regulatory DNA, for example, the Encyclopedia of DNA Elements (ENCODE) integrates DNase-seq and ChiP-seq data sets across hundreds of cell types to define cis-regulatory elements (ENCODE Project Consortium 2004; Sloan et al. 2016; ENCODE Project Consortium et al. 2020). Other approaches look at evolutionary conservation (Kuderna et al. 2023; J. W. Thomas et al. 2003; R. Chen et al. 2001) or constraint in disease to identify important DNA (Rentzsch et al. 2019; Kircher et al. 2014; S. Chen et al. 2023). These efforts resulted in a rich map of almost one million DNA annotations which demonstrated that putative enhancers far outnumber genes and are highly tissue and cell-type-specific.

1.2.2 Functional analysis of gene regulation

While biochemical maps of DNA and histone modifications and evolutionary constraints helped us build catalogs for possible regulatory elements, we are still far from understanding their function. The next section will delve into experimental strategies to characterize regulatory elements and the understanding we are deriving from these experiments (Figure 1.3).

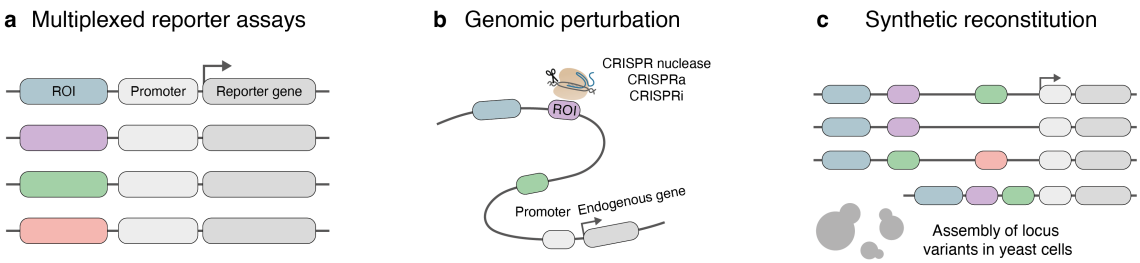

Figure 1.3: Experimental strategies to characterize regulatory elements. a. Multiplexed reporter assays: Thousands to millions of regions of interest (ROI) are cloned in front of a minimal promoter to test their ability to drive the expression of a reporter gene. **b.** Genomic perturbation: CRISPR nuclease, CRISPR activators, or CRISPR inhibitors are targeted to regulatory elements, and their effect on the expression of the cognate target gene is assessed. **c.** Synthetic reconstitution: Variants of the locus of interest are first assembled in vitro and in yeast and subsequently introduced into a mammalian cell to map their expression levels.

A high throughput way to experimentally test the regulatory potential for millions of DNA sequences is with multiplexed reporter assays (MPRAs). Sequences of interest are cloned next to reporter constructs and their ability to drive reporter expression is measured by deep sequencing (Melnikov et al. 2012; Patwardhan et al. 2012; Arnold et al. 2013; Gordon et al. 2020; de Boer et al. 2020). However, MPRAs generally do not account for the genomic context and chromatinization of enhancers, how they interact with nearby regulatory elements, and their compatibility with promoters. Two recent studies expanded the MPRA to test the compatibility of promoters and enhancers at a large scale (Martinez-Ara et al. 2022; Bergman et al. 2022). They discovered that the majority of enhancers and promoters are compatible and drive the interaction of target genes multiplicatively. The exceptions are promoters of housekeeping genes that are less susceptible to enhancers, and promoters of variably expressed genes that are more susceptible to regulation by enhancers (Bergman et al. 2022). MPRAs are powerful in revealing the regulatory potential of individual sequences but fundamentally strip those sequences of the context in which they usually operate.

An alternative method to understand enhancers in their endogenous context is through genomic perturbation. For example, enhancers can be mutagenized with clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated (Cas) systems or activated/repressed through Cas9 fused to transcriptional modulators (CRISPR activation or CRISPR inhibition, discussed later) (Lopes, Korkmaz, and Agami 2016). This approach is attractive to map the functional relevance of individual enhancers but the types of possible perturbations are limited and local. Lin et al. attempted to expand on this by screening the effects of all single and pairwise

interactions of seven *MYC* enhancers (Lin et al. 2022). They discover that nearby enhancers tend to interact additively while far-away enhancers interact synergistically. Genome perturbations delivered clean insights into the necessity of individual regulatory elements in their relevant endogenous context. While we learned which parts of existing regulatory circuits are crucial for normal function, perturbations fundamentally cannot show us how to build a new gene regulatory circuit from scratch.

To build new circuits, we need to understand for example how the physical spacing of regulatory elements across endogenous loci affects their behavior, a question that none of the methods above could address. Two recent studies use clever approaches to address this question. Zuin et al. integrated a reporter transgene that was split by a transposon containing the cognate enhancers into a genomic region with low regulatory complexity (Zuin et al. 2022). Transposase expression mobilizes the enhancers locally and the corresponding reporter expressions can be measured for enhancers of varying genomic distance to and contact frequencies with the promoter. The enhancer strength decreased with increasing genomic distance from the transcription start site but less rapidly than DNA contact probability and the presence of a single CTCF site decreased transcription levels across it by 60%. Thomas et al. instead inserted orthogonal recombinase sites at varying distances into another region of low regulatory complexity to address the same question (H. Thomas et al. 2023). They subsequently integrated enhancers into these landing pads and also observed a loss of enhancing potential with increasing distance. The degree to which the potential decreased was correlated to the intrinsic strength of the enhancers, whereby the potency of weaker enhancers decayed more quickly with distance. Intriguingly, for simultaneous integration of multiple enhancers at different distances, weak enhancers that could not activate transcription from intermediate distances greatly boosted the strength of distal enhancers by acting as a ‘stepping stone’ between the promoter and distal enhancer. Together, a picture emerges where promoter distance, insulation by CTCF, enhancer strengths, as well as enhancer interactions together orchestrate gene activation. These approaches also highlight the potential of genome engineering with transposases and recombinases to dissect enhancer logic (discussed later).

All of the examples above focus on some aspects of enhancer biology and study them in isolation. The proof of a complete understanding is the ability to design complex regulatory regions to a specification, followed by constituting and testing them in endogenous contexts. Early synthetic reconstitutions create several variations of a regulatory region that are synthesized and assembled (e.g. in yeast) followed by the integration into the human genome (Brosh et al. 2021). Reconstitution allows the redesign of virtually any aspect of a locus and has been used to disentangle the contributions of retinoic acid response elements and distal enhancers in the regulation of *HOX* genes (Pinglay et al. 2022), to understand the context-dependency of the *H19*

and *SOX2* enhancers (Ordoñez et al. 2023), and to discover a type of regulatory sequence that facilitates the action of other enhancers in the alpha-globin and *SOX2* loci (Brosh et al. 2023; Blayney et al. 2023), reminiscent to the interaction observed by enhancer integration into multiple landing pads (H. Thomas et al. 2023). To more comprehensively learn and predict regulatory grammar, we will likely have to probe many more variants in many more loci. While synthetic reconstitution is powerful, it is also labor intensive, requiring the utilization of three different model organisms (yeast, bacteria, and the target mammalian cell line) as well as molecular biology and DNA synthesis and is therefore not easily scalable. An ideal method could generate a large number of complex variants directly in mammalian cells (discussed in Chapter 3).

1.3 Highly rearranged genomes

Another way to understand genome organization and gene regulation is through the study of highly rearranged genomes that provide many natural examples of the consequences of moving DNA sequences out of their normal context. Ideally, these rearranged genomes would have a counterpart that is still in wild-type configuration so that direct comparisons can be drawn. Three very different organisms and model systems can currently provide such insight: (1) balancer chromosomes in the fruit fly, (2) chromothripsis in cancers, and (3) the scramble system built into synthetic yeast genomes (Figure 1.4).

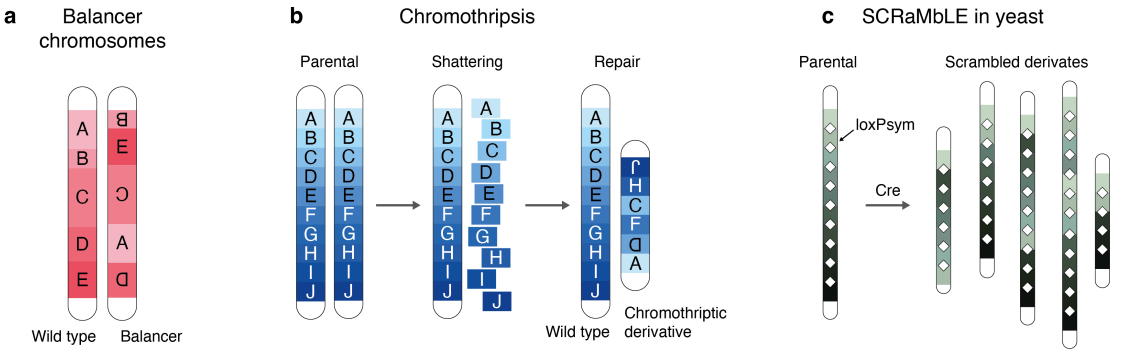

Figure 1.4: Examples of highly rearranged genomes. a. Balancer chromosomes in fruit flies have several large and nested inversions and are only viable with a normal chromosome. **b.** Chromothripsis is a mutational event where usually one allele of a chromosome arm becomes shattered into hundreds of fragments which are subsequently lost or stitched together in random order **c.** Synthesized yeast chromosomes have hundreds of loxPsym sites integrated (white diamonds) and Cre induction can result in vastly changed chromosome architectures.

1.3.1 Balancer chromosomes

Balancer chromosomes in Drosophila are heavily rearranged chromosomes that suppress the recombination with wild-type chromosomes in meiosis (Oster, 1956). Comparing gene expression from the wild-type and balancer alleles is an opportunity to understand the consequences of structural changes (Ghavi-Helm et al. 2019). The structural variants on the balancer chromosomes caused extensive changes to chromatin topology, but only a handful of genes around variant breakpoints changed expression as a consequence (Ghavi-Helm et al. 2019). It remains an open question why some genes are sensitive to chromatin topology while others are not. While balancer chromosomes have more variants than usual Drosophila chromosomes, the majority of results from the study came from only eight large and nested inversions. Further, balancer chromosomes are highly selected to behave like wild-type chromosomes and might thus be depleted from changes with drastic effects on gene expression. In contrast, chromothriptic derivative chromosomes can harbor thousands of variants.

1.3.2 Chromothripsis

Chromothripsis is a catastrophic mutational event where the arms of one or more chromosomes are shattered into hundreds or thousands of pieces. Some fragments are lost in the process and the remaining pieces are re-joined in a seemingly random order and orientation (Stephens et al. 2011). As a result, cells that survived chromothripsis will have heavily rearranged chromosomes that show a characteristic copy number oscillation around two or three copy number states (Korbel and Campbell 2013). Chromothripsis can amplify oncogenes or inactivate tumor suppressor genes and is pervasive in cancers, affecting around 30% of whole-genome sequenced cancer samples (Cortés-Ciriano et al. 2020). One mechanism through which chromothripsis arises is from lagging chromosomes in mitosis which can become encapsulated in aberrant nuclear structures called micronuclei. Within micronuclei, DNA replication is delayed and DNA repair impaired (Krupina, Goginashvili, and Cleveland 2021). Rupture of micronuclei can additionally expose lagging chromosomes to cytosolic factors. Together these factors can lead to the shattering of the chromosome(s) within nuclei (Krupina, Goginashvili, and Cleveland 2023; Hatch et al. 2013; C.-Z. Zhang et al. 2015). Attempts to repair the shattered pieces can result in chromothriptic chromosomes that are eventually re-incorporated into the main nucleus after mitosis (Ly et al. 2019; Crasta et al. 2012).

The high density of rearrangements in chromothripsis provides unique examples of genes that are repositioned to novel genomic neighborhoods and thus makes it an ideal model system to study the importance of gene order, orientation, as well as transcriptional and epigenetic neighborhoods on gene expression. Since chromothripsis usually only affects one of two alleles, the wild-type

copy can serve as a reference for 'normal' gene expression. Ijaz et al. reconstructed a chromothriptic chromosome 6 from an esophageal cancer organoid and resolved 916 structural variations (Ijaz 2021). They discovered that more genes were differentially expressed between the chromothriptic and wild-type chromosomes compared to two wild-type chromosomes and that differentially expressed genes reside closer to structural variants than non-differentially expressed genes. Structural variants changed TAD structures and moved genes into regions with different chromatin marks. For example, chromothripsis moved the *AKA12* gene from active chromatin (H3K27ac and H3K4me3 marks) into more repressive chromatin (H3K27me3 marks) which resulted in a 570-fold downregulation. These results demonstrate that the integrated study of structural variants, genome topology, histone and DNA modifications, and gene expression can yield insights into the grammar of genome usage.

Both balancer and chromothriptic chromosomes are post-selection systems. It is unclear how often complex rearrangements occur that produce derivative chromosomes incapable of survival and are therefore never seen. To understand the effects of structural variants on gene expression and cell viability better, an ideal system would allow us to take a snapshot of all variants that were generated after a period of genome instability and then additional snapshots as cells with lethal variants are depleted. Synthetic yeast chromosomes with hundreds of recombinase sites enable that.

1.3.3 Scrambled yeast genomes

Our capacity to synthesize and assemble DNA from the ground up has advanced remarkably over the last three decades. The prospect of crafting synthetic genomes for complete cells has come within reach. Following the synthesis of the compact genome of the *Mycoplasma mycoides* bacterium (Gibson et al. 2010), an international consortium aimed to synthesize the substantially more complex genome of the eukaryotic single-celled yeast, *Saccharomyces cerevisiae* (Richardson et al. 2017; J. S. Dymond et al. 2011). The design diverged from the wild-type genome in five ways: All TAG stop codons were reassigned to TAA, repeat elements, tRNAs, and many introns were removed or relocated, short recoded sequences (PCR tags) were introduced into open reading frames (ORFs) and symmetrical loxP sites (loxPsym) were inserted into the 3' UTRs of all non-essential ORFs. Each of the hundreds of loxPsym sites in synthetic yeast chromosomes could act as an anchor for recombination by Cre recombinase. Collectively, these sites form the substrate of an inducible genome evolution system called SCRaMbLE (synthetic chromosome rearrangement and modification by loxP-mediated evolution) (J. Dymond and Boeke 2012). Recombinase expression will trigger thousands of deletions, inversions,

duplications, and translocations across the yeast population creating a myriad of diverse genome architectures and phenotypes.

The diversity created in a SCRaMbLE experiment can be harnessed to evolve cells towards desired phenotypes or to understand genome biology. SCRaMbLE has demonstrated its effectiveness across diverse contexts, successfully enhancing the biosynthesis of carotene, lycopene, violacein (Blount et al. 2018; W. Liu et al. 2018; Gowers et al. 2020; Juan Wang et al. 2018; Jia et al. 2018; Wu et al. 2018), bolstering the resistance to acetic acid, alkali, caffeine, ethanol, heat, and salt (Kang et al. 2022; Ma et al. 2019; W. Liu et al. 2018; Luo et al. 2018), as well as facilitating growth on xylose as an alternative carbon source (Blount et al. 2018). Key to the success of SCRaMbLE in finding genomic architectures with beneficial phenotypes is the large combinatorial rearrangement potential. For example, a study in a yeast strain with six synthetic chromosomes detected over 260,000 unique rearrangements following Cre induction (Zhou et al. 2022). Studying heavily rearranged genomes also uncovered novel biology. Through the careful examination of 612 SCRaMbLE-induced gene repositions, Brooks et al. discovered that transcriptional neighborhoods predictably regulate transcript isoform lengths and expression levels (Brooks et al. 2022).

The yeast genome is relatively simple, 250 times smaller than the human genome and 73% are protein-coding (Mewes et al. 1997). In contrast, the human genome is vast, mostly repetitive, non-coding, and poorly understood. There is likely an expanse of novel genome biology to discover if we implemented a SCRaMbLE system in human cells. SCRaMbLE in yeast was achieved through complete genome synthesis but the size of the human genome still puts it out of reach for *de novo* synthesis. Instead, to scramble human cells, we need to engineer existing genomes and for this, we need versatile and effective tools.

1.4 Tools for the programmable manipulation of genomes

While single-gene knockouts and the introduction of point mutations have been effective in making sense of the coding genome (e.g. (Findlay et al. 2018; Meyers et al. 2017; Hart et al. 2015; Costanzo et al. 2010; Winzeler et al. 1999; Findlay et al. 2014)), the sparsity of the non-coding genome demands tools that can delete/invert and transpose sequences at a much larger scale. Making these types of programmable changes has been historically challenging. However, in recent years a convergence of technologies has enabled this type of genome manipulation at an unprecedented scale. These technologies are (1) CRISPR prime editing, (2) highly multiplexed editing, (3) site-specific recombinases. Each is discussed in more detail below.

1.4.1 CRISPR/Cas systems

While several approaches to engineering genomes (summarized in (Klug 2010; Bogdanove and Voytas 2011)) existed before CRISPR/Cas systems, these approaches lacked rapid programmability and each additional target site required the design of a novel protein. The aspiration for the genome engineering field was to find an enzyme that could be programmed to make precise manipulations in the genome at any desired location without the need to create a protein from scratch for each new target. The discovery and characterization of a remarkable adaptive microbial immune system was a major leap toward this goal (Bhaya, Davison, and Barrangou 2011). Upon phage infection, CRISPR/Cas systems can integrate small nucleotide sequences (protospacers) from the invading phage genome into the CRISPR locus of the host genome. Once expressed, the protospacers will serve as guides to recruit Cas nucleases to cut the DNA of invading phages and prevent infection.

By now, many diverse CRISPR systems have been characterized across the tree of life (including eukaryotes (Saito et al. 2023; K. Jiang et al. 2023)), with functions beyond adaptive immunity (Meers et al. 2023). We will focus on the most widely characterized CRISPR/Cas9 system (Jinek et al. 2012). Here, the Cas9 protein binds to a complex of a transactivating guide RNA (tracrRNA) and a crRNA which guides the ribonucleoprotein complex to target sites with complementarity to the protospacer next to a short protospacer-adjacent motif (PAM – NGG for *Streptococcus pyogenes* Cas9). At the target site, the Cas9 enzyme cuts both DNA strands, destroying the invader. Excitingly, the tracrRNA:crRNA complex could be replaced with an artificial RNA sequence which made it possible to guide Cas9 to any desired DNA target (Jinek et al. 2012). In 2020, Doudna and Charpentier were awarded with the Nobel Prize for their characterization of CRISPR systems as programmable nucleases.

The programmability of the CRISPR/Cas9 system made it an ideal candidate for the development of a mammalian genome editor. This was achieved through a combination of innovations: fusion of the crRNA and tracrRNA into a single guide RNA (sgRNA) (Jinek et al. 2012), the endowment with a nuclear localization signal, and codon optimization for mammalian expression (Cong et al. 2013; Mali et al. 2013; Jinek et al. 2013). Cas9 could now in principle make a double-strand break at any position in the genome next to a PAM site (Figure 1.5a).

Mammalian cells can repair double-strand breaks in one of three ways. Non-homologous end-joining and microhomology-mediated end-joining are error-prone and typically create short insertion and deletion mutations around the cut site which could be exploited to knock out target genes or disrupt the binding of regulatory factors. This is beneficial to study the function of genes

in the laboratory or to disrupt pathogenic proteins in disease. Alternatively, homology-directed repair (HDR) can precisely fix the lesion by using complementary information from the sister chromatids or homologous chromosomes. HDR can be leveraged to install precise mutations including long insertions by supplying exogenous donor DNA templates encoding the edits and homologies to the cut site (Cong et al. 2013). However, HDR is inefficient in non-dividing cells (Rothkamm et al. 2003) and is usually accompanied by many alternative alleles with mutations. In addition, more recent work has demonstrated that double-strand breaks can result in catastrophic events including large, on-target deletions, chromothripsis (discussed later), and aneuploidies (Nahmad et al. 2022; Adikusuma et al. 2018; Papathanasiou et al. 2021; Weisheit et al. 2020). Nevertheless, the versatility and ease of use propelled CRISPR into countless molecular biology research laboratories and the first CRISPR-based therapeutics are showing great promise in clinical trials.

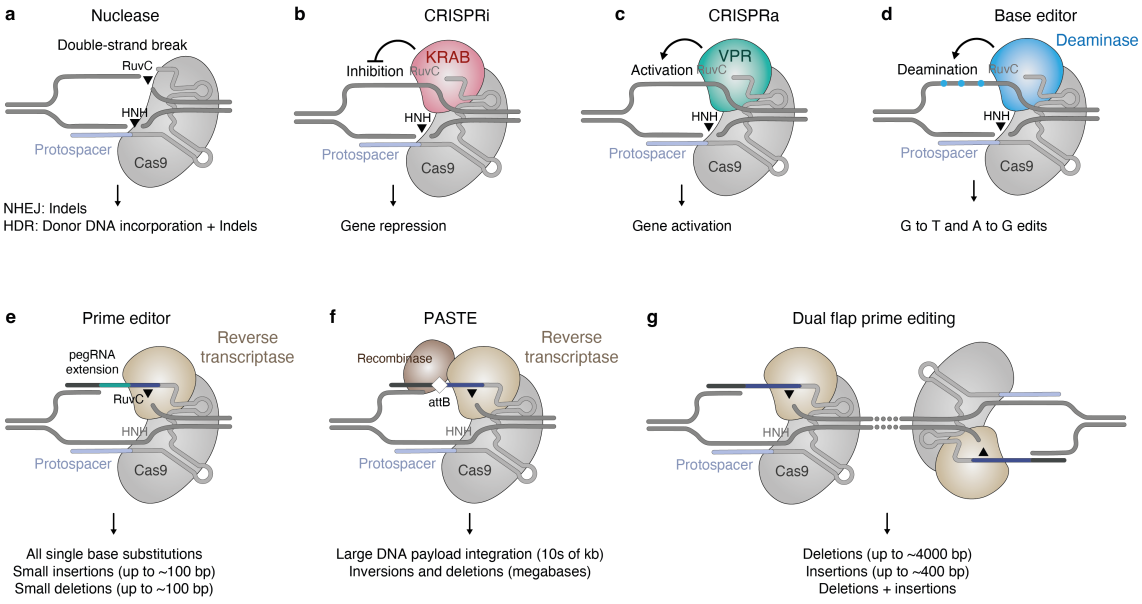

Figure 1.5. The CRISPR/Cas toolbox. a. The Cas9 nuclease recognizes DNA sequences complementary to the protospacer sequence and next to a protospacer adjacent motif to induce two DNA breaks. **b-c.** CRISPR inhibitors and activators fuse dead Cas9 with a co-repressor or co-activator recruiting domain (here KRAB or VPR) to repress or activate nearby genes. **d.** Base editors fuse a deaminase domain to nicking Cas9 to mediate base conversions. e. RuvC and HNH: endonuclease domains of Cas9. **e.** Prime editors fuse a reverse transcriptase to nicking Cas9 and also employ a 3' extended guide RNA (pegRNA) to install precise substitutions, short insertions, or short deletions. **f.** PASTE systems use the prime editor platform to introduce a recombinase site and co-deliver recombinase for large cargo insertion. **g**. In dual flap prime editing strategies both DNA strands are targeted. Repair of the lesion can result in large deletions and insertions.

The downside of Cas9 for engineering is the lack of control and the dependency of the mutation outcome on stochastic repair processes (Hussmann et al. 2021; Pallaseni et al. 2023; Cong et al. 2013). Advancements have been made to nudge the outcomes of repair (Riesenberg et al. 2023; Robert et al. 2015) and predict the types of mutations that are generated at a given target site (M. W. Shen et al. 2018; Allen et al. 2018), but the precise installation of desired mutations or the insertion of novel sequences generally comes with undesired by-products and depends on co-delivery of an external DNA donor (Mali et al. 2013). These limitations make Cas9 a great workhorse for knocking out genes but a blunt tool for precise genome manipulations.

1.4.2 Precision genome engineering

A key insight to developing the next generation of genome editors was the separation of targeting and effector functions of Cas proteins. Cas9 contains two nuclease domains (RuvC and HNH) that can be inactivated separately or together, resulting in nicking and catalytically dead versions of Cas9 (nCas9 and dCas9). These modified proteins can still be programmably targeted but no longer make double-strand breaks and represent perfect canvases for the addition of novel effector domains. For example, transcriptional repression and activation domains can be added to create epigenetic editors that silence or activate their target genes (Figure 1.5b-c) (Qi et al. 2013; Gilbert et al. 2013). An ingenious next step was the fusion of cytidine or adenosine deaminases to nicking or dead Cas9 which made it possible to directly install point mutations (C to T or A to G – and their reverse complement) at desired genomic locations (Figure 1.5d) (Komor et al. 2016; Gaudelli et al. 2017). Together this toolkit complemented and expanded the use cases of standard Cas9 but it was still not possible to precisely install transversion mutations, deletions, insertions, or larger structural changes.

Prime editors are a newly developed tool that is poised to finally fulfill our aspirations to programmably install any type of precise edits anywhere in the genome (Anzalone et al. 2019). Prime editors consist of the Cas9 nickase fused to a reverse transcriptase and a prime-editing guide RNA (pegRNA). The pegRNA serves two purposes; it specifies the target and the desired edit. At the target site, Cas9 nicking releases one strand of the target DNA which hybridizes to the primer binding site of the pegRNA (Figure 1.5e). The reverse transcriptase then copies the edit specified in the pegRNA into the genome which can become fixed through a series of DNA repair processes.

Prime editing is effective in making substitutions and short insertions or deletions but struggles to make edits > 50 bp. The ability to make larger variants is arguably necessary to understand the organization of gene regulatory regions and genome organization principles. To expand the

capabilities of prime editing, two clever strategies have been implemented. In the first, prime editors insert one or multiple recombinase recognition sequences (discussed below) and subsequent recombinase expression catalyzes large-scale changes such as deletions, inversions, or integration of recombinase site flanked cargo DNA (Figure 1.5f) (Anzalone et al. 2022; Yarnall et al. 2022). With the second strategy, two prime editors are targeted to the opposite DNA strands to synthesize complementary flaps (Figure 1.5g) (Anzalone et al. 2022; Choi, Chen, Suiter, et al. 2022; T. Jiang et al. 2022). The flaps can anneal with one another and replace the original DNA sequence between the two target sites after DNA repair. Dual flap prime editing has been successfully implemented to delete hundreds of DNA bases or install insertions of up to 400 bp (Jinlin Wang et al. 2022).

The emerging field of prime editing is advancing rapidly, and we are beginning to understand the complex series of events required to make an edit, resulting in several improvements to the PE system (P. J. Chen and Liu 2022). The extensions of pegRNAs are vulnerable to degradation by cellular nucleases. Engineered pegRNAs (epegRNAs) fuse a structured RNA motif to the extension, which helps to stabilize it from degradation, increasing prime editing efficiencies by 1.5-4-fold (Nelson et al. 2022). In addition, improving the pegRNA scaffold for better binding to Cas9 and more effective transcription from polymerase III promoters further increases editing rates (Jost et al. 2020; B. Chen et al. 2013). At the genome, mismatch repair (MMR) limits the incorporation of smaller edits and prime editing strategies that circumvent MMR through the co-delivery of dominant MMR proteins greatly enhancing editing (P. J. Chen et al. 2021; Ferreira da Silva et al. 2022). Finally, refining the PE protein architecture through codon optimization, Cas9 mutations, linker choice, and an optimized nuclear localization sequence further elevates the editing rate (P. Liu et al. 2021; P. J. Chen et al. 2021).

Despite advances to the PE system, an important remaining downside of prime editing is its highly variable efficiency (Doman et al. 2022). pegRNA design is challenging due to many independent parameters that could be optimized and interact with one another (e.g. primer binding site length, reverse transcriptase template length, target site, edit type, edit position, spacer efficiencies, and so on). Fueled by data from large prime editing screens, several pegRNA prediction tools were developed (Weller et al. 2023; G. Yu et al. 2023; Mathis, Allam, Kissling, et al. 2023; H. K. Kim et al. 2021) that attempt to aid the design of active pegRNAs. However, due to the large combinatorial complexity of the parameters, each tool comes with separate advantages and drawbacks. A crucial gap in the field was that none of the data sets included a comprehensive analysis of short insertions (1-100 nt), which we attempted to fill (Chapter 2). Programmable insertion of short sequences is particularly useful for tagging proteins, fixing deletion mutations

by incorporating the missing sequence, and inserting recognition sites for site-specific recombinases.

1.4.3 Site-specific recombinases

While prime editing facilitates the precise insertion of short sequences (1-200 bp) and paired prime editing the deletion of moderately long sequences (thousands of bp), neither method is suitable for inversions or precise insertions/deletions at the scale necessary to interrogate genomes (thousands or millions of base pairs). Site-specific recombinases can fill this gap (Kilby, Snaith, and Murray 1993). Similar to restriction enzymes and CRISPR systems, site-specific recombinases were discovered as a component in the ancient arms race between bacteria and bacteriophages (Sternberg and Hamilton 1981; Abremski and Gottesman 1982). They mediate the circularization, integration, or excision of phage genome from host cells by recognizing two short DNA sequences on the phage and bacterial genomes, breaking them apart, and rejoining the ends in a new orientation (Sadowski 1986). Based on the active site amino acid residue used to catalyze this reaction, recombinases generally fall within two categories (Figure 1.6).

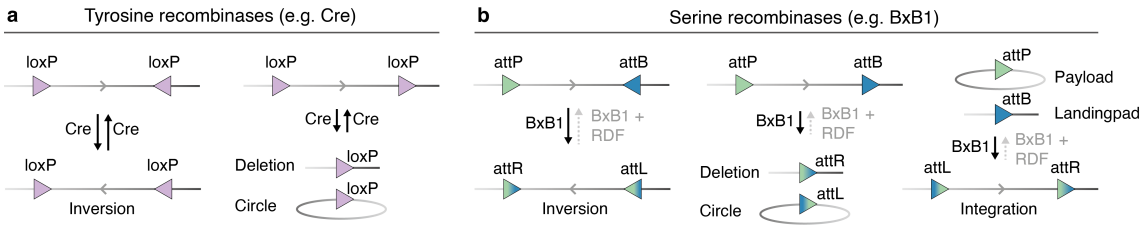

Figure 1.6: Rearrangements mediated by tyrosine and serine recombinases. a. Cre recombinase can make inversions (left) or deletions (right) depending on the orientation of the flanking loxP sites. The recombinase site remains unchanged throughout the process and therefore the reaction is reversible. **b.** Serine recombinases catalyze a rearrangement between an attB and an attP sequence, converting them into attL and attR hybrid sequences. Depending on the orientation and identity of the sites this can result in inversions (left), deletions (middle), or integrations (right). The rearrangement is unidirectional in the absence of recombinase directionality factors (RDFs). For clarity, only rearrangements in G1 and between the same molecule or a circular cargo are shown.

Tyrosine recombinases include the most widely used Cre recombinase derived from the P1 bacteriophage (Sternberg and Hamilton 1981). Cre drives the recombination between two 34 bp, near palindromic loxP sequences (locus of X-over P1). The outcomes can be different based on the orientation of the DNA ends joined and the relative location of the sites. If two loxP sites are on one DNA molecule, Cre will generate a deletion-minicircle between two sites of the same

orientation or an inversion for ones in opposite orientations. If the two sites are located on two different DNA molecules, recombination will generate translocations or integration of episomal DNA. More complicated events are possible in the S and G2 phases of the cell cycle owing to recombination between sister chromatids (Ramírez-Solis, Liu, and Bradley 1995). LoxP sites derive their orientation from seven central nucleotides that are non-palindromic. For an engineered version of the loxP site with palindromic central nucleotides (loxPsym), Cre stochastically catalyzes deletions or inversions (Hoess, Wierzbicki, and Abremski 1986). Additional mutations in the loxP site (lox5171 and lox2271) were engineered that recombined with themselves but not the wild-type sequence (G. Lee and Saito 1998), enabling orthogonal Cre-lox circuits. Other mutations can enable unidirectional rearrangements (similar to serine recombinases discussed below) (Araki, Araki, and Yamamura 1997). Their combined use with several similar but orthogonal recombinases (Flp, Dre, Vika) enables the construction of intricate genetic circuits (Golic and Lindquist 1989; Sauer and McDermott 2004; Karimova et al. 2013).

Most tyrosine recombinases rearrange two identical sites and the reaction is therefore bidirectional. In contrast, large serine recombinases recombine two distinct sequences, attB (attachment site from the bacteria) and attP (attachment site from the phage), resulting in a pair of recombinant sites (attL and attR) (Brown et al. 2011). This reaction is unidirectional in the absence of additional phage-encoded recombination directionality factors. Unidirectionality is ideal for integrating large DNA constructs into the genome or for making inversions (Liberante and Ellis 2021). Xu et al. characterized 15 different recombinases and determined BxB1 as the best candidate for integrating DNA into human genomes (Xu et al. 2013). More recent work systematically mined nucleotide sequences across the tree of life for novel recombinases and discovered additional candidates that are potentially more efficient than BxB1 or have other desirable properties such as having attachment sides that are already present in the human genome (Durrant et al. 2023).

Recombinases are a powerful tool to induce defined rearrangements into genomes. In early work, Ramires-Solis et al. demonstrated that it is possible to create megabase scale deletions, inversions, and duplications in mouse embryonic stem cells by targeting loxP site containing cassettes to the desired breakends and inducing Cre (Ramírez-Solis, Liu, and Bradley 1995). Similarly, rearranging two non-homologous chromosomes was leveraged to make mouse models of human leukemia-associated translocations (Buchholz et al. 2000; E. C. Collins et al. 2000). Subsequent work established that recombination efficiency on the same chromosome is ~10% for sites less than 10 Mb apart and drops to ~ 0.01% at 60 Mb, still much higher than for sites on non-homologous chromosomes (~0.001%) (Zheng et al. 2000). Such low-efficiency events can be enriched by adding split selection markers on the loxP cassettes that are reconstituted upon

successful recombination (Y. Yu and Bradley 2001). Recombination is also useful to make mouse models with conditional knockouts (Plück 1996). Here a gene is flanked by two loxP sites ('floxed') and the recombinase is expressed under the control of a tissue-specific or inducible promoter resulting in selective gene loss upon Cre expression.

One of the perhaps most impressive displays of recombinase-mediated genome engineering is the work by Lee et al. who completely humanized the mouse immunoglobulin locus (E.-C. Lee et al. 2014). To enable this, the authors first inserted lox sites and selection markers into the mouse immunoglobulin locus in mouse embryonic stem cells, creating a landing pad. Next, they retrieved bacterial artificial chromosomes (intermediates from the human genome sequencing project) that contain the human immunoglobulin locus, installed lox sites and selection markers, and integrated them into the mouse landing pad. Through clever design, integration of the payload creates a novel landing pad and mobilization with the PiggyBac transposase can excise used selection markers, freeing them up for the next round of integration. Thus, it was possible to iteratively integrate 2.7 Mb of human DNA into the mouse genome. Finally, to prevent the mouse variable segments from forming functional antibodies, the authors used Cre and targeting of recombinase sites to invert a ~20 Mb region at the heavy chain locus. This technology formed the backbone for the therapeutic antibody company Kymab which was acquired by Sanofi for $1.1 billion in 2021.

1.5 Engineering genomes at scale

Weaving the threads together, we have a genome and a solid map of its contents but an incomplete understanding of how this content underpins function. More specifically, we do not know how dispensable the vast non-coding and repetitive sequences between and within the islands of instructions are, and neither do we know how precisely the order and spacing of regulatory sequences and genes interplay to ensure faithful expression. However, building on decades of previous work, we now possess powerful molecular machines to precisely manipulate genomes, and combining these technologies might provide a path forward to address the most pressing unanswered questions.

The first powerful combinations are prime editing and recombinases. Recombinases have been held back by the need to first laboriously integrate their recognition sequences which prime editors could address, enabling structural manipulation of the genome at scale. In addition, this strategy is attractive because it avoids double-strand DNA breaks, causes few undesired on-target mutations, and, unlike HDR, works in non-dividing cells. Yarnall et al. optimized prime editor-recombinase fusion constructs for the one-pot insertion of a serine recombinase attachment site

followed by recombination and payload integration. This method, dubbed PASTE, enabled 10-30% integration efficiencies of payloads up to 36 kb (Yarnall et al. 2022). Anzalone et al. explored a similar strategy but focussed on using dual flap prime editing to install attachment sites and did not fuse the recombinase and prime editor (Anzalone et al. 2022). In addition, they demonstrated the successful inversion of a 40 kb region with single-digit efficiencies by co-targeting attB and attP sites to the variant breakends (Anzalone et al. 2022).

To effectively combine prime editing and recombinases, we first need to understand the determinants that govern how prime editors insert small sequences. In chapter two I use prime editing to systematically install thousands of small sequences into the genome, including recombinase recognition sites, and measure the efficiencies of each insert. Based on these data I could work out determinants of efficiency for writing small sequences with prime editing and uncover novel biology on how cells repair insertion prime edits.

The second ingredient for manipulating gnomes at scale is to make lots of changes at once. With CRISPR systems, each edit is usually specified by one unique guide or pegRNA. It is possible to make several edits to the genome by introducing multiple guides into one cell. For example, Chen et al. converted 33 instances of the TAG stop codon into TAA with base editing and a single transfection (Y. Chen et al. 2022). In chapter four I use this strategy to tile an enhancer cluster in the *OTX2* locus with recombinase sites to study the consequence of hundreds of inversion and deletions on gene expression.

Simultaneous transfections are fundamentally limited by the number of guides that can be introduced into the cell at once. An alternative to introducing many edits into the same genome is to target a promiscuous sequence. This strategy has been used to inactive 62 copies of porcine endogenous retroviruses in the pig cell genome to pave the way for safer trans-species transplants (Yang et al. 2015) and to understand the dynamics of DNA repair following Cas9 cleavage (Zou et al. 2022). In the most ambitious example of multiplexed editing, Cory et al. target catalytically dead base editors to LINE-1 retrotransposons to make more than 10,000 edits in a single genome. In chapter three, I target prime editors to LINE-1s to insert hundreds of recombinase sites into a single genome. Cre induction generates thousands of rearrangements throughout the genome which we could trace over time, illuminating the features of variants that survive compared to those that do not. Finally, surviving clones provided a unique resource with defined structural changes whose effect on gene expression we could study.

2

Determinants of efficiency for short sequence insertion using prime editing

Prime editing was just published before I started my PhD and we sought to leverage it to make structural changes in genomes at scale. While examples in the initial publication showed that it was possible to insert recombinase sites (Anzalone et al. 2019), no work had been done to systematically establish the cellular and sequence determinants governing short sequence insertions and how to optimize editing reagents for maximum efficiency. Our group had just published a manuscript on the prediction of mutations generated by Cas9 (Allen et al. 2018) and was working on another story about predicting base editing outcomes (Pallaseni et al. 2022). It was therefore natural for me to take on a project to deeply characterize short sequence insertions with prime editing. I designed a library of 3,604 sequences of various lengths and measured the frequency of their insertion into four genomic sites in three human cell lines, using different prime editor systems in varying DNA repair contexts. I found that length, nucleotide composition, and secondary structure of the insertion sequence all affect insertion rates. I also discovered that the 3′ flap nucleases TREX1 and TREX2 suppress the insertion of longer sequences. I designed the experiments and analyzed the data but had help from Elin Madli Peets for the execution of the experiments and teamed up with Juliane Weller who built a computational model to predict editing efficiencies based on the experimental results. Much of this chapter is adapted from ‘Prediction of prime editing insertion efficiencies using sequence features and DNA repair determinants’ on which I am the lead author (Koeppel et al. 2023).

2.1 Introduction

The efficient insertion of short DNA sequences into genomes could change the course of biotechnology and medicine (Anzalone et al. 2022; Yarnall et al. 2022). Small insertions can encode protein tags for purification and visualization, or manipulate protein function by altering protein localization, half-life, or interaction profiles. Integrating sequences for transcription factor binding sites and splicing modulators provides control over gene expression while introducing structural elements or recombinase sites can change DNA conformation and provide a substrate for large-scale engineering. For therapeutic opportunities, over 16,000 small deletion variants have been causally linked to disease (Landrum et al. 2016, 2018), and could in principle be restored by inserting the missing sequence (Geurts et al. 2021; Schene et al. 2020). A prominent example is cystic fibrosis, where 70% of cases are caused by a 3 nucleotide (nt) deletion (Drumm, Ziady, and Davis 2012; Zielenski and Tsui 1995). To enable reversing these mutations in practice, a technology must integrate insertions efficiently, accurately, and safely, avoiding the unintended outcomes and double-strand break stress that hampers existing Cas9-based therapies (Leibowitz et al. 2021; Allen et al. 2018; Kosicki, Tomberg, and Bradley 2018).

Prime editors can insert short DNA sequences without generating double-strand breaks or requiring an external template. They consist of a nicking version of Cas9 fused to a reverse transcriptase domain, which is complexed with a prime editing guide RNA (pegRNA) (Anzalone et al. 2019). The pegRNA comprises a primer binding site homologous to the sequence in the target, and a reverse transcriptase template that includes the intended edit, all in the 3’ extension of a standard CRISPR/Cas9 guide RNA. At the target site, Cas9 nicks one strand of the genomic DNA, which then anneals to the primer binding site on the pegRNA, and is extended by the Cas9-fused reverse transcriptase using the pegRNA-encoded template sequence. Next, DNA repair mechanisms resolve the conflicting sequences on the two DNA strands, ultimately writing the intended edit into the genome. Where CRISPR/Cas9 was compared to molecular scissors capable of disrupting target genes, and base editors were seen as molecular pencils for their ability to substitute single nucleotides, prime editors can be described as molecular word processors, able to perform search and replace operations directly on the genome (Anzalone, Koblan, and Liu 2020; G. Liu et al. 2022; P. J. Chen and Liu 2022; Doman et al. 2022).

The prime editing system is complex, and the determinants of its efficiency are not fully understood. Several partly independent steps, including three DNA binding events and successful DNA repair, are needed to produce an edit, each potentially influenced by the introduced sequence. In the largest study to understand these biases (at the time of initiating this chapter), Kim et al. comprehensively tested the consequences of varying the reverse transcription templates

and primer binding site lengths using a library of 55,000 pegRNAs. The editing rate increased with Cas9 sgRNA activity, GC content, and melting temperature of the primer binding site. While further optimization of sequences was possible, primer binding sites of 11-13 nt and reverse transcriptase templates of 10-12 nt had the highest average editing efficiencies (H. K. Kim et al. 2021).

The majority of libraries used by Kim et al. contained the same single nucleotide substitution 5 nt upstream of the nick site. Similarly, nearly all investigations of prime editing efficacy to date have predominantly focused on single nucleotide substitutions (Anzalone et al. 2019; H. K. Kim et al. 2021; Kweon et al., 2021; Y. Liu et al. 2020; P. J. Chen et al. 2021; Nelson et al. 2022). Of the many possible useful sequences in molecular biology, only a handful have been introduced with prime editing. Therefore, in contrast to a relatively deep understanding of Cas9 mutagenesis (Allen et al. 2018; Doench et al. 2016; Meier, Zhang, and Sanjana 2017; H. K. Kim et al. 2019) and base editing outcomes (Pallaseni et al. 2022; Arbab et al. 2020; Song et al. 2020) very little is known about how the inserted sequence affects efficiency, and the length range of insertions feasible by prime editing has not been defined.

I measured the insertion efficiency of 3,604 sequences in several target sites and a variety of cellular and repair pathway contexts and found that insertion sequence length, nucleotide composition, and secondary structure all affect insertion efficiency. Moreover, I define the precise effect of mismatch repair on thousands of insertion sequences and discover that overexpression of the 3’ flap nucleases *TREX1* and *TREX2* abolished the insertion of longer sequences. Together, sequence features and repair pathway activity explained most of the variation in insertion rate. Juliane Weller then used these insights to train a sequence-based prediction model informed by mismatch repair efficiency that predicts editing outcomes for novel sequences with high accuracy and demonstrated the models' usefulness for the selection of optimal reagents for new insertions.

2.2 Results

To systematically assess insertion rates using prime editing, I first established faithful prime editing systems in HEK293T and HAP1 cells for the integration of a single loxP site. Once these were in place, I could scale the experiment to measure the incorporation of thousands of sequences into several genomic target sites and use the resulting data set to understand the sequence, target site, repair, and prime editing system features that govern insertion efficiencies.

2.2.1 Establishing prime editing systems for short sequence insertion

I first tested the ability of prime editing to insert recombinase site versions into the *HEK3* locus in HEK293T cells by transient transfection of the prime editing reagents. I assessed editing efficiency using capillary gel electrophoresis (Figure 2.1a) and amplicon sequencing (Figure 2.1b). The 34 bp core loxP site was integrated at 23% efficiency. LoxP site versions that were extended by 10 bp on one or both sides were integrated at substantially lower rates (6% for a 44 bp version and 3% for a 54 bp version, Figure 2.1a,b). Notably, indel rates were <1% and indistinguishable from a non-transfected control, demonstrating the high on-target precision of prime editing (Figure 2.1b). The discrepancy in rates highlighted the importance of the edit sequence on incorporation efficiencies and I sought to characterize the rules governing short sequence insertions with prime editing more generally.

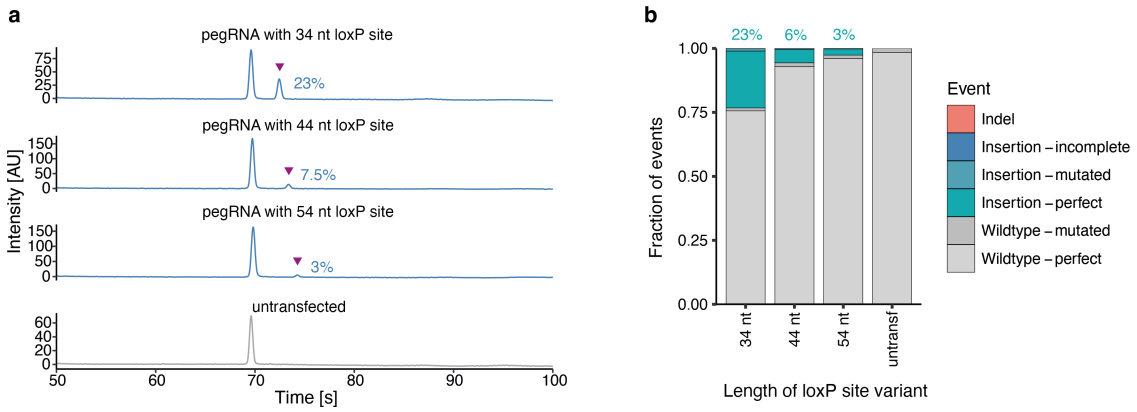

Figure 2.1. Variable insertion efficiencies for loxP site variants. a. Bioanalyzer traces for PCR amplicons of cells that were transfected with prime editor and the indicated pegRNA constructs (panels). The purple triangle indicates the peak corresponding to an insertion and the number next to it is the molecular frequency of the insertion product to overall DNA. **b.** Quantification of the frequency (y-axis) of mutation types (colors) from next-generation sequencing of PCR amplicons (x-axis, as in a).

Besides the desired sequence, the target site as well as the editing cell line and DNA repair context could all influence insertion efficiencies. Therefore, I decided to also test prime editing in an additional cell line and two different DNA repair contexts. The near haploid HAP1 cell line is interesting for genome engineering due to the absence of a confounding second allele (Blomen et al. 2015). Based on the finding that mismatch repair inhibits the incorporation of short edits (P. J. Chen et al. 2021; Ferreira da Silva et al. 2022), I additionally sought to establish prime editing in HAP1 cells that are knockout for the mismatch repair protein MLH1 (hereafter referred to as HAP1 Δ*MLH1*). Since HAP1 cell lines are difficult to transfect, I designed a doxycycline-inducible prime editor on a PiggyBac transposon (Figure 2.2a) and clonal HAP1 and HAP1 Δ*MLH1* cell lines with stably integrated prime editor were derived. To test the ability of these cell lines to make edits, Mélanie Gouley, a master student I supervised, infected them with a lentivirus with an engineered pegRNA encoding for a loxP site integration into the *HEK3* locus. Both HAP1 and HAP1 Δ*MLH1* cells achieved integration efficiencies between 55-78% on day 10 and 72-85% on day 13 respectively (Figure 2.2b) demonstrating high prime editing activity.

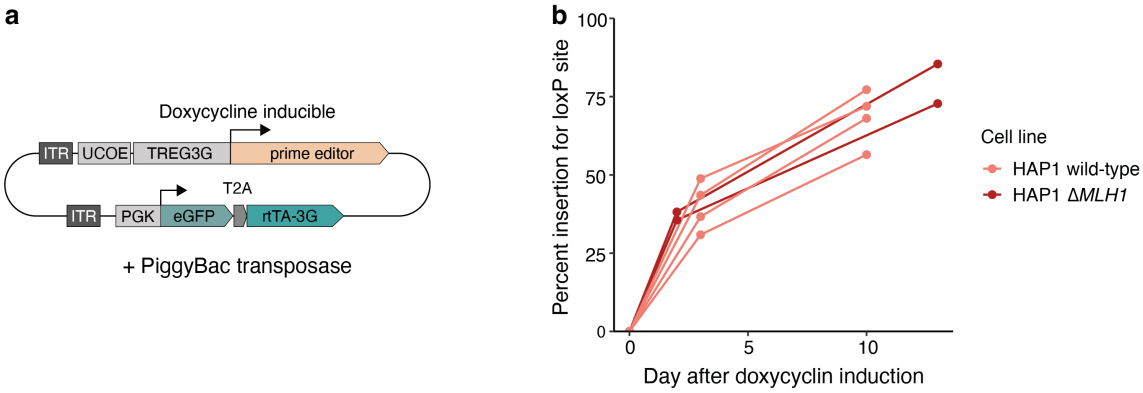

Figure 2.2. Stable prime editing cell lines. a. Schematic of the doxycycline-inducible prime editor expression construct on a PiggyBac transposon. *UCOE*: Universal chromatin opening element. *TREG3G*: Doxycycline-inducible promoter. rtTA-3G: transactivator protein. ITR: Inverted terminal repeats. **b.** Integration efficiencies (y-axis) for a loxP site into the *HEK3* locus over 13 days (x-axis) in six clones (markers and lines) stratified by cell line (colors).

2.2.2 Systematic characterization of insertion efficiencies

To systematically characterize how the length and composition of inserted sequence, as well as cell line, target site, and the version of the prime editor system, affect insertion rates, I designed 3,604 pegRNAs encoding insertions immediately upstream of the nick site. These comprise 270 sequences useful for molecular biology (e.g. His-6 tag, recombinase sites, and mNeonGreen11 (Feng et al. 2017)), 1,957 eukaryotic linear motifs (Dinkel et al. 2014, 2016; Puntervoll et al. 2003), 439 sequences with variable secondary structure, all single nucleotides, dinucleotides,

trinucleotides, tetranucleotides, and one hundred random sequences of each length between 5 and 10 nt (Figure 2.3a). Insertions ranged from the length of 1 to 69 nt and varied in GC content (Figure 2.3b), while the primer binding site and homology arm lengths in the pegRNA were fixed to 13 and 34 nt, respectively. The libraries were delivered by lentivirus against four target sites (three previously tested: *HEK3*, *EMX1*, *FANCF* (Anzalone et al. 2019) and the safe-harbor *CLYBL* locus (Cerbini et al. 2015)) in two cell lines (HEK293T and HAP1), followed by transient transfection of the prime editor 2 plasmid (HEK293T cells) or doxycycline induction of PiggyBac transposase integrated prime editor (HAP1 cells, Figure 2.2a), five days of selection, and sequencing of two amplicons from the cell pool, one of the targeted locus and one of the pegRNA locus (Figure 2.3c). I calculated insertion efficiencies as the fraction of reads in the target site amplicon with a given insertion divided by the fraction of reads for the pegRNA encoding it in the pegRNA amplicon and analyzed them as the main statistic in the rest of the study.

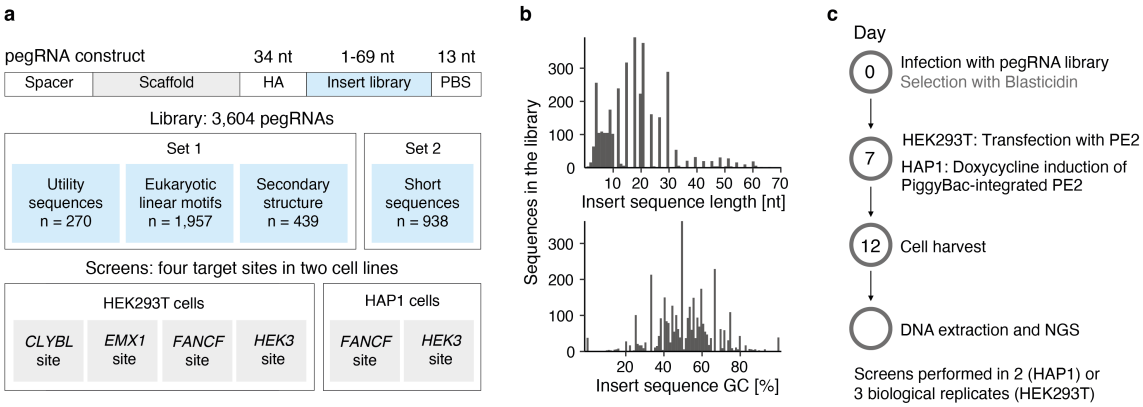

Figure 2.3. A method for high-throughput measurement of prime insertion efficiencies. a. Screen setup. Set 1 and Set 2 libraries were screened separately and data merged (Methods); panels d-f reflect Set 1 results only. **b.** Library composition. The number of sequences in the library (y-axis) with different insert sequence lengths (x-axis, top panel) and %GC content (x-axis, bottom panel). **c.** Experimental design.

Insertion efficiencies of sequences varied widely. The top 5% of templates were inserted 27-134 times more efficiently than the bottom 5% across the various target site and cell line combinations (Figure 2.4), indicating substantial sequence-dependent variation. The insertion rates were consistent across two biological replicates in HAP1 and HAP1 Δ*MLH1* cells and three biological replicates in HEK293T cells (median R=0.70; Figure 2.5), but differed in magnitude across screens (average across pegRNAs 0.18% for the *CLYBL* locus in HEK293T to 6.7% for the *HEK3* locus in HEK293T cells, Figure 2.6a). The *CLYBL* and *EMX1* target sites with lower overall insertion efficiencies, and subsequently lower effective coverage, showed weaker replicate correlations compared to the *HEK3* locus with high insertion rates (Figure 2.5).

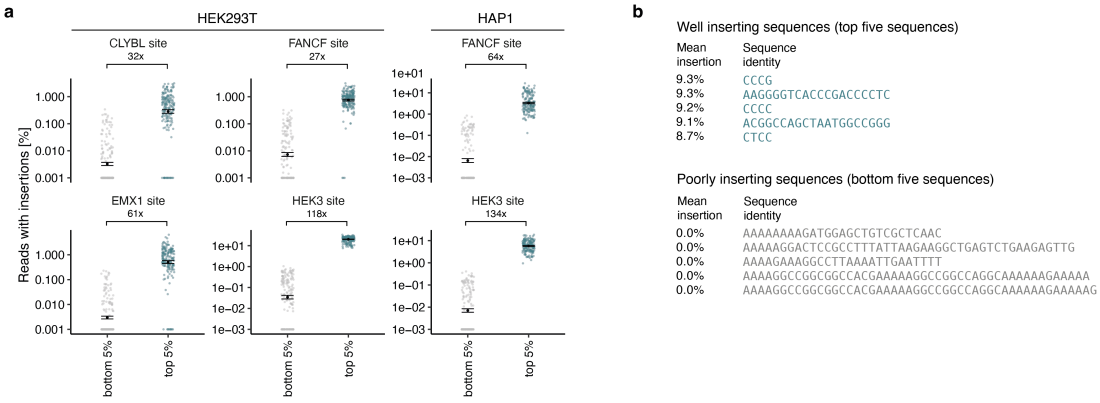

Figure 2.4. Insertion rates differ substantially between sequences. a. Screen normalized insertion efficiency (y-axis) for the top 5% of pegRNAs with the highest insertion rates across all screens and the bottom 5% of pegRNAs (x-axis, colors). Markers are individual pegRNAs. Data are presented as mean values +/- standard error of mean. *n* = 3 biological replicates. **b.** Example of well and poorly inserting sequences with their respective mean insertion rates across screens.

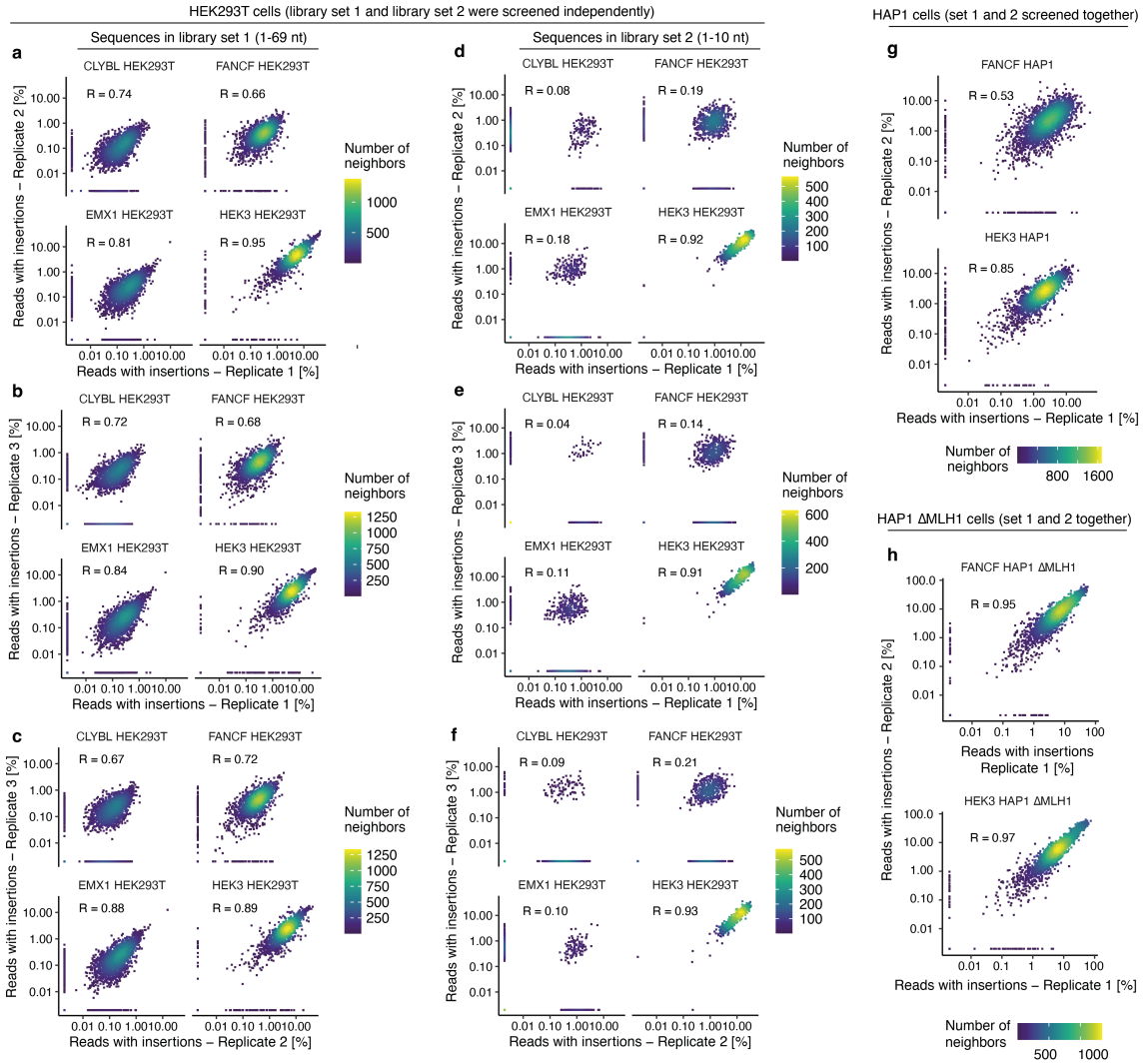

Figure 2.5 Reproducibility across large-scale prime editing insertion screens. a. Percent insertion in replicate 1 (x-axis) compared to percent insertion in replicate 2 for insert sequences of the library Set1 (markers) for different target sites (panels) in the HEK293T cell line. **b-c.** As in (a) but for different replicate comparisons. **d-f.** As in (a-c) but for the library Set2. **g.** As in (a) but for the HAP1 cell line and screening library set 1 and library set 2 together. **h.** As in (a) but for the HAP1 *ΔMHL1* cell line and screening library set 1 and library set 2 together.

To understand the consistency of insertion efficiencies across contexts, I next compared them between replicates, cell lines, and target sites. Insertion rates into the same target site in different cell lines were more correlated (mean R=0.52) than into different target sites in the same line (mean R=0.38). The correlation was weakest when both the target site and cell line differed (mean R=0.17, Figure 2.6b-c), demonstrating both target sequence-specific and cell line-dependent biases on insertion rates.

Unintended editing outcomes included single base mutations, small insertions, and deletions around the nicking site, deletions overlapping the primer binding site and reverse transcription template, insertion of mutated library sequences, duplications of the reverse transcription template, as well as partial scaffold integrations (Figure 2.6a, Figure 2.7a-c). These outcomes were rare overall (0.06%-0.45%). Base changes at the target site (potentially arising from errors of the reverse transcriptase) were infrequent in reads with and without insertions (0.038% vs 0.030%) but slightly elevated upon insertion immediately downstream of the nick site and for the first nucleotides after the end of the homology arm (Figure 2.7d-f). Overall, the intended insertions were the dominant mutations generated, and I do not consider the unintended edits further.

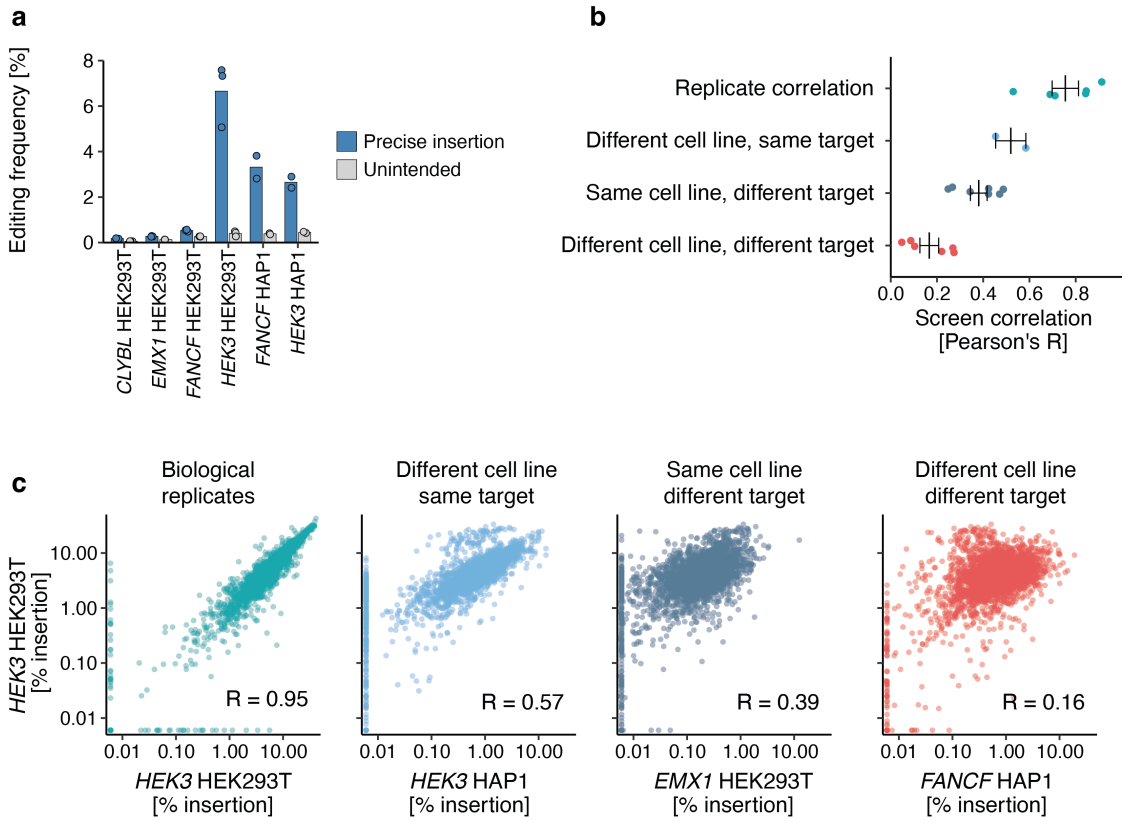

Figure 2.6. Target-site and cell line affect prime insertion efficiencies. a. Editing frequencies. Average mutation frequency (y-axis) for different screens (x-axis) stratified by mutation type (blue: insertions; gray: unintended outcomes). Markers represent one replicate and bars the average across $n = 3$ biological replicates. **b.** Pearson's R between replicate correlations or insertion rates in two screens (x-axis) for different comparisons (y-axis, colors). Markers: correlation value of one pair of screens (for replicate correlations, mean of pairwise comparison across $n = 3$ biological replicates); line and whiskers: mean and standard error of the mean. **c.** Examples of categories from (b). Percent insertion in the *HEK3* locus in HEK293T cells (y-axis) compared to values (x-axis) in other contexts (panels, colors) for insertion sequences (markers). Left panel: comparison of biological replicates; other panels: comparison of replicate averages. Label: R of values in linear scale. Colors: as in (b).

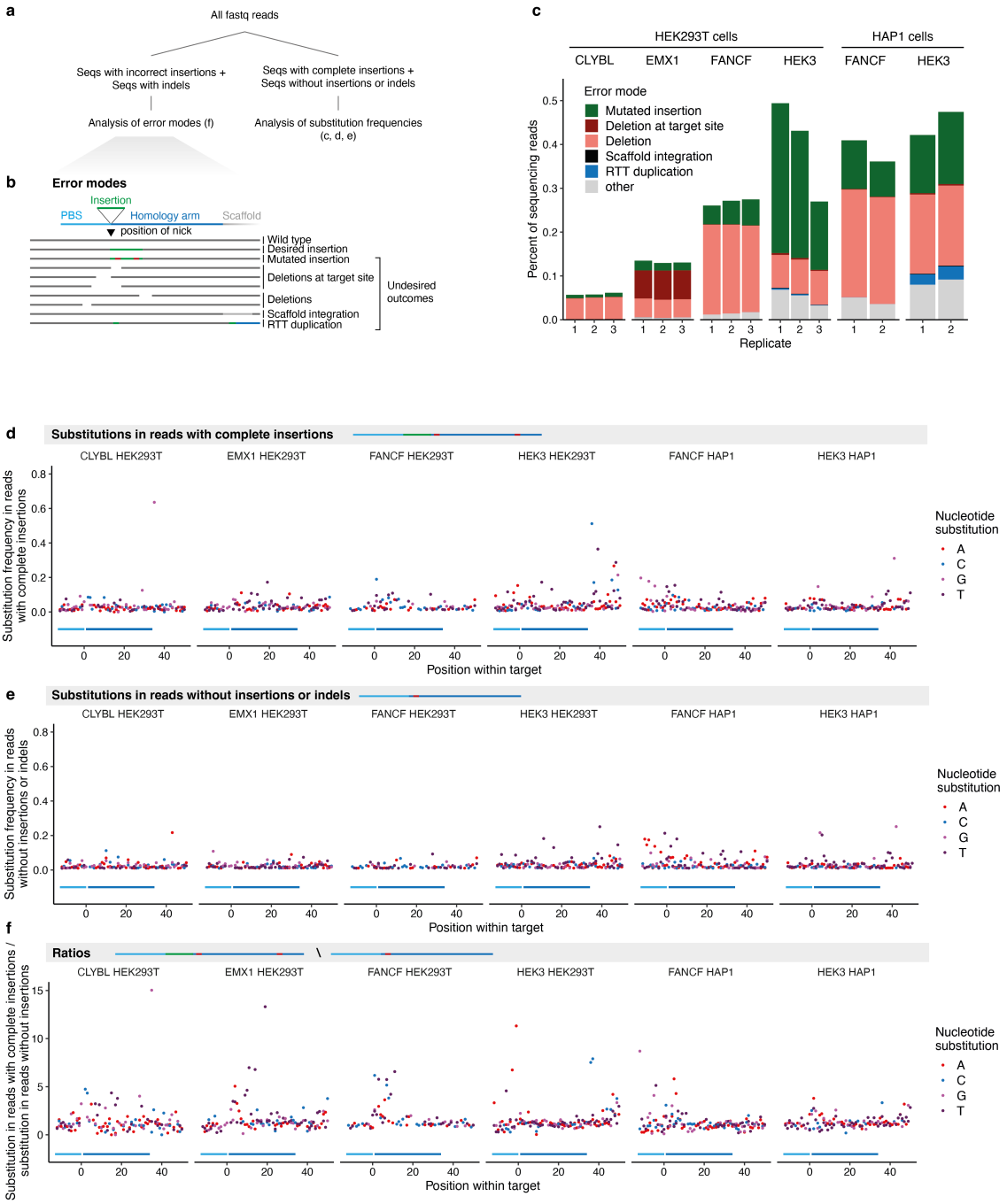

Figure 2.7 Error modes in prime editing screens. a. Schematic of the analysis for unintended outcomes. **b.** Schematic of the various analyzed error modes. RTT: reverse transcriptase template. **c.** Frequencies of unintended outcomes (y-axis) stratified by error types (colors) for replicates (x-axis) at various target sites and cell lines (panels). **d.** The average percentage of sequencing reads with complete library insertions (y-axis) with a non-reference sequence nucleotide (colors) at positions relative to the nicking site (x-axis). $n = 3$ biological replicates for HEK293T cells and $n = 2$ biological replicates for HAP1 cells. **e.** As (d) but instead showing reads without insertions or indels. **f.** As (d) but displaying the fold-changes between the averages for reads with complete insertions and for reads without insertions or indels.

2.3.1 Insert size and mismatch repair activity effects

Given the repeatable sequence-dependent variation in insertion rates that spans over three orders of magnitude, I sought to understand the responsible features, starting with insert length. Insertion frequency did not decrease monotonically with insert length in HEK293T cells, but instead, had two modes of high values. First, sequences of 3 and 4 nt were inserted on average 2.0-4.1 times more efficiently than others across the four targeted sites (Figure 2.8a). Second, sequences between 15 and 21 nt were inserted on average 1.3-1.6 times more efficiently than 10-14 nt ones, and 1.5-2.0 times more efficiently than sequences longer than 21 nt (Figure 2.8a). These relative biases in efficiency were shared between all target sites, despite a 20-fold range of their average insertion rates. Inserts longer than 45 nt were incorporated less frequently, at a screen average rate that is 4 to 8 times lower than that of sequences shorter than 45 nt. The longest sequence that was inserted at > 1% frequency (1.4%, *HEK3* site in HEK293T cells) was 66 nt, demonstrating that integration of moderately long sequences is feasible with prime editing.

In contrast to HEK293T cells, the insertion frequency of the short 1-4 nt sequences was not substantially higher than that of longer ones in HAP1 cells (0.60-1.27 times, Figure 2.8b). This reduced the concordance of insertion rates in the two cell lines at the same site (R = 0.41 for *FANCF* and 0.54 for *HEK3*, Figure 2.8c-d) compared to replicates (median R=0.78, Figure 2.5). One possible explanation is mismatch repair (MMR) proficiency, since HEK293T cells are partly MMR deficient due to promoter methylation of the *MLH1* gene (Trojan et al. 2002), while HAP1 cells are not. The MMR pathway recognizes and excises short mismatches of less than 13 nt and could therefore remove short insertions in HAP1 cells before the nicked strand is re-ligated (Gupta, Gellert, and Yang 2011). Indeed, mismatch repair antagonizes prime editing for substitutions and short insertions (P. J. Chen et al. 2021; Ferreira da Silva et al. 2022). Consistent with this explanation, I observed strong correlations between insertion rates in HAP1 and HEK293T cells for sequences longer than 13 nt that are not affected by mismatch repair (R=0.78 for the *FANCF* locus and 0.91 for the *HEK3* locus, Figure 2.8c-d).

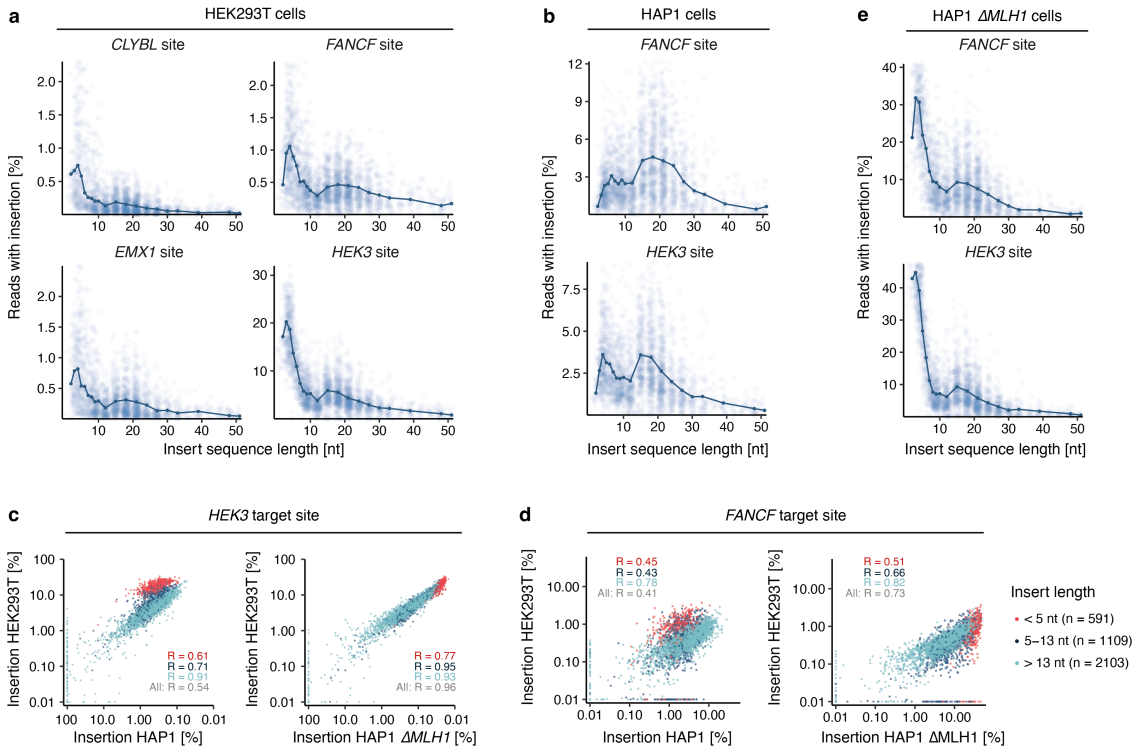

Figure 2.8 Prime insertion efficiency depends on insert length and MMR. a. Insertion rate in HEK293T cells. Percent reads with insertion (y-axis, cut-off at 3 standard deviations above mean) for different insert sizes (x-axis) of individual sequences (blue markers) and averages for lengths with at least 30 measured sequences (dark blue line and markers) at different target sites (panels). Data represent the average of $n = 3$ biological replicates. **b.** As (a), but for HAP1 cells. **c.** Insertion rate in one cell context (y-axis) compared to in another context (x-axis) at the *HEK3* target of individual sequences (markers), comparing HEK293T to HAP1 cells (left panel) and HEK293T cells to HAP1 *ΔMLH1* cells (right panel). Red: short sequences (up to 4 nt); blue: medium sequences (5-13 nt); teal: longer sequences (>13 nt). Label: R between rates. The data are an average from $n = 3$ biological replicates (HEK293T) or $n = 2$ biological replicates (HAP1). **d.** As (c) but for the *FANCF* target. **e.** As (a), but for HAP1 *ΔMLH1* cells.

To experimentally test the hypothesis that rates of inserting short sequences differ between cell lines due to mismatch repair activity, the *HEK3* and *FANCF*-targeted libraries were screened in HAP1 cells that are knockout for *MLH1* (HAP1 Δ*MLH1*, Figure 2.5h, Figure 2.8e). I found that the average insertion rates of 1-4 nt sequences were most affected by the knockout, increasing by 7.2-11 fold, while the rates of 5-13 nt sequences increased by 2.1-2.7 fold (Figure 2.9a). Overall, 66% (*HEK3*) and 67% (*FANCF*) of the variance in the fold changes (Figure 2.9b-c) were explained by a model where the loss of MMR increases the insertion rate of 1 nt sequences by 23-28 fold and drops 40-48% for every additional nucleotide. The low correlations of insertion rates between HEK293T and wild-type HAP1 cells (R = 0.41-0.54) also improved to close to replicate concordance when matching MMR status (R = 0.73-0.96 between HEK293T and HAP1 Δ*MLH1*

cell lines, Figure 2.9d-e). In summary, our findings highlight that MMR proficiency is the major source of independent variation between the tested cellular contexts for prime insertion of short sequences.

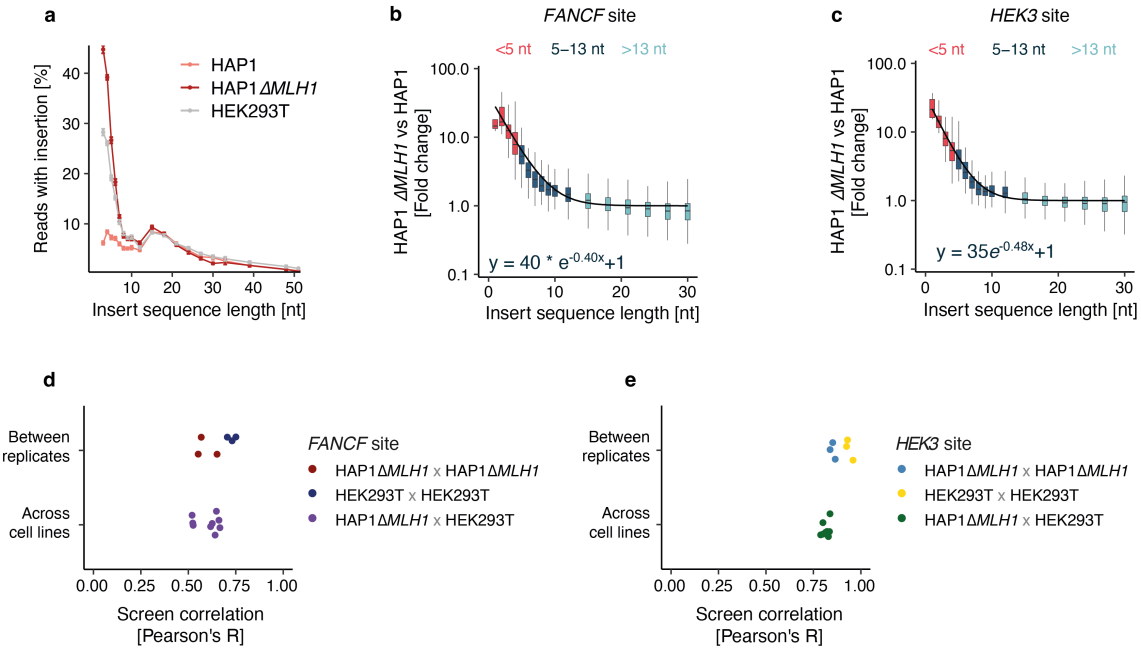

Figure 2.9. The effects of mismatch repair on sequence insertion with prime editing. a. Average insertion rates (y-axis) across insert lengths (x-axis) with at least 30 measured sequences in various cell line contexts (colors). Data are presented as mean values +/- standard error of mean. *n* = 3 biological replicates (HEK293T) or *n* = 2 biological replicates (HAP1). **b.** The ratio of relative insertion rates (Methods) at the *FANCF* locus between HAP1 *ΔMLH1* and HAP1 cells (y-axis) for different lengths (x-axis) stratified by insert sequence lengths (colors). Box: median and quartiles; whiskers: least extreme of 1.5 times the interquartile range from the quartile and most extreme values. Line: fit from an exponential model (ratio ~ a * exp(-b*length)+1). *n* = 2 biological replicates. **c.** As (b) but for the *HEK3* site. **d.** Replicate concordance and concordance between different cell lines for insertions into the *FANCF* target. Pearson's R between insertion rates in two screens (x-axis) for different comparisons (y-axis, colors). Markers: correlation value of one pair of screens. **e.** As (d) but for the *HEK3* target.

2.2.2 Effects of prime editing steps

Having confirmed mismatch repair as a length-dependent determinant of insertion efficiency, I next sought to understand how different steps of prime editing affect insertion rates of sequences in the library. Specifically, I dissected the contributions of (1) pegRNA expression (2) reverse transcription by two different reverse transcriptases (3) the presence of a nicking guide, and (4) overexpression of 3' and 5' flap nucleases (Figure 2.10).

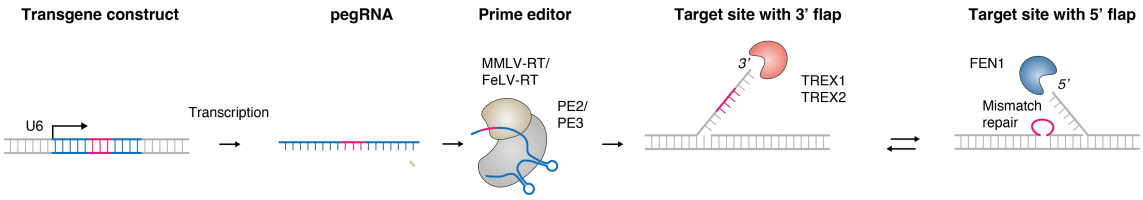

Figure 2.10. The molecular steps of prime editing. Schematic of molecular steps involved in prime editing.

I first assessed expression levels of pegRNAs targeting the *HEK3* site in HEK293T cells using deep sequencing. Abundance in the transcriptome was well correlated between replicates (median R=0.97, Figure 2.11a) and with the DNA-derived read-count frequency (R=0.56, Figure 2.11b). The exceptions were sequences that resulted in four or more consecutive thymines on the pegRNA cassette (adenines in the inserted DNA), which act as transcription terminators for RNA polymerase III (B. Chen et al. 2013; Porrua, Boudvillain, and Libri 2016). Upon removing pegRNAs with terminator motifs, the correlation between measured DNA and RNA sequence coverage increased to 0.59 (Figure 2.11b). Sequences with four or more consecutive adenines were 4.8-fold less expressed and accordingly, their average insertion rate was 4.8-fold lower compared to other sequences (Figure 2.11b-c). Overall, 23 of the 24 inserts (96%) that were not observed in any screen contained at least one run of four or more adenines, highlighting this feature as a useful filter in pegRNA design.

Second, to disentangle the contribution of the reverse transcription step, Fabio Liberante cloned a prime editor construct with the nicking Cas9 fused to an engineered feline leukemia virus reverse transcriptase (MashUp RT - pipettejockey.com) with similar fidelity to the murine leukemia virus (MMLV) one used in PE2. I included this construct into the prime insertion screens and observed 6.7-fold lower average insertion rates compared to the standard PE2 (0.72% and 4.86% respectively; Figure 2.12a-d) but with high correlation to PE2 (R = 0.80; Figure 2.12e). Therefore, the effects of the insert sequence on insertion are not specific to the murine reverse transcriptase used in PE2 and highlight the possibility of performing prime editing experiments with alternative constructs.

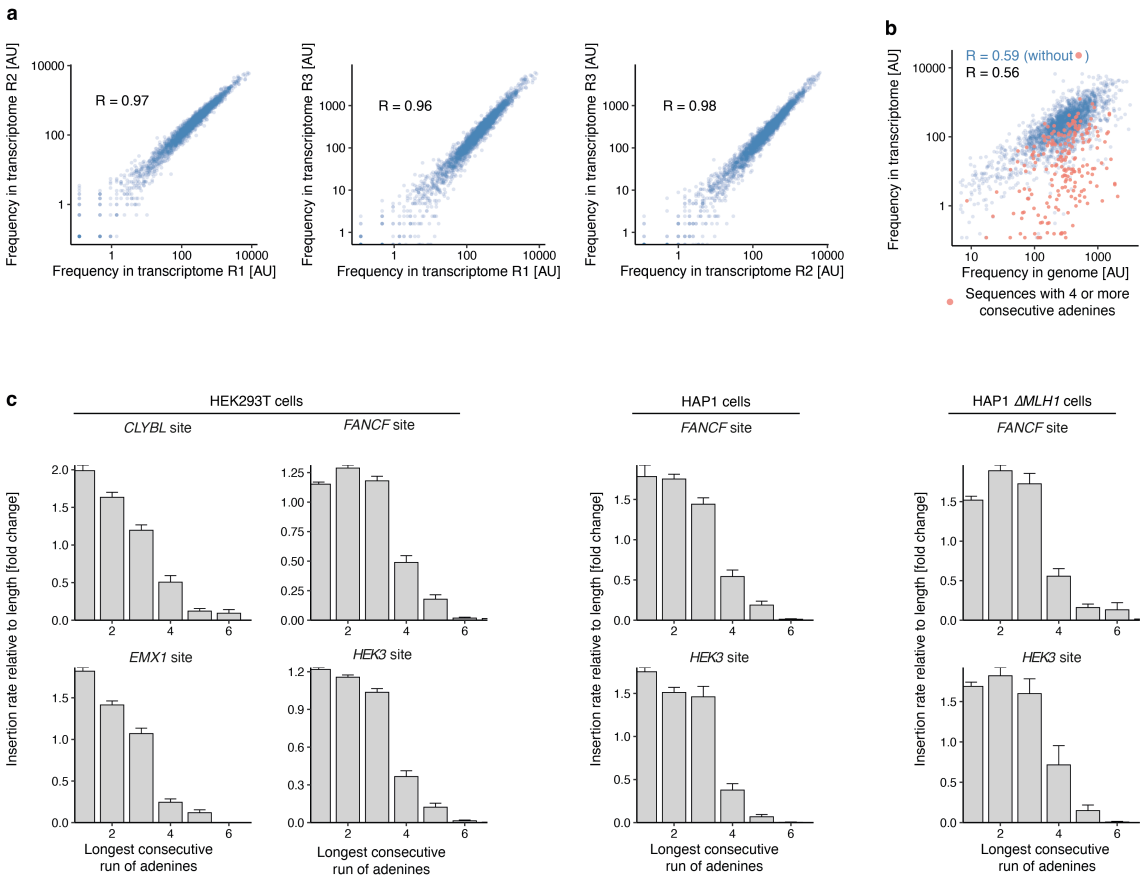

Figure 2.11. Consecutive runs of adenines in the insert decrease pegRNA expression and insertion efficiency. a. Normalized pegRNA read counts from the transcriptome in one replicate (x-axis) compared to another replicate (y-axis) for insert sequences (markers) and different pairwise combinations (panels). **b.** Normalized pegRNA count derived from sequencing of PCR amplicons from genomic DNA (x-axis) or PCR amplicons from RNA (y-axis) for the *HEK3* site in HEK293T cells for individual pegRNAs (markers). Pink: inserts with four or more consecutive adenines. Data represent the average of $n = 3$ biological replicates. **c.** Average insertion rate relative to length bin median (y-axis) for inserts stratified by the longest consecutive run of adenines (x-axis). Panels show various target sites and cell lines. Data are presented as mean values +/- standard error of mean. $n = 3$ biological replicates for HEK293T cells and $n = 2$ biological replicates for HAP1 and HAP1 *ΔMLH1* cells.

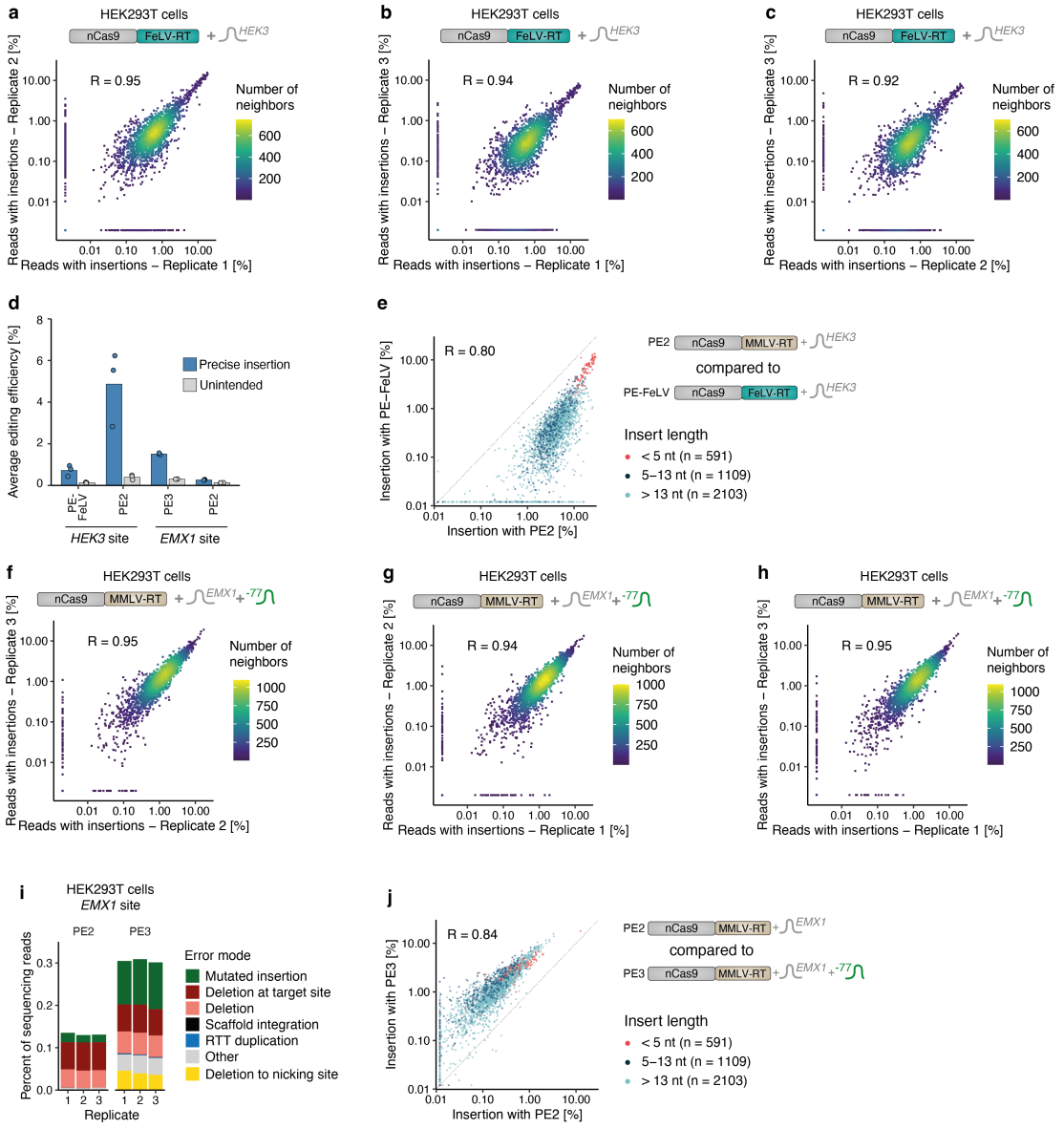

Figure 2.12 Prime editor compositions affect overall but not relative insertion rates. a. HEK293T cells expressing nicking Cas9 fused to the Feline Leukemia Virus (FeLV) Reverse Transcriptase. Comparing percent reads with insertions in replicate 1 (x-axis) to replicate 2 (y-axis) for library Set1 insert sequences targeting the *HEK3* site (markers). **b.** As (a) but comparing replicates 1 and 3. **c.** As (a) but comparing replicates 2 and 3. **d.** Editing frequencies for alternative prime editing systems. Mutation frequency (y-axis) for three biological replicate screens (markers) using different prime editor systems (x-axis) stratified by mutation type (blue: insertions; gray: unintended outcomes). Bar: average of markers. **e.** Insertion frequencies at the *HEK3* site in HEK293T using the standard MMLV reverse transcriptase (PE2, x-axis) and the FeLV reverse transcriptase (PE-FeLV, y-axis) for different insertion sequences (markers). Colors: number of neighboring points. $n = 3$ biological replicates. **f.** HEK293T cells expressing PE2 and a nicking guide RNA that targets 77 nt downstream. Comparing percent reads with insertions in replicate 1 (x-axis) to replicate 2 (y-axis) for library Set1 insert sequences targeting the EMX1 site. **g.** As (f) but comparing replicates 1 and 3. **h.** As (f) but comparing replicates 2 and 3. **i.** Frequencies of unintended outcomes (y-axis) stratified by error types (colors) for replicates (x-axis) at the EMX1 target sites comparing cells without nicking guide RNA (PE2) and with nicking guide RNA (PE3) (panels). **j.** As (e) but comparing PE3 and PE2 at the *EMX1* site.

The PE3 system includes an additional sgRNA to nick the non-edited strand, which increases editing efficiency as well as indel formation rate (Anzalone et al. 2019). I explored how the addition of this extra nicking sgRNA affects the insertion frequencies of sequences in the library. I chose the *EMX1* locus in HEK293T cells where the insertion efficiencies of 0.28% on average were poor without the nicking guide RNA and co-transfected a nicking guide RNA that targets 77 nt downstream of the pegRNA target (P. Liu et al. 2021). The extra nick increased the average insertion rate by 5.6-fold to 1.5% (Figure 2.12d, Figure 2.12f-h), and increased the indel rate by 2.3-fold to 0.31%, including deletions between the nick sites of the pegRNA and sgRNA that were not observed for PE2 (Figure 2.12i). Importantly, the relative insertion rates for sequences in the library were highly concordant between PE2 and PE3 in HEK293T cells (R=0.84, Figure 2.12j).

An important step in prime editing is to resolve between the intermediates with a 5' flap (containing the wild-type sequence) or a 3' flap (containing the insertion) that compete. The activity of the respective flap nucleases might steer the balance between the two outcomes. To test this, the 5' flap nuclease *FEN1* and the 3' flap nucleases *TREX1* and *TREX2* were overexpressed in the context of the *HEK3* site targeting screen in HEK293T cells. As a control, *eGFP* was overexpressed in the same backbone used for the nucleases. The insertion rates after *FEN1* or *eGFP* overexpression were highly correlated to those measured in screens without overexpression (R=0.93 and 0.97, Figure 2.13, Figure 2.14a,b) with similar length dependence. Intriguingly, *TREX1* and *TREX2* overexpression abolished the insertion of longer sequences. For cells that did not overexpress nucleases or overexpressed *eGFP*, the average insertion rate for sequences longer than 4 nt was 4.4-6.0% which is 4.4-5.8 times less than for shorter sequences. The relatively high insertion rates of longer sequences are in contrast to cells overexpressing *TREX1* and *TREX2*, where the average insertion rate for sequences > 4 nt was only 0.66% or 0.97%; 25.3-26.7 fold lower than that of shorter ones (Figure 2.14c,d).

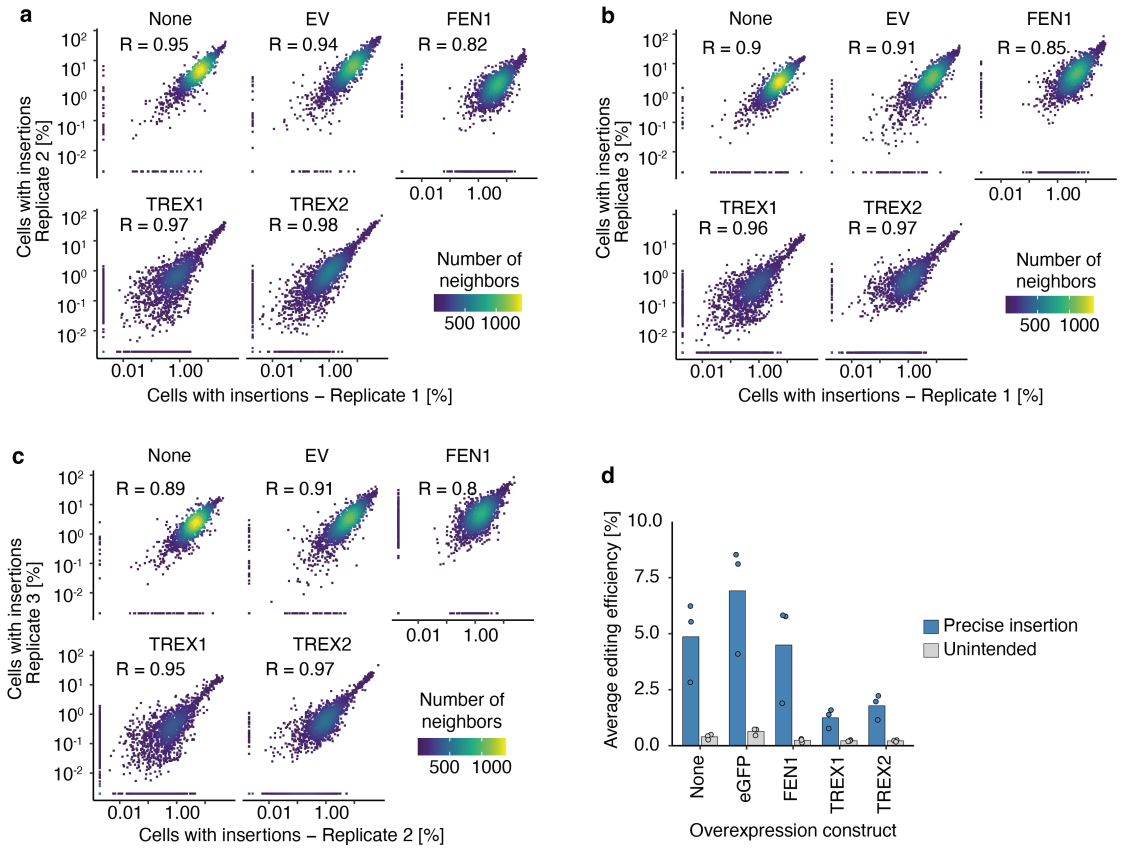

Figure 2.13. Reproducibility across prime editing insertion screens with flap nuclease overexpression. a. HEK293T cells overexpressing various constructs (panels). Comparing percent reads with insertions in replicate 1 (x-axis) to replicate 2 (y-axis) for library Set1 insert sequences targeting the *HEK3* site. **b**. As (a) but comparing replicates 1 and 3. **c.** As (a) but comparing replicates 2 and 3. **d.** Editing frequencies for screens with overexpression constructs. Mutation frequency (y-axis) for three biological replicate screens (markers) using different prime editor systems (x-axis) stratified by mutation type (blue: insertions; gray: unintended outcomes). Bar: average of markers.

To confirm that *TREX1* and *TREX2* antagonize prime insertions in a length-dependent manner, I co-transfected HEK293T cells with overexpression constructs encoding eGFP, *TREX1*, or *TREX2* and individual pegRNAs targeting the *HEK3* site encoding a 1, 3, 9, or 30 nt insertion (C, CAG, *BCL6* binding site, and Myc-tag) in the context of 25 or 34 nt homology arms. Overexpressing *TREX1* and *TREX2* decreased editing rates across all insert and homology arm lengths, but disproportionately more for longer inserts (1.6-3.0 fold for the 1 nt insertion compared to 20-108 fold for the 30 nt insertion, (Figure 2.15a-b). This effect could be driven by the length of the insert sequence alone or of the entire 3' flap (corresponding to insertion + homology arm). In line with the results from the pooled screens (Figure 2.14d), I observed a strong correlation between the log fold change of insertion rates for *TREX1/2* over *eGFP* with the insert sequence length (R=0.97) which decreased when considering the total extension length (R=0.86-0.92, Figure 2.15c), suggesting a more important role for the insertion length than the overall flap length.

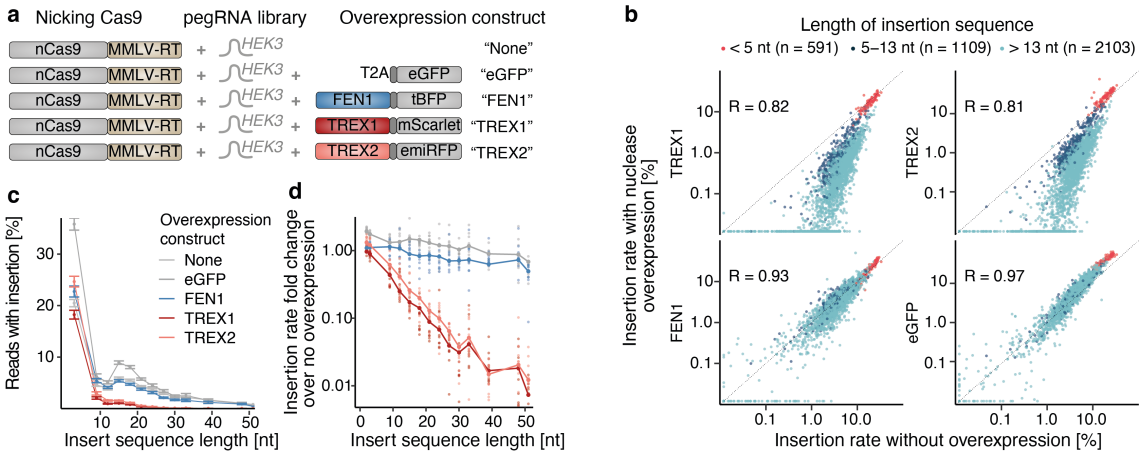

Figure 2.14. TREX1 and TREX2 antagonize the insertion of long sequences. a. Schematic of screens with overexpression constructs. **b.** Insertion frequencies for different overexpressions (y-axis and panels) compared to no overexpression (x-axis) for three biological replicate screens (markers) stratified by insertion sequence lengths (colors). **c.** Average insertion rates (y-axis) across insert lengths (x-axis) with at least 30 measured sequences for overexpression constructs (colors). Data are presented as mean values +/- standard error of mean. $n = 3$ biological replicates. **d.** As (c) but instead displaying the insertion rate fold changes of screens with overexpressions compared to no overexpression (y-axis), calculated from the ratio of sums of all sequences (lines) or 10 randomly sampled sequences.

The *HEK3* locus in HEK293T contains a single nucleotide variation at position 9 after the prime editor nick site. The pegRNA homology arm encodes a G for this position, while one of the three chromosome copies encodes an A. If a 3' flap containing the edit and at least 9 nucleotides of the homology arm was fixed into the genome, I would expect a decreased frequency of the A allele. Indeed, for both pooled and validation screen conditions without *TREX1/2* overexpression, I only observed 0.95-1.6% (screen averages) of reads with library insertions containing A in the +9 position compared to 33-36% for unedited reads (Figure 2.15d). This is in contrast to screens overexpressing *TREX1/2* where the percentage of the A allele increased to 3.4-6.9%, suggesting a higher proportion of flaps where the homology arm was digested to below 9 nt (Figure 2.15d). Taken together, this data demonstrates that *TREX1/2* antagonize the insertion of longer sequences with prime editing, presumably by digesting the 3' flap intermediate containing the edit.

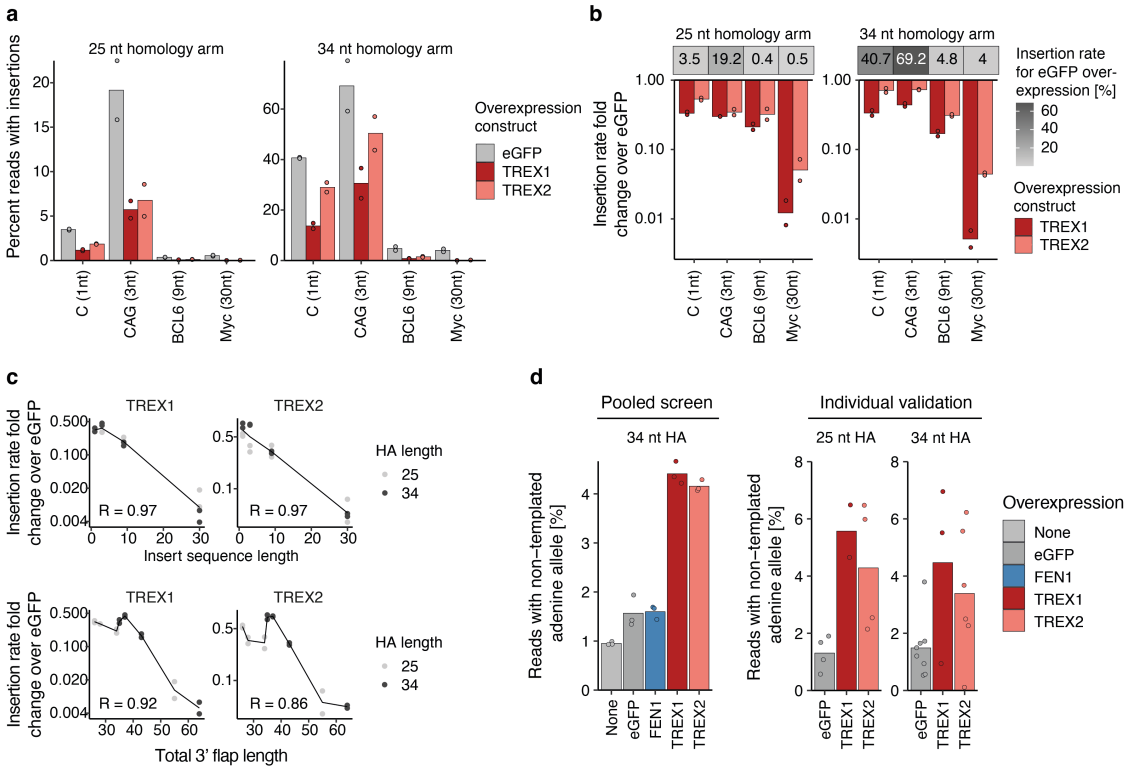

Figure 2.15. *TREX1* and *TREX2* degrade the 5'flap based on insertion length. a. Insertion frequencies (y-axis) of four sequences with varying insert lengths (x-axis) for two biological replicates (markers) while overexpressing *eGFP*, *TREX1*, or *TREX2* (colors), stratified by homology arm lengths (panels). Bar: average of markers. **b.** Top: Average insertion frequency (grayscale) of four sequences with varying lengths (x-axis) when overexpressing eGFP stratified by homology arm lengths (panels). Bottom: Insertion rate fold changes compared to eGFP (y-axis) when overexpressing *TREX1* and *TREX2* (colors). $n = 2$ biological replicates. **c.** Insertion rate fold changes (y-axis) over *eGFP* in cells overexpressing *TREX1* or *TREX2* (columns) for sequences stratified by insert sequence length (x-axis, top row) or by the length of the total 3'flap (x-axis, bottom row) from two biological replicates (markers). Bar: average of markers. **d.** Fraction of the non-templated adenine allele at the +9 position (y-axis) for cells with overexpression constructs (x-axis and colors) stratified by experiment and homology arm lengths (panels). Markers show screen averages from three biological replicates for the pooled screen or from separate pegRNAs for the individual validation experiment.

2.2.3 Sequence content effects on insertion efficiency

I next examined sequence content-dependent variation in insertion rate. To address this in a length-independent way, I calculated the insertion rate of each insert relative to sequences with the same or similar length by dividing the insertion rate by the median insertion rate for sequences of the same length (for sequences < 10 nt) or the median across length bins for the sparser longer sequences. I then measured the correlation of length-adjusted insertion rates with sequence features, computed from the perspective of the written sequence (i.e. the reverse complement of the pegRNA molecule sequence). I observed a consistent cytosine preference across all four target sites and cell lines (Figure 2.16 a,b), with each extra percent cytosine in the insert increasing the relative insertion rate by an average of 2.2%. Conversely, the percent of adenine and thymine decreased insertion rates for all loci and cell lines (Figure 2.16a,b).

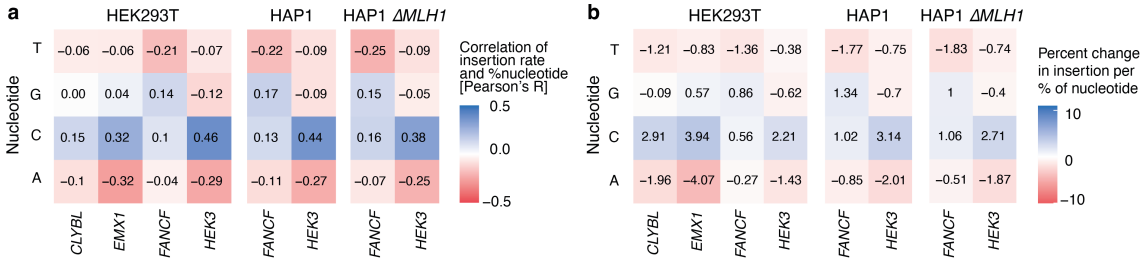

Figure 2.16 The impact of nucleotide composition on insertion efficiencies. a. Correlation of length-normalized insertion rate with nucleotide frequency in the insert (colors) for each nucleotide (y-axis) in each screen (x-axis). Data represent the average of $n = 3$ (HEK293T) or $n = 2$ (HAP1) biological replicates. **b.** As (a) but instead showing percent change in length-normalized insertion rate per % increase in insert nucleotide content.

The observations of nucleotide content effect were limited to four target sites, and moderately variable. To confirm whether the sequence influences hold more broadly, an additional set of screens was performed in HEK293T cells, targeting the original *HEK3* site and five novel sites within 1 kb of the *HEK3* site (dubbed *HEK3*-S2 to *HEK3*-S6) with pegRNA libraries encoding 356-388 18 nt inserts on pegRNAs with 15 nt homology arms (average insertion rate 3.2%, median R between replicates 0.81, Figure 2.17). Reassuringly, the sequence preferences were recapitulated in this experiment, with a strong preference for cytosines (average R between insertion rate and cytosine fraction = 0.47, Figure 2.18a,b).

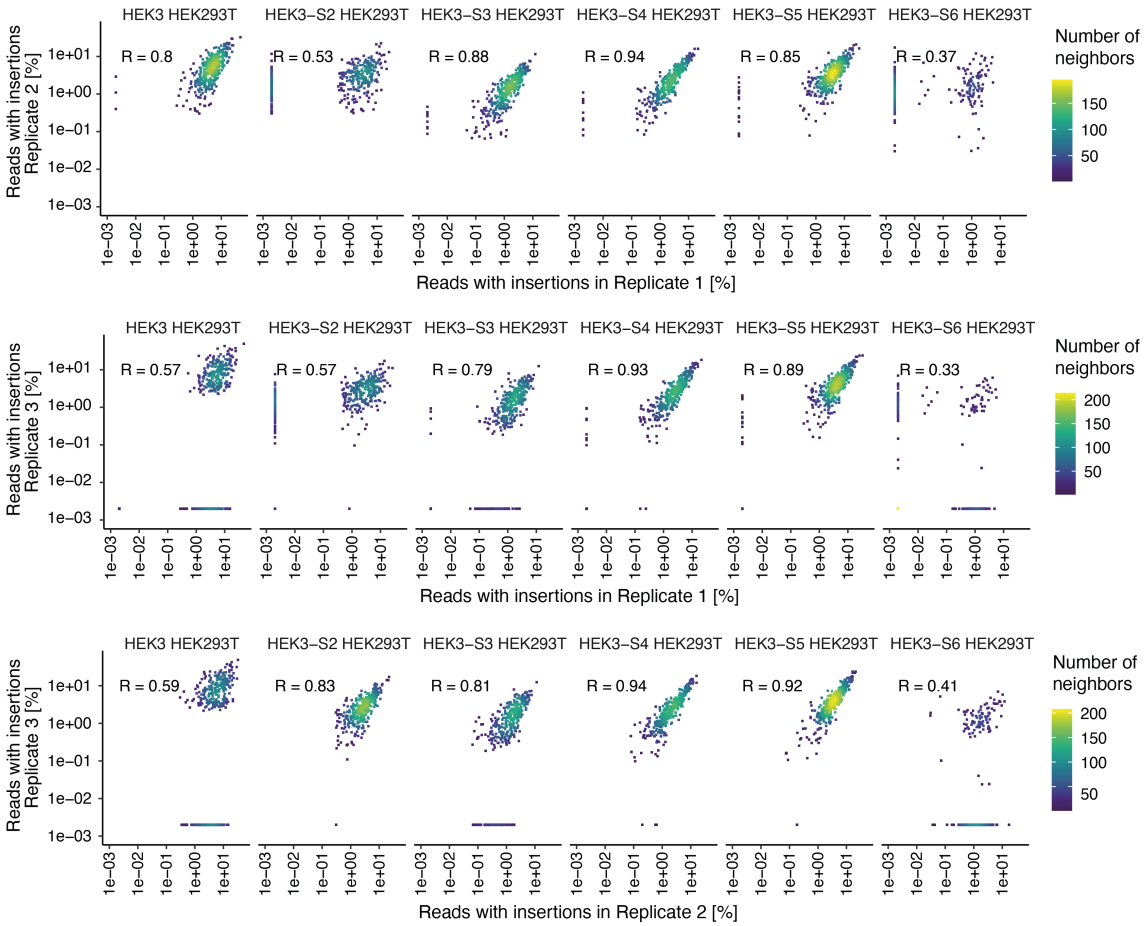

Figure 2.17. Reproducibility across prime editing screens inserting 18 nt sequences in novel target sites. Percent insertion in one replicate (x-axis) compared to percent insertion in another replicate for the new set of screens with 18 nt insert sequences (markers) at different target sites (rows) within 1 kb of the *HEK3* site in HEK293T cells. c. As (a) but for a new set of screens with 18 nt inserts and 15 nt homology arms targeting five novel sites within 1 kb of the *HEK3* site.

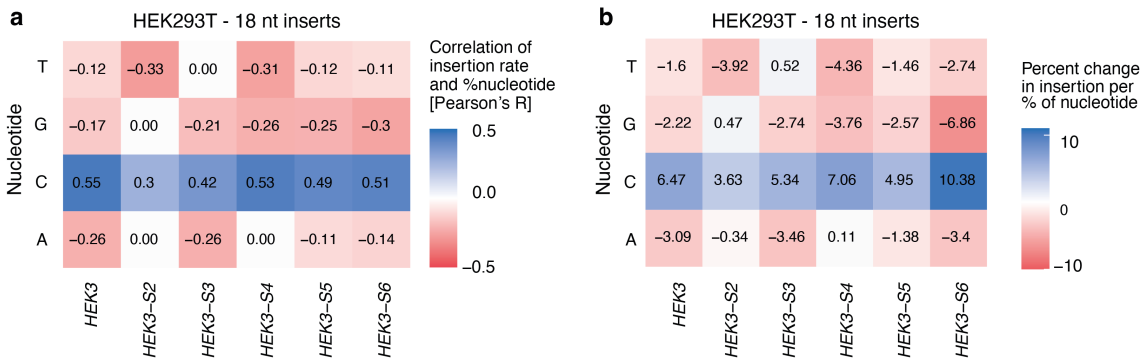

Figure 2.18 The preference for cytosines in the insert is consistent across five new target sites. a. Correlation of length-normalized insertion rate with nucleotide frequency in the insert (colors) for each nucleotide (y-axis) for a set of screens with 18 nt inserts and 15 nt homology arms targeting *HEK3* and five novel sites within 1 kb of the *HEK3* site (x-axis). Data represent the average of $n = 3$ (HEK293T) or $n = 2$ (HAP1) biological replicates. **b.** As (a) but instead showing percent change in length-normalized insertion rate per % increase in insert nucleotide content.

I next sought to understand how pegRNA secondary structure affects insertion rates. As the strength of the structure depends on the length of the insert, the secondary structure's free energy was calculated relative to a large sample of sequences of the same length (Methods). I observed that sequences with relatively stronger structures were more efficiently inserted (R=0.46, Figure 2.19a). To better understand this effect, I considered which combination of the pegRNA parts (primer binding site, insert, and homology arm) gives predicted free energies that best reflect insertion efficacy. I observed the strongest correlation when the structure was calculated from the reverse transcribed portion of the extension (i.e. the combination of insert sequence and homology arm; average R across screens = 0.38), and the additional inclusion of the primer binding site sequence decreased correlation (Figure 2.19b,c). Further, the free energies of pegRNA extensions designed for one target site always predicted insertion efficiency better at the same site than at other target sites (Figure 2.19d). Since the homology arm is specific to the target, this also explains some of the differences in insertion rates I observed across the target sites.

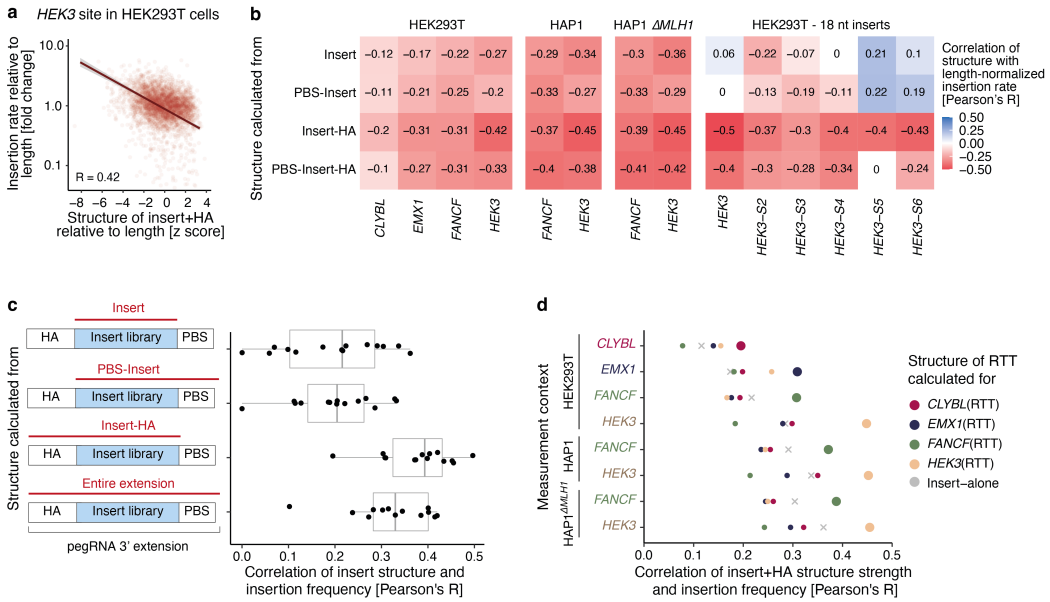

Figure 2.19 Structure in the reverse transcribed portion of the pegRNA improves the editing rate. a. Insertion rates at the *HEK3* site in HEK293T cells relative to length bin median (y-axis) for inserts (markers) with calculated Gibbs free energy (ΔG) from ViennaFold (x-axis). Line: linear regression fit; shaded area: 95% posterior confidence interval of the fit. Data represent the average of $n = 3$ biological replicates. **b.** Correlation of length normalized insertion rate (colors) with structure calculated from different parts of the extension (y-axis) in each screen (x-axis, grouped by cell line and screen sets). Data represent the average of $n = 3$ (HEK293T) or $n = 2$ (HAP1) biological replicates. **c.** Correlation (x-axis) between insertion rates and insert sequence free energy calculated from different parts of the 3' extension (y-axis). Box: median and quartiles; whiskers: least extreme of 1.5 times the interquartile range from the quartile and most extreme values. $n = 3$ (HEK293T) or $n = 2$ (HAP1) biological replicates. **d.** Correlation (x-axis) between insertion efficiency in different contexts (y-axis) and pegRNA 3' extension structure free energy calculated for pegRNAs against different target sites (colored markers), or the insert sequence alone (gray cross).

Structure in the insert and homology arm could increase prime editing efficiency by protecting the pegRNA itself from nuclease degradation, a strategy explored in engineered pegRNAs (epegRNAs) which contain structured RNA elements to the 3' of the primer binding site (Nelson et al. 2022; Xiangyang Li et al. 2022; G. Zhang et al. 2022). If this was the case, I would expect structured pegRNAs to be more abundant in the transcriptome, which was not the case (Figure 2.20a,b), suggesting an alternative mechanism. Engineered pegRNAs increase editing rates by stabilizing the 3'-extension with a structured cap. To understand if structure in the insert acts independently from the structured cap, I screened 439 inserts of varying free energy from the original pegRNA library in the epegRNA construct, targeting the *HEK3* site in HEK293T cells (Figure 2.20c-e). I found that the additional structure in the insert and homology arm also increased insertion rates for epegRNAs (R=-0.34) but to a lesser extent than for regular pegRNAs (R=-0.53, Figure 2.20f), and that the insertion rates between regular and epegRNAs were highly correlated (R=0.79, Figure 2.20g). Together, this implies that structure past the protective cap still influences insertion rates independently of transcript abundance and that the results on insertion efficiencies are relevant for epegRNAs as well.

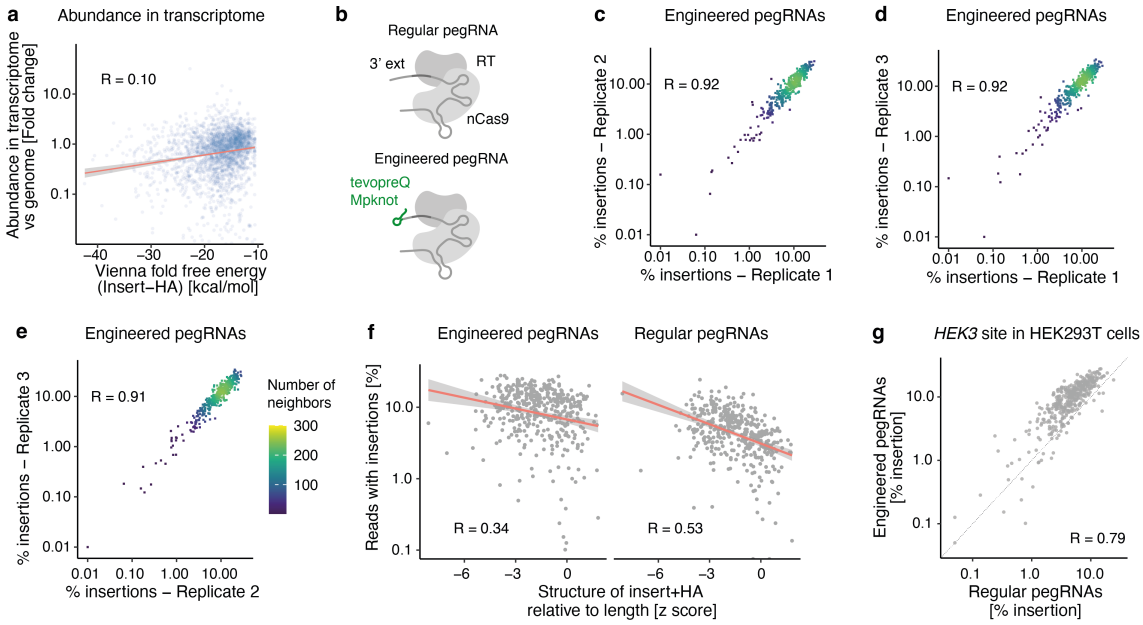

Figure 2.20 Structure past the protective cap influences insertion rates independently of transcript abundance. a. Fold-change at the *HEK3* site in HEK293T cells between read counts in the transcriptome and the genome for inserts (markers) with calculated Gibbs free energy (ΔG) from ViennaFold (x-axis). Line: linear regression fit; shaded area: 95% posterior confidence interval of the fit. Data represent the average of $n = 3$ biological replicates. b. Schematic comparing regular pegRNAs to engineered pegRNAs. **c.** Percent insertion in replicate 1 (x-axis) compared to percent insertion in replicate 2 for engineered pegRNAs encoding 379 structured

inserts (markers) in the HEK293T cell line. **d-e.** As in (c) but for different replicate comparisons. **f.** Insertion rates at the *HEK3* site in HEK293T cells relative to length bin median (y-axis) for 379 structured inserts (markers) with calculated Gibbs free energy (ΔG) from ViennaFold (x-axis) stratified by engineered and regular pegRNAs. Line: linear regression fit; shaded area: 95% posterior confidence interval of the fit. Data represent the average of $n = 3$ biological replicates. **g.** Insertion rates for sequences (markers) at the *HEK3* site in HEK293T for pegRNAs (x-axis) and engineered pegRNAs (y-axis). Data represent the average of $n = 3$ biological replicates.

I further noticed that structure in the reverse transcribed portion of the pegRNA was not correlated to the insertion rates of sequences < 5 nt, but was well correlated for longer sequences (Figure 2.21a-e). Since insertion rates of longer sequences are more impacted by overexpression of *TREX1* and *TREX2*, I speculated that the structure protects the reverse transcribed 3' DNA flap containing the edit from degradation (Figure 2.21f). Indeed, I observed that structure has a 2.4-2.6 fold stronger effect for cells overexpressing *TREX1* or *TREX2* compared to cells overexpressing *FEN1*, *eGFP*, or nothing (Figure 2.21a-e).

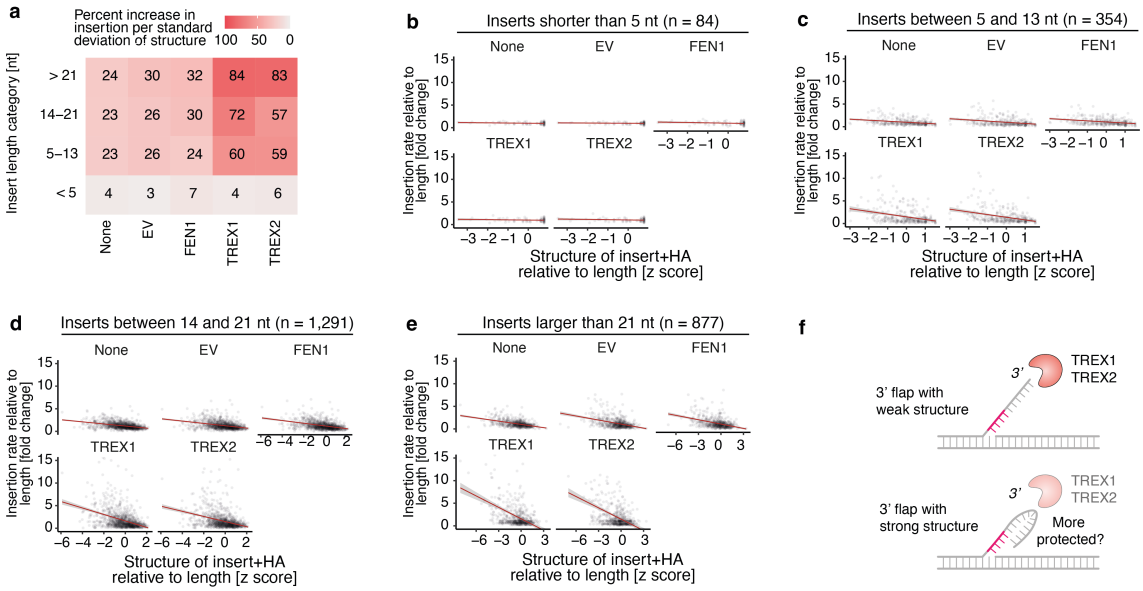

Figure 2.21 Structure is more protective for longer sequences and in the presence of *TREX1* and *TREX2*. a. Percent increase in insertion rate with each standard deviation increase in structure strength (colors) for different overexpression constructs (x-axis) and insertion sequence lengths (y-axis). **b.** Correlation of insertion rates at the *HEK3* site in HEK293T cells relative to length bin median (y-axis) for 84 inserts < 5 nt (markers) with z scores of calculated Gibbs free energy (ΔG) from ViennaFold (x-axis) relative to a large sample of random sequences of the same length. Stratified by overexpression constructs (panels). **c-e.** As (b) but for sequences of different lengths. **f.** Schematic of a model whereby structured inserts could protect from *TREX1/2* digestion.

Structure plays a role in other parts of the pegRNA molecule as well. For instance, the 13 nucleotides of the primer binding site are perfectly complementary to the protospacer (positions 5-17) and can therefore hybridize with each other. If the first nucleotides of the insert create further base pairing with the protospacer and scaffold, the strength of this structure is enhanced, and the protospacer could be sequestered from base pairing with the target site or ribonucleoprotein complex formation with Cas9 could be impaired (Figure 2.22a). To test if this additional pairing affects insertion rates, I predicted minimum free energy configurations of the primer binding site and the first three insert nucleotides with the spacer and the first guanine of the scaffold and observed 27% lower editing rates for inserts with extended base-pairing three nucleotides into the protospacer compared to no extension (Figure 2.22a). Finally, I tested if the disruption of the structural scaffold loops, which are required for association with Cas9, by the insert sequence reduces insertion rates. I calculated the minimum free energy configuration of the insert with the scaffold and observed 26% lower average editing for the pegRNAs with the first scaffold loop disrupted (screen range 10-43%) and 20% with the second and third loops (screen range 11-35%) compared to other inserts of the same length (Figure 2.22b). This loop dependence agrees with recent findings that scaffold variants with additional point mutations to stabilize the stem-loops can increase prime editing efficiencies (Xiaosa Li et al. 2022).

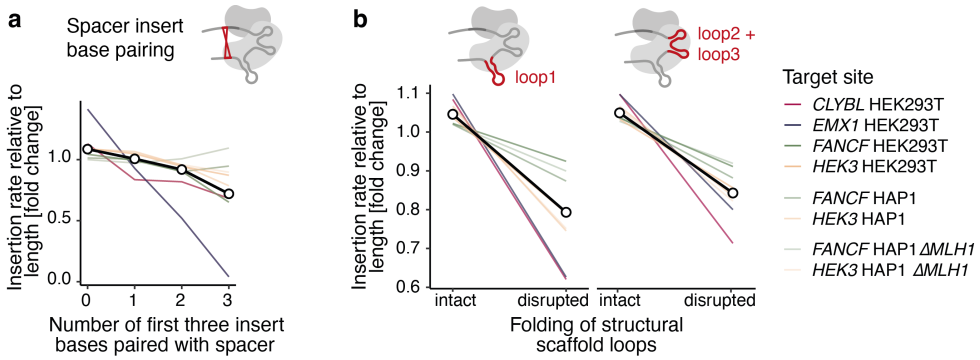

Figure 2.22 Extensive base pairing of insert sequences with the protospacer of scaffold loops inhibits prime editing. a. Insertion rates relative to length bin median (y-axis) for sequences with a different number of bases of the protospacer pairing to the first three nucleotides of the insert (x-axis). Colored lines show screen medians and the thicker black lines and dots show the median across all screens. **b.** As (a) but comparing sequences that disrupt or preserve (x-axis) scaffold loops (panels).

Combining effects of insert sequence length, cytosine content, and structure explained why some sequences are inserted much better than others. For example, the long 66 nt ELMI003108 sequence that was inserted in the *HEK3* locus at 1.39% insertion frequency (0.66% on average for the other 10 sequences > 66 nt) formed a strong structure together with the *HEK3* homology

arm (minimum free energy = -35.2 kcal/mol; 1.5 standard deviations lower than the average free energy of 66 nt sequences, Figure 2.23a). Other longer sequences that were inserted frequently relative to their size were recombinase sites that are often near-palindromic and therefore form strong structures (Figure 2.23b,c) which will be beneficial for larger genome engineering efforts discussed in chapters 3 and 4. Finally, the library included eight codon variations of the His6-tag in forward and reverse orientations. The average insertion difference between the best codon variant and the worst was 13.3-fold, with the highest insertion rate for the cytosine-richest CAC histidine codons (Figure 2.23d). This directly demonstrates the practical utility of this new understanding for guiding the codon choice for tags to insert.

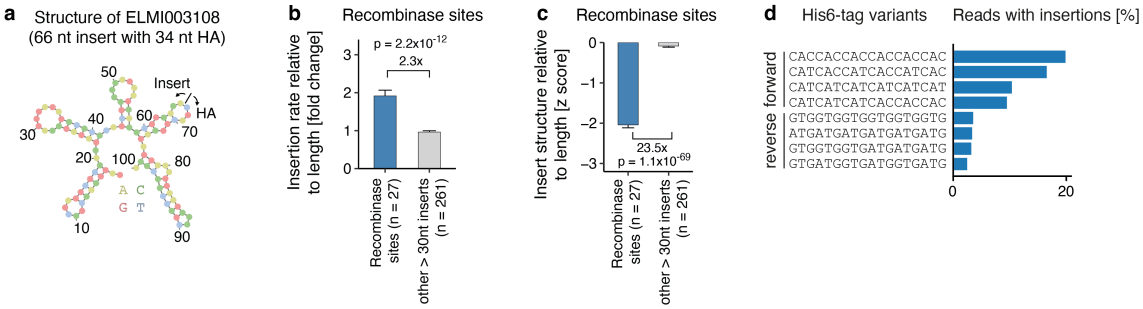

Figure 2.23 Sequence length, composition, and structure explain why some sequences are inserted better than others. The predicted secondary structure of a 66 nt insert sequence (ELMI003108) with the *HEK3* homology arm. **b.** Recombinase sites are efficiently inserted relative to their size. Average insertion rate relative to the length bin (y-axis) for recombinase sites or other insert sequences larger than 30 nt (x-axis). Bars: Median and standard error of median Comparison: ratio of blue to grey bar height. p-value: 1.1×10^{-69}; two-sided Student's t-test. $n = 3$ biological replicates. **c.** As (b), but for the secondary structure of the inserts. **d.** The average percent of reads with insertions (x-axis) for different codon versions of the His6 tag in forward and reverse orientation (y-axis) at the *HEK3* site in HEK293T cells.

2.2.4 Predicting insertion rates

The careful dissection and annotation of pegRNAs with sequence features presented an opportunity to predict the relative efficiencies of inserting different sequences into the same site. Juliane Weller, a PhD student in our lab, took on this challenge and built a computational model that could predict insertion rates with a correlation of 0.68 (Pearson's R) on held-out data. We called the model MinsePIE (Modeling insertion efficiency for Prime Insertion Experiments) and incorporated it into a package available at https://github.com/julianeweller/MinsePIE and produced a web application to predict prime editing insertion rates at https://elixir.ut.ee/minsepie/. Detailed descriptions of the model architecture, training, and validation, as well as use cases are published in (Koeppel et al. 2023).

2.3 Discussion

In this chapter, I presented a comprehensive analysis of prime editing insertion efficiencies using 3,604 pegRNAs and diverse follow-up experiments. I found determinants of insertion efficiencies that include target sites, repair contexts, prime editor systems, cytosine content, and the tendency of insert sequences to form secondary structures. I confirmed that active mismatch repair antagonizes the insertion of shorter sequences and discovered that the overexpression of the 3' flap nucleases *TREX1* and *TREX2* inhibited the insertion of longer sequences.

I uncovered a complex relationship between insertion sequence features and efficiency that is shaped by DNA processing and repair mechanisms. For the shortest sequences of up to 10 nt, it is increasingly appreciated that MMR proficiency is a strong factor (P. J. Chen et al. 2021; Ferreira da Silva et al. 2022), and I directly and comprehensively reaffirm this connection here. Surprisingly, sequences between 15 and 21 nt could insert at higher rates than shorter ones in MMR-proficient cells, and elongating the insertion can improve its insertion efficacy. This effect is likely due to a combination of antagonization by MMR for the shortest sequences, and potential steric issues for the 10-14 nt ones.

Sequences longer than 30 nt are incorporated less frequently. This could partly be explained by the discovery that the 3' flap nucleases *TREX1* and *TREX2* antagonize prime editing in an insert sequence length-dependent way. One explanation, supported by my observation that more structured long sequences insert at higher frequencies due to factors beyond RNA stability is that DNA flaps with longer insertions and less structure likely spend more time in a non-hybridized state and expose more single-stranded DNA even when hybridized, thus making them more vulnerable to nuclease degradation. This demonstrates that flap nucleases modulate prime editing, which motivates strategies for the next generation of long sequence insertions.

I further discovered that a stronger secondary structure of the reverse transcribed portion of the pegRNA led to higher insertion efficiency. This effect was evident when comparing different inserts into the same target, but also explained variable rates when attempting to write the same sequence into different target sites. I also observed strong correlations between structure and insertion rates in the context of epegRNAs, and correlation was highest when the structure was confined to the insert and the homology arm, indicating that the effects of structures in these two regions are separate. Therefore, I hypothesize that while the epegRNA structure improves editing rates by preventing degradation of the RNA 3'-extension, the structure in the transcribed template does so by preventing degradation of the ssDNA flap intermediate by flap nucleases. Indeed, flap nucleases had a smaller impact on insertions that resulted in more structured flaps. Alternatively,

structured inserts could ease the pairing of the edited strand with the non-edited strand due to being sterically smaller via folding onto themselves.

The improved understanding of insertion efficiency using the prime editing system naturally leads to recommendations for experimental design. First, I suggest choosing sequences with high cytosine content that are prone to form secondary structures. Inserts with runs of adenines should be avoided when using the U6 promoters for pegRNAs. For sequences shorter than 14 nt, transiently inhibiting mismatch repair (as implemented in PE4 or PE5 systems) (P. J. Chen et al. 2021), or knocking out *MLH1* will drastically improve insertion rates in MMR-proficient cells. If mismatch repair inhibition is undesired, padding the sequences to 18 nt or installing additional silent mutations on the reverse transcriptase template can increase insertion rates.

I measured the insertion efficiencies for thousands of sequences into 9 target sites in three cell lines across four prime editor systems. Nevertheless, there are several factors that I did not assess. First, I focused on precise on-target insertions but did not assay genome-wide off-target editing and only briefly touched on rare, undesired on-target mutations. Second, to comprehensively understand the determinants of prime editing efficiency requires the systematic assessment of several variables. Among the ones that are shaping up to be most important are (1) edit type, (2) target site, (3) PBS and RTT optimization, (4) repair context, (5) chromatin context, and (6) prime editor optimizations. Edit types and repair context were the most important dimensions for the complex genome manipulations I describe in chapters 3 and 4, and were among the least understood aspects of prime editing when I started to work on this chapter. Therefore, I deliberately focussed on insertions and repair but did not comprehensively characterize the other dimensions. Other efforts have inserted a smaller number of edits into a large number of synthetic target sites while exhaustively screening PBS and RTT variation (H. K. Kim et al. 2021; G. Yu et al. 2023; Mathis, Allam, Kissling, et al. 2023) and more recent work has focused on chromatin context (Xiaoyi Li et al. 2023) as well as prime editor optimization (Doman et al. 2023) and repair context (P. J. Chen et al. 2021). An ideal model would combine all these elements to predict optimal prime editing reagents for any given experiment. Such a model would require generating very large data sets. A preprint combining chromatin context, target sites, and edit types that came out at the time of writing this thesis is a first step in this direction (Mathis, Allam, Tálas, et al. 2023).

The prime editing field is moving rapidly (Scholefield and Harrison 2021; P. J. Chen and Liu 2022). Diverse applications are already emerging (Erwood et al. 2022), and some of the most exciting ones are specifically built around the insertion of short sequences. Examples include insertion of recombinase sites using prime editing to enable directed insertion of large DNA cargo

of up to 36 kb(Anzalone et al. 2022; Yarnall et al. 2022), creating long deletions and insertions using paired pegRNAs (Jinlin Wang et al. 2022; Anzalone et al. 2022; T. Jiang et al. 2022; Choi, Chen, Suiter, et al. 2022; Kweon et al. 2022), as well as clever utilization of short sequence insertion to generate a molecular recorder for sequential cellular events (Choi, Chen, Minkina, et al. 2022; Loveless et al. 2021; W. Chen et al., 2021). A better understanding of how cellular determinants and pegRNA features affect prime editing rates provides a foundation for these advances. The work presented in this chapter adds the important dimension of short sequence insertion in different DNA repair contexts, which holds promise in enabling both sophisticated genome engineering and the correction of thousands of pathogenic mutations.

2.4 Methods

Mammalian cell culture

I purchased the human HEK293T cell line from AMS Biotechnology (EP-CL-0005) and the HAP1 Δ*MLH1* cell line from Horizon discovery (HZGHC000343c022). The HAP1 WT cell line was provided by Andrew Waters (Wellcome Sanger Institute). HEK293T cells were cultured in DMEM (Invitrogen) and HAP1 cells in IMDM (Invitrogen), both supplemented with 10% FCS (Invitrogen), 2 mM glutamine (Invitrogen), 100 U/ml penicillin and 100 mg/ml streptomycin (Invitrogen) at 37 ℃ and 5% CO2. For cryopreservation, pellets of 1-10 million cells were resuspended in 1 ml of full media supplemented with 10% DMSO and frozen at -80˚C in MrFrosties (Nalgene) or CoolCell (Corning) freezing containers before transfer to liquid nitrogen for long-term storage.

Plasmid cloning

Important plasmids and benchling links to annotated sequences are listed in Table 2.1. pCMV-PE2-P2A-PuroR was cloned by replacing eGFP from pCMV-PE2-P2A-GFP (Addgene 132776) with the puromycin resistance gene. To do so, a gene fragment containing parts of the MMLV reverse transcriptase and the puromycin resistance gene was ordered from Integrated DNA Technologies. The gene fragment and pCMV-PE2-P2A-GFP were digested using AgeI, purified with the Monarch PCR & DNA Cleanup Kit (New England BioLabs), and ligated with T4 DNA ligase (New England BioLabs). The ligation product was transformed into XL10-Gold Ultracompetent Cells (Agilent) and plasmid DNA isolated using the Plasmid Plus Midi Kit (Qiagen).

pCMV-PE-FeLV-P2A-EGFP was generated by replacing the *MMLV* coding sequence between the XTEN linker and the 2A cleavage peptide with a synthesized gene fragment from Integrated DNA Technologies using Gibson Assembly. The gene fragment contained a human codon-

optimized version (codon optimization by Integrated DNA Technologies) of the MashUp reverse transcriptase (pipettejockey.com) that was engineered from the Feline Leukaemia Virus (UniProt Q85521). Fabio Liberante conceived the idea and helped with cloning.

pLentiGuide-BlastR was generated by replacing the puromycin resistance gene from Lenti_gRNA-Puro (Addgene 84752) with a blasticidin resistance gene. A gene fragment containing parts of the EF1a promoter and the blasticidin resistance gene was ordered from Twist Biosciences. The gene fragment and Lenti_gRNA-Puro were digested using FseI (New England BioLabs) and MluI-HF (New England BioLabs), purified with the Monarch PCR & DNA Cleanup Kit (New England BioLabs), and ligated with T4 DNA ligase (New England BioLabs) and transformed into XL10-Gold Ultracompetent Cells (Agilent). Plasmid DNA was isolated using the Qiagen Spin Miniprep Kit.

pPB-TREG3G-PE2-rtTA3G-P2A-eGFP was generated by fusing three gene fragments with restriction cloning. The first part contains the ITR sequences for the PiggyBac transposase, the second part contains prime editor 2 under the control of the third-generation doxycycline-inducible rtTA3G promoter, and the third part was synthesized by Twist Biosciences and contains a PGK promoter followed by the rtTA3G protein, a P2A sequence and eGFP.

pTwist_FEN1-T2A-tagBFP, TREX1-T2A-mScarlet, TREX2-T2A-emiRFP670, and Acceptor-T2A-eGFP were ordered from Twist Biosciences in a pTwist EF1 Alpha cloning vector. The protein sequences encoded by the primary transcripts of *FEN1, TREX1,* and *TREX2* were identified on ensembl.org (July 2022), fused with the T2A sequence, and the respective fluorophores and reverse translated into codon-optimized nucleotide sequences (Twist Biosciences).

Table 2.1. Important plasmids used in this chapter

Name	Description	Benchling link
pCMV-PE2-P2A-PuroR	Prime editor expression plasmid with puromycin resistance	https://benchling.com/s/seq-JxYVGybwwOovqgONpITH?m=slm-7R0qOn9t8xIHTPi7UZno
pLentiGuide-BlastR	Lentiviral acceptor vector for pegRNAs with blasticidin resistance	https://benchling.com/s/seq-of3MsHcYymrO04VXMqN5?m=slm-6njEl8yUYq48oeEWe8nG
pLentiGuide-BlastR-Library	Example of a library vector containing the loxP site and targeting the *FANCF* locus	https://benchling.com/s/seq-FjvxjpC95r4xbyJUBhQd?m=slm-iW7NNuOXt9FzJyXv5YBb

pPB-TREG3G-PE2-rtTA3G-P2A-eGFP	Piggybac vector with doxycyclin-inducible prime editor	https://benchling.com/s/seq-rCcJG0pk2TUvOSVljikl?m=slm-2LxVK7M5LvfREDcBRfgX
pLenti-PEG-HEK3-loxP	Lentiviral guide RNA vector to insert a loxP sequence into the *HEK3* locus	https://benchling.com/s/seq-hcqrsiKZ655IuGHrVd8Q?m=slm-hGVxYWhbjf6QFKFLYWaO
pTwist_EF1a_FEN1-T2A-tagBFP	FEN1 and tBFP overexpression vector	https://benchling.com/s/seq-P9kog1NtZ4NGlP84RcPL?m=slm-uxtWyuiTq9hk2EKYig0o
pTwist_EF1a_TREX1-T2A-mScarlet	TREX1 and mScarlet overexpression vector	https://benchling.com/s/seq-bDzcTrQqGtapgEJDOxLy?m=slm-VuDGiBXtGTrXBejWriiA
pTwist_EF1a_TREX2-T2A-emiRFP670	TREX2 and emiRFP670 overexpression vector	https://benchling.com/s/seq-fE0LXpErRwfbgEGF5frx?m=slm-h6ggd0mB9n7BCDbjVYMK
pTwist_EF1a_Acceptor-T2A-eGFP	Overexpression vector with a cloning site and eGFP	https://benchling.com/s/seq-C6ZKV8n8oCzml5oufiuw?m=slm-Mt6rXWRjdQZvWbdVrmgw

Generating HAP1 cell lines that stably express prime editor

HAP1 cell lines expressing prime editors were generated by co-transfecting pCMV-hyPBase (Yusa et al. 2011) and pPB-TREG3G-PE2-rtTA3G-P2A-eGFP. 500,000 HAP1 WT and 500,000 HAP1 Δ*MLH1* cells were each seeded into one well of a six-well plate one day before transfection. For each transfection, 3 µg of each plasmid were mixed with 6 µl of Plus reagent and 7.5 µl of Lipofectamine LTX (Invitrogen) reagent, incubated for 30 minutes, and then added to the cells. Two weeks post-transfection, cells were sorted into single clones based on eGFP expression. Two different individual clones were used for each screen. To test prime editing on a single locus, I infected cells with a lentivirus containing a pegRNA targeting the *HEK3* site and encoding a loxP sequence (Table 2.1). We selected for guide integration with 2µg/ml puromycin and then induced prime editor expression with 1 µM doxycycline and determined editing rates by visualizing amplicons (P7/8 and P9, Table 2.2) on the bioanalyzer (Agilent). I had help in the creation of HAP1 and HAP1 Δ*MLH1* cell lines from Mélanie Gouley, a master student I supervised.

Table 2.2. Sequences of oligonucleotides used in this chapter

ID	Name	Sequence	Purpose
P1	CLYBL_S2_F	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGAAGACCCCAGTGATTCATGCCTC	NGS of *CLYBL* target site
P2	CLYBL_S4_F	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACGAAGACCCCAGTGATTCATGCCTC	NGS of *CLYBL* target site
P3	CLYBL_iPCR_R	GAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTGGCTTGACTAGGGCTGGATGAT	NGS of *CLYBL* target site
P4	1114_EMX1_S1_F	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACAGCTCAGCCTGAGTGTTGA	NGS of *EMX1* target site

P5	1115_EMX1_S7_F	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGAC GACACAGCTCAGCCTGAGTGTTGA	NGS of *EMX1* target site
P6	1116_EMX1_ R	GAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTC CGATCTCTCGTGGGTTTGTGGTTGC	NGS of *EMX1* target site
P7	912_HEK3_F_S0	ACACTCTTTCCCTACACGACGCTCTTCCGATCTATG TGGGCTGCCTAGAAAGG	NGS of *HEK3* target site
P8	913_HEK3_F_S8	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGG ACACAATGTGGGCTGCCTAGAAAGG	NGS of *HEK3* target site
P9	995_NGS_HEK3 _R	GAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTC CGATCTCCCAGCCAAACTTGTCAACC	NGS of *HEK3* target site
P10	1000_FANCF_F_S 3	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGA attgcagagaggcgtatca	NGS of *FANCF* target site
P11	1001_FANCF_F_S 6	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTA GAAattgcagagaggcgtatca	NGS of *FANCF* target site
P12	1002_FANCF_R	GAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTC CGATCTGGGGTCCCAGGTGCTGAC	NGS of *FANCF* target site
P13	962_Seqlib_S2_F	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGAT GGCTTTATATATCTTGTGGAAAGGACGAAACACC	NGS of pegRNA library
P14	963_Seqlib_S6_F	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTA GAATGGCTTTATATATCTTGTGGAAAGGACGAAACA CC	NGS of pegRNA library
P15	965_P7_Broad_R	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTC TACTATTCTTTCCCCTGCACTGT	NGS of pegRNA library
P16	Goose_F_Universal	TTAAGCAAGCAAGCGAGCACTC	Amplification of oligos from pool
P17	Goose_CLYBL_R	GCGTCAATTCAGGCAACGAAGA	Amplification of oligos from pool
P18	Goose_EMX1_R	GCTTCAAGCCCTAGTGTCCTCA	Amplification of oligos from pool
P19	Goose_FANCF_R	ACAACTCTCTGAATGCGCTTGC	Amplification of oligos from pool
P20	Goose_HEK3_R	CCCGAGGAAATGATAGGGCGAT	Amplification of oligos from pool
P21	PE1.0	AATGATACGGCGACCACCGAGATCTACACTCTTTCC CTACACGACGCTCTTCCGATC*T	Forward primer indexing PCR
P22	Rev primer	CAAGCAGAAGACGGCATACGAGATN10GAGATCGGT CTCGGCATTCCTGCTGAACCGCTCTTCCGATCT	Reverse primer for indexing PCR

DNA oligonucleotides were designed using Benchling (https://www.benchling.com/) and ordered from IDT or Sigma Aldrich with standard desalting for purification.

Library design

Set 1: The insert sequence libraries contained 2,666 unique sequences, made up of useful molecular biology sequences, the Eukaryotic Motif Library, and sequences with strong secondary structures. I designed four separate versions of this library with identical insert sequences to target the *CLYBL*, *EMX1*, *FANCF*, and *HEK3* sites. The pegRNAs contained a 13 nt PBS and a 34 nt homology arm on the RT template. The utility sequences were hand-picked for their usefulness in molecular biology. The Eukaryotic Motif Library instances library with the corresponding fasta file of the genes was downloaded from elm.eu.org/instances.html?q= (Puntervoll et al. 2003; Dinkel et al. 2016, 2014) on 2020/11/19 and filtered to only contain sequences from "*homo sapiens*" that are longer than 1 amino acid. The amino acid motifs were extracted from the fasta file based on the indicated start and end sites. Finally, the amino acid motifs were reverse translated into DNA sequence using the 'reversetranslate' R package (version 1.0.0) and using the most frequent codon from the homo sapiens codon table. For the secondary structure library,

100,000 random DNA sequences of 20 and 30 nt length were generated (RBioinf::randDNA function; version 1.48.0), and their secondary structure was calculated (see insert sequence structure section). The sequences were distributed into 10 bins based on the strength of their secondary structure and 20 sequences were randomly picked from each structure bin to be included in the library. Finally, 30 random perfect 20 and 30 nt RNA hairpins were generated and amended to the secondary structure library. The combined library of insert sequences is deposited as Supplementary Data 1 in (Koeppel et al. 2023). The insert sequences were then flanked with primer binding sites, random nucleotide stuffer sequences for shorter inserts, BsmBI sites, and target vector compatible overhangs, resulting in 11,166 sequences of 199 nt. The oligonucleotide library was ordered from Twist Biosciences.

Set 2: This set of insert sequences was focused on short sequences between 1 and 10 nt. It included all 1, 2, 3, and 4 nt sequences and 100 random sequences (RBioinf::randDNA function; version 1.48.0) respectively of 5 to 10 nt, and 61 sequences < 10 nt from Set1 for a total of 999 unique inserts (938 were recovered in screens). The libraries were endowed with target-site-specific adaptor sequences and ordered the same way as Set1.

18 nt insert sequence libraries: This set of sequences consisted of six sublibraries that were designed to target the *HEK3* site and five additional nearby sites (within 1 kb), dubbed *HEK3-2, HEK3-3, HEK3-4, HEK3-5, and HEK3-6*. The sublibraries shared 100 identical, randomly generated (RBioinf::randDNA function; version 1.48.0) 18 nt insert sequences and 256-288 sublibrary-specific 18 nt insert sequences that were picked based on their ability to form secondary structure in the RT template. In contrast to Set1 and Set2, I ordered oligos for this set of sequences that already included the spacer (20 nt), improved scaffold (86 nt, sequence: gtttaagagctatgctggaaacagcatagcaagtttaaataaggctagtccgttatcaacttgaaaaagtggcaccgagtcggtgc), PBS (13 nt), insert (18 nt), and HA (15 nt). The oligos were endowed with BsmBI sites, overhangs for cloning, and primer binding sites for amplification of the oligo pool. The oligonucleotide library was ordered from Twist Biosciences.

Library cloning

Set1 and Set2: First, a separate, site-specific backbone was cloned for each target site. A gene fragment was ordered containing the protospacer, guide RNA scaffold, parts of the reverse transcriptase template and primer binding site, a stuffer sequence flanked with BsmBI sites for insert library insertion, and the T7 terminator motif. 100 ng of the gene fragments were digested with BsaI-HFv2 (New England BioLabs) and purified with the Monarch PCR & DNA Cleanup Kit (New England BioLabs). The pLentiGuide-BlastR plasmid was digested with BsmBI-V2

(New England BioLabs) at 55°C for 8h followed by 20 min heat inactivation at 80°C and gel purification using the QIAEX II Gel Extraction Kit (Qiagen). The gene fragments were ligated into the backbone using T4 DNA ligase (New England BioLabs) and transformed into XL10-Gold Ultracompetent bacteria (Agilent). The plasmids were purified with a Qiagen Spin Miniprep Kit.

Second, pegRNA insert libraries were inserted into the site-specific backbones. The insert libraries were synthesized as oligonucleotide pools and amplified using KAPA HiFi HotStart ReadyMix (Roche). Libraries for individual target sites were amplified with separate primers (Table 2.2). The products were purified using the Monarch PCR & DNA Cleanup Kit, digested with BsmBI-v2 at 55°C for 4h, and heat-inactivated at 80°C for 20 min alongside 5 µg of site-specific plasmids. The digested oligos were purified using the Monarch PCR & DNA Cleanup Kit. The vectors were treated with quick CIP (New England BioLabs) for 15 minutes at 37°C and then purified using QIAquick PCR Purification Kit (Qiagen). Inserts were ligated into vectors using Golden Gate assembly. A 1:3 molar ratio of insert and vector was mixed with BsmBI-v2 and T4 DNA ligase and incubated in a thermocycler for 30 cycles, alternating between 5 minutes at 42°C and 5 min at 16°C and finishing with a heat inactivation step at 60°C for 5 min. The ligation products were purified with Monarch PCR & DNA Cleanup Kit and electroporated into MegaX DH10B T1R Electrocomp Cells (ThermoFisher). The bacteria were grown overnight in liquid culture and plasmid was extracted using the Plasmid Plus Midi Kit.

epegRNA libraries were cloned by first generating a *HEK3* site-specific epegRNA backbone with a stuffer sequence for the insert libraries (as above). The tevopreQ sequence was added to the fragment containing the protospacer, guide RNA scaffold, parts of the reverse transcriptase template and primer binding site, a stuffer sequence flanked with BsmBI sites for insert library insertion, and the T7 terminator motif by PCR (using P42, P43, Table 2.2). Next, the 379 sequences with strong structures were amplified from the Set1 oligopool by PCR and cloned into the epegRNA *HEK3* backbone as described above.

18 nt inserts and codon variation libraries: pLentiGuide-BlastR plasmid was digested with BsmBI-V2 (New England BioLabs) at 55°C for 8h followed by 20 min heat inactivation at 80°C and gel purification of the vector using the QIAEX II Gel Extraction Kit (Qiagen). Amplification, purification, digestion, and repurification, were performed as described above. The oligo sequences were ligated into pLentiGuide-BlastR using Golden Gate assembly, the ligation product was purified and transformed into bacteria, and the plasmid was extracted after an overnight culture as above. I had help with library cloning from Elin Madi Peets, a technical specialist in our laboratory, and co-first author of the manuscript.

Lentivirus production

Lentivirus was produced in HEK293FT cells that were transfected with Lipofectamine LTX (Invitrogen). 5.4 µg of a lentiviral vector, 5.4 µg of psPax2 (Addgene 12260), and 1.2 µg of pMD2.G (Addgene 12259) were mixed in 3 ml Opti-MEM together with 12 µl PLUS reagent and incubated for 5 min at room temperature. 36 µl of the LTX reagent was added and the mix was incubated for another 30 min at room temperature. 3 ml of the transfection mix was then added to 80% confluent cells in 10 ml DMEM media in a 10-cm dish. After 48h the supernatant was collected and stored at 4°C. Fresh media was added to the cells and harvested 24 hours later. The two harvests were kept separate. For virus titration, Lenti-X GoStix Plus (Takara) was used following the manufacturer's protocol. I had help with lentivirus production from Elin Madi Peets.

pegRNA insertion screens in HEK293T cells

Infection with pegRNA library. cells were infected with the pegRNA library (separate infections for each target site and library set) aiming at a multiplicity of infection of 0.5 and a guide coverage of > 1000x. Each screen was performed in 3 biological replicates and independently infected. To achieve this, $6x10^6$ cells were plated in three wells of a six-well plate and spin infected for 15-30 mins at 2000 rpm. Following infection, cells were resuspended and plated at $2x10^4$ cells/cm^2. Cells were cultured for 7 days and selected for pegRNA integration with 10 µg/ml blasticidin.

Transfection with prime editors. HEK293T cells were seeded at a concentration of $6.9x10^4$ cells/cm^2 in a 15-cm dish. The next day the media was replaced with fresh media and the cells were transfected using Lipofectamine LTX reagent. 72 µg PE-Puro or PE-FeLV plasmid were mixed with 8 µg pCS2-GFP and 40 µl Lipofectamine P3000 (Invitrogen) in 3.2 ml Opti-Mem (Gibco). In another tube, 40 µl of Lipofectamine 3000 and 160 µl Lipofectamine LTX were mixed in 3.2 ml Opti-Mem. The solutions were combined, incubated for 30 minutes at room temperature, and then added to the cells. For PE3, an additional 6 µg of nicking guide RNA was added. For screens with nuclease overexpression, an additional 30 µg of flap-nuclease or eGFP plasmid in the pTwist vectors was added. I had help with the pegRNA insertion screens from Elin Madi Peets.

pegRNA insertion screens in HAP1 and HAP1 *ΔMLH1* cells

Infection with pegRNA library. The pegRNA library viruses for all target sites and sets were individually quantified using the Lenti-X GoStix Plus (Takara) kit and then combined into one virus pool. The HAP1 Δ*MLH1* and HAP1 Δ*MLH1* cells with PiggyBac-integrated PE2 were infected with the virus pool aiming at a multiplicity of infection of 0.5 and a pegRNA coverage

of > 1000x. Each screen was performed in 2 biological replicates with separate PiggyBac prime editor clones and independently infected. To achieve this, $6x10^6$ cells were plated in three wells of a six-well plate and spin infected for 15-30 mins at 2000 rpm. Following infection, cells were resuspended and replated at $2x10^4$ cells/cm^2. Cells were cultured for 7 days and selected for pegRNA integration with 10 µg/ml blasticidin.

For each replicate, 30 million cells were seeded into 5-layer flasks and induced with 1 µM doxycycline (Selleckchem). The cells were split once at day 4 and the doxycycline was refreshed. Finally, cells were harvested on day 7 post-induction. I had help with the pegRNA insertion screens from Elin Madi Peets.

DNA extraction and library preparation for next-generation sequencing

Genomic DNA extraction and sequencing library preparation for screens were done as described in Allen et al., 2018 (Allen et al. 2018). Briefly, cell pellets were resuspended in TAIL BUFFER A (100 mM Tris-HCl, 5 mM EDTA, 200 mM NaCl) and then mixed with 1 volume of TAIL BUFFER B (100 mM Tris-HCl, 5 mM EDTA, 200 mM NaCl, 0.4% SDS) supplemented with freshly thawed Proteinase K (20 mg/ml final). The lysate was incubated overnight at 56˚C. On the next day, RNase A was added to a final concentration of 10 µg/ml and incubated at 37˚C for 30 min - 4 h. One volume of isopropanol was added and the DNA spooled on a sterile inoculation loop. The DNA was washed three times by dipping it into consecutive 5 ml tubes containing 70% ethanol. The DNA was air-dried for 5-10 mins and resuspended in TE buffer (pH 8.0).

For each screen, two independent amplicons were generated by PCR using Q5 Hot Start High-Fidelity 2X Master Mix (New England BioLabs). One amplicon for the targeted locus and one amplicon for the pegRNA locus (primers in Table 2.2). To maintain high coverage for each sample, 40 µg of gDNA was used as the template and each PCR reaction was run in 50 µl aliquots containing no more than 5 µg DNA. The PCR reactions were column-purified using the QIAquick PCR Purification Kit (Qiagen). Sequencing adaptors and barcodes were added with a second round of PCR using the KAPA HiFi HotStart ReadyMix (Roche), primers P3 and P4 (Table 2.2), and 1 ng of template DNA. Amplicons were purified with Agencourt AMPure XP beads in 0.7:1 ratio (beads to PCR reaction volume) and quantified with the Quant-iT™ High-Sensitivity dsDNA Assay Kit (Invitrogen). The amplicons were pooled together and sequenced on the Illumina HiSeq 2500 using HiSeq Rapid SBS Kit v2 (500 cycles, 250 paired-end). I had help with DNA extraction and library preparation from Elin Madi Peets.

Reverse transcription of pegRNA libraries.

Frozen cell pellets containing 4.5-6.1 million cells from screens targeting the *HEK3* site in HEK293T cells were washed with 500 µl PBS and the RNA was extracted using the mirVana™ miRNA Isolation Kit (Invitrogen). 8.4 µg-16.6 µg of template RNA split across 8 reactions were used for gDNA digestion and cDNA synthesis with the SuperScriptTM IV VILO™ Master Mix with ezDNase (Invitrogen). For cDNA synthesis, a primer was used that was the reverse complement to the 13 nt PBS with extra nucleotides on the 5' end to provide additional base-pairing for PCR amplification (ATCGAGTTTCAGACTGAGCACG, Table 2.2). pegRNAs were amplified from the cDNA mixture by 27 cycles of PCR using KAPA HiFi HotStart ReadyMix (Roche) and primers P39 and P40 (Table 2.2). Library preparation and sequencing were performed as described in the "DNA extraction and library preparation of next-generation sequencing" section (page 59). I had help with RNA extraction and library preparation from Elin Madi Peets.

Generating read count tables

Paired forward and reverse reads from Illumina sequencing were merged using PEAR v0.9.11. Data for different sequencing lanes of the same screen were concatenated. The resulting merged fastq files were processed as follows: First, DNA sequences were trimmed to contain the 10 nt up and downstream of the nick site (for target site amplicon) or to contain 15 nt up and downstream of the nick site (pegRNA amplicon). On average, 98% of reads were matched for the target site amplicon and 84% for the pegRNA amplicon. The trimmed sequences were then matched to each insert in the pegRNA library flanked by 10 nt of target site sequence (for target site amplicon) or flanked by 15 nt pegRNA plasmid sequence (pegRNA amplicon), requiring no mismatches. Adding the flanking sequences is to ensure that only insertions at the correct location are considered. On average 92% of reads were matched to the unedited locus or an insertion for both the target site amplicon and the pegRNA amplicon.

Combining replicates

pegRNAs where any replicate had fewer than 20 reads in the pegRNA amplicon mapping to it were filtered out. Insert counts were normalized to frequencies by dividing the reads for each insert by the number of reads in each screen. Insertion efficiencies were calculated for each replicate and screen by dividing the target insert frequency by the pegRNA insert frequency (Note: calculating insertion frequencies this way likely underestimates them, as it does not take cells that were not infected with the library into account. In addition, an average of 16% of reads in the pegRNA amplicons did not match to any sequence in the library). Finally, insertion efficiencies were averaged across replicates.

Mutation rates around the insertion site and indel detection

The fastq reads of the target sites were trimmed by matching a stretch of ten nucleotides directly upstream of the PBS and 60 nt downstream of the insertion site (*CLYBL*: CTGAATGGTG, CAGAGTTCCA; *EMX1*: GGGCCTGAGT, ATGGGGAGGA; *FANCF*: CCTCATGGAA, AGCACCTGGG; *HEK3*: CCTTGGGGCC, AGCTTTTCCT). The occurrence of library insertions was detected by pattern-matching the trimmed reads for library sequences. Indel detection: The trimmed reads were filtered in a series of steps. First, sequences with insertions at the nick site that perfectly match a sequence in the insert libraries were removed (this also means that the method cannot detect single/double/triple nucleotide insertions at the nick site because the library contains all possible singlets/doublets/triplets). Second, sequences that contained 'N' were removed. Third, sequences with a perfectly preserved sequence around the cut site were removed. Fourth, sequences that are 83 nt long were removed (83 nt corresponds to the length of a sequence without indels). The remaining sequences were annotated according to the indel type. Scaffold integrations were sequences that contained five or more nucleotides of the scaffold (GCACC) directly downstream of the RTT. Mutated insertions were sequences that matched any sequence > 10 nt in the library with no more than 3 mismatches (fuzzyjoin R package 'v0.1.6', optimal string alignment method). Duplications were sequences that contained two or more copies of the homology arm sequence. Deletions at the target sites were deletions that overlapped up to 10 nt up and/or downstream with the nick site. Other deletions were deletions that did not overlap with the nick site and all remaining sequences are classified as 'other'.

SNV detection: Going from the outside to the inside of the trimmed sequence (with the nicking site being between the two innermost nucleotides), the occurrence of the four nucleotides was counted at every position. Non-reference nucleotides were classified as mutations except a non-reference SNP (A) in HEK293T cells for 1 of 3 alleles at position +9. The RT template on the pegRNA corresponds to the sequence of the major allele (G).

Data analysis and feature generation

Merging data from Set1 and Set2: For each target site and cell line, the insertion rates in Set2 were multiplied by the ratio of the mean insertion rate of the shared sequences in Set1 and the mean insertion rate in Set2. For the 140 shared insert sequences, the mean insertion rate between both sets was calculated. Length-normalized insertion rates: Length residuals were calculated by dividing the insertion rate by the median insertion rate for sequences of the same length (for sequences < 10 nt) or by dividing sequences into length bins. The length bins consisted of sequences from 10-14, 15-19, 20-24, 25-29, 30-39, 40-49, 50-59, and 60-69 (sequences with lengths above 30 nt were divided into length bins of 10 nt because there were fewer longer

sequences in the library). Juliane Weller calculated the tendency of insert sequences (alone or in the context of PBS and/or HA) to form secondary structures using the RNA fold (version 2.4.16) algorithm of the ViennaRNA (2.5.0a) package (Gruber et al. 2008; Hofacker 2003). The free energy was normalized to the mean and standard deviation (z score) of 1000 random sequences with the same length and in the same context.

Comparison of HAP1 and HAP1 *ΔMLH1* lines

To account for screen batch effects for direct comparisons, the mean insertion rates across wild-type and *MLH1* knockout HAP1 cell lines were scaled to be identical for > 13 nt sequences that are not affected by MMR. The fold changes of the scaled insertion efficiencies between HAP1 Δ*MLH1* and HAP1 lines were then calculated for each sequence in the library.

Validation of nuclease overexpression with individual pegRNAs.

I chose four different insertions (C, CAG, a BCL6 recognition sequence: TTCTAGGAA, and a Myc-tag: GAGCAGAAGCTGATCAGCGAAGAGGACCTC) from the pooled library for validation and cloned them into *HEK3* site targeting pegRNAs endowed with 25 or 34 nt homology arms. One day before transfection, HEK293T cells were seeded in two 24-well plates at 50,000 cells per well. All transfections were done in replicates and each well was transfected with 500 ng of pCMV_PE2_P2A_PuroR, 150 ng pTwist nuclease or eGFP overexpression constructs, and 100 ng prime editing guide RNA using Lipofectamine LTX according to the manufacturer's protocol. Successful transfection one day later was confirmed by fluorescence microscopy and 2 µg/ml puromycin was added one day later. Cells were harvested five days post-transfection by direct lysis of cell pellets using home-made quick extract buffer (1 mM CaCl2, 3 mM MgCl2, 1 mM EDTA, 1% Triton X-100, 10 mM Tris pH 7.5) with freshly added proteinase K (0.2 mg/ml) followed by 15 min incubation at 65°C and 20 min incubation at 95°C. 1.5 µl of the lysate was directly added to 25 µl of amplicon PCRs. Sequencing adaptors and barcodes were added by a second round of PCR and the purified products were sequenced on Illumina Miseq (300 cycles). Correctly edited reads were identified by pattern matching for the insert sequence flanked by 10 nt of the target site to each end. Unedited sequences were detected by matching the 20 nt of wild-type sequences around the nick site. The insertion rate was calculated by dividing the number of edited reads by the number of wild-type reads.

Statistics and reproducibility

The n numbers denoted in the figure legends refer to independent experiments that were separately infected with the pegRNA library. Measurements were always taken from distinct samples. No statistical methods were used to predetermine the sample size. The experiments were not

randomized and the investigators were not blinded to allocation during experiments and outcome assessment. Wherever correlations were indicated, Pearson's R was used. T-tests were performed as two-sided tests. Normal distribution of the underlying data was assumed and no adjustments for multiple comparisons were made.

Software

BaseSpaceCLI (1.4.0); Geneius codon optimization webtool from Eurofins Genomics (accessed 2022); benchling (accessed between 2019-2023); PEAR (0.9.11); Python (3.8.10); Python packages: Biopython (1.79), more-itertools (8.5.0), pandarallel (1.6.1), scikit-learn (0.24.2), scipy (1.5.3), shap (0.39.0), statannot (0.2.3), XGBoost (1.4.0); R (4.0.2); ViennaRNA (2.5.0); R packages: Broom (0.7.9), fuzzyjoin (0.1.6), ggpointdensity (0.1.0), RBioinf (1.48.0), reversetranslate (1.0.0), ShortRead (1.46.0), spgs (1.0-3), Tidyverse (1.3.1), Viridis (0.6.1).

3

Randomizing the human genome by engineering recombination between repeat elements

Nick Lane writes in his 2015 book, What is Life: “If we had to sum up the architectural constraints on genomes in a single phrase, it would have to be ‘anything goes’”. But is this the case and how could we test ‘what goes’? One way would be by creating and studying the phenotypes for many versions of a genome that all differ in organization and content.

In this chapter, I lay out a strategy to create deletions, inversions, translocations, and extrachromosomal circular DNA at scale by highly multiplexed insertion of recombinase recognition sites into repetitive sequences. Combining learnings on short sequence insertions with prime editing and targeting of LINE-1 elements, I derived stable human cell lines with several thousand clonal insertions. Subsequent recombinase induction generated an average of more than one hundred megabase-sized rearrangements per cell, and thousands across the whole population. The ability to detect rearrangements as they are generated and to track their abundance over time allowed me to measure the selection pressures acting on different types of structural changes. I observed a consolidation towards shorter variants that preferentially delete growth-inhibiting genes and a depletion of translocations. I isolated and characterized 21 clones with multiple recombinase-induced rearrangements. These included viable haploid clones with deletions that span hundreds of kilobases and triploid HEK293T clones with aneuploidies and fold-back chromosomes. I mapped the genetic changes to RNA expression to decipher how structural variants affect gene regulation. The genome scrambling strategy developed here makes it possible to delete megabases of sequence, move sequences between and within chromosomes, and implant regulatory elements into new contexts which will shed light on the genome organization principles of humans and other species.

I had help from Gareth Girling in the LINE-1 editing time courses and the derivation of scrambled clones. I collaborated with Raphael Ferreira from the Church lab on this project, and only my results are presented in this chapter. The results will be preprinted in early 2024 and I will be the lead author on the manuscript.

3.1 Introduction

Only ~1% of our genomes are made up of protein-coding sequences, ~8% consists of non-coding sequences with biochemical marks correlating with regulatory activity (ENCODE Project Consortium et al. 2020), and the majority (54%) is repetitive (Hoyt et al. 2022). It remains unclear how much DNA in our genomes is dispensable for cellular survival and how the expression of genes changes when manipulating the order and position of nearby sequences.

The importance of genome structure, order, and content can be probed by comparing how alternative genome configurations behave in cells. Experimentally, site-specific recombinases can induce diverse DNA sequence alterations such as deletions, inversions, and translocations which have shed light on genomic organization principles (Tian and Zhou 2021; Olorunniji, Rosser, and Stark 2016; Zheng et al. 2000). However, the application of recombinases has been primarily confined to investigating individual loci and not the entire human genome. Alternatively, complex genome-shattering events can be observed in pathologies such as chromothripsis in cancer, and illuminate the space of viable genome configurations (Stephens et al. 2011; Forment, Kaidi, and Jackson 2012). These drastic alterations can lead to an unexpected increase in cellular adaptability including drug resistance (Shoshani et al. 2021). Despite the instructive nature of these natural genomic phenomena, our observations are confined to post-selection results and we remain without the means to initiate these rearrangements at will or monitor them in real-time.

A method to generate variants at the genome scale and in an experimentally controllable way has been pioneered in the baker's yeast, *S. cerevisiae*, through the incorporation of hundreds of symmetrical loxP sites (loxPsym) into synthetic chromosomes. Each of the sites could act as an anchor for recombination by Cre recombinase. Collectively, these sites form the substrate of an inducible genome evolution system called SCRaMbLE (synthetic chromosome rearrangement and modification by loxP-mediated evolution) (J. Dymond and Boeke 2012) (J. S. Dymond et al. 2011; Zhou et al. 2022). This approach is uniquely targeted, avoids double-strand breaks, and preserves exon integrity. A wealth of knowledge about genome rearrangements has been obtained through this system. For example, scrambling one synthetic chromosome and moving genes across varying expression neighborhoods affected isoform selection (Brooks et al. 2022) while scrambling six synthetic chromosomes linked rearrangement patterns with chromatin accessibility and 3D arrangements, connecting chromatin structure with genome evolutionary dynamics (Zhou et al. 2022). However, the human genome is vastly larger compared to the yeast genome, putting the synthesis of chromosomes for SCRaMbLE currently out of reach. In addition, the sparsity of genes in the human genome would likely result in a different landscape of tolerated variation.

Scrambling human cells would require inserting recombinase sites into the genome at scale. Prime editing is a novel genome-editing technique that makes it possible to precisely engineer small sequence changes without double-strand DNA breaks and external DNA donor templates (Anzalone et al. 2019; Anzalone, Koblan, and Liu 2020; P. J. Chen and Liu 2022). To achieve the necessary scale, recurring patterns of repetitive elements offer a unique opportunity to insert many recombinase sites simultaneously (Yarnall et al. 2022). Long-interspersed elements-1 (LINE-1) are a class of transposable elements that make up 17% of the human genome. Previous studies have targeted these abundant repetitive elements with base editors and demonstrated the feasibility of highly multiplexed genome editing (Smith et al. 2020; Zou et al. 2022).

Here, I combine the SCRaMbLE and prime editing technologies to derive stable cell lines with hundreds of loxPsym insertions into LINE-1 elements, despite the considerable toxicity associated with manipulating numerous loci. I characterize these prime-edited cell lines and explore the features of LINE-1 elements that make them amenable to editing. Scrambling cell lines with hundreds of integrated recombinase sites generates thousands of diverse structural variants. By comparing variants generated shortly after induction and those remaining in culture after two weeks, I map the selection pressures acting on these variants. In the process, I discover megabase scale deletions that can survive in haploid genomes and characterize changes in gene expression in cell clones with multiple Cre-induced rearrangements. These findings help elucidate the selective forces shaping structural changes of the genome and provide a strategy to manipulate mammalian genomes at previously unseen scales.

3.2 Results

The strategy described in this section consists of four steps (Figure 3.1). (1) Targeting prime editors to high copy number LINE-1 retrotransposons to insert hundreds of recombinase sites into the genomes of cell lines. (2) Treating cells with Cre recombinase to induce and sequence thousands of structural variants. (3) Keeping cells in culture for two weeks and sequencing the persisting variants to learn about the selection pressures acting on rearrangements. (4) Deriving single-cell clones with Cre-induced rearrangements and characterizing them deeply.

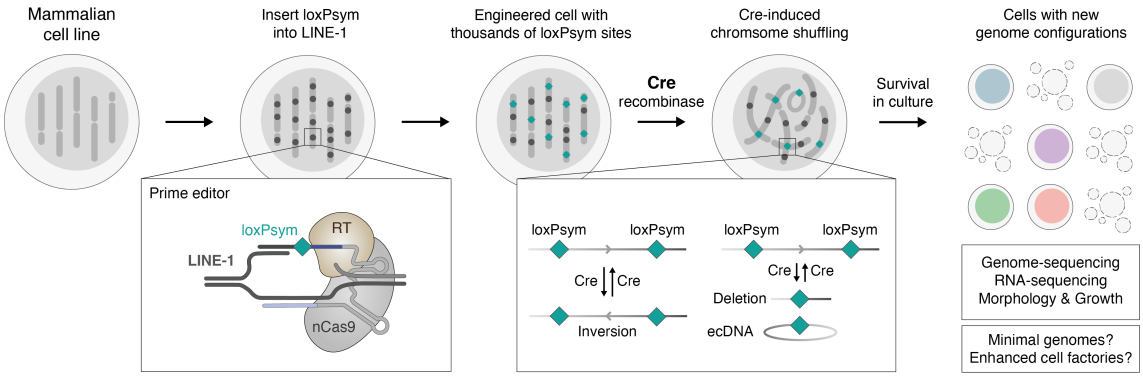

Figure 3.1 A strategy to scramble human genomes. Schematic of a strategy to scramble mammalian cells. Thousands of loxPsym sites are inserted into LINE-1 with prime editing. Induction of Cre recombinase shuffles the chromosomes. Surviving derivative clones are sequenced and phenotyped.

3.2.1 Stable cell lines with hundreds of recombinase site insertions

I hypothesized that prime editors could simultaneously insert hundreds of recombinase sites into the genome of a cell if targeted to high copy-number elements. The feasibility of large-scale genome editing at repeating sequences in mammalian cell lines has been demonstrated by Smith et al., who targeted base editors to LINE-1s introducing more than 10,000 mutations in a single genome (Smith et al. 2020). Their best-performing sgRNA targeted a highly conserved region in the second open reading frame (ORF) of the LINE-1 retrotransposon (Figure 3.2) and had 5,832 perfect matches to the human reference genome (17,496 in a triploid genome). I converted this sgRNA into a prime editing gRNA (pegRNA) by adding a 3' extension containing a reverse transcriptase template with 34 nt overlap to the LINE-1 sequence, a loxP site, and a 13nt primer binding site (Figure 3.2) and co-transfected the pegRNA with prime editor 2 (PE2) into HEK293T cells (Figure 3.3a). I observed successful editing of 1.5% of all targeted LINE-1s (average of 200-300 insertions in triploid genomes of the population) after two days. However, the fraction of edited elements plummeted to only 0.02% by day 11 (Figure 3.3b), indicating that prime editing of a high copy number element was detrimental to survival.

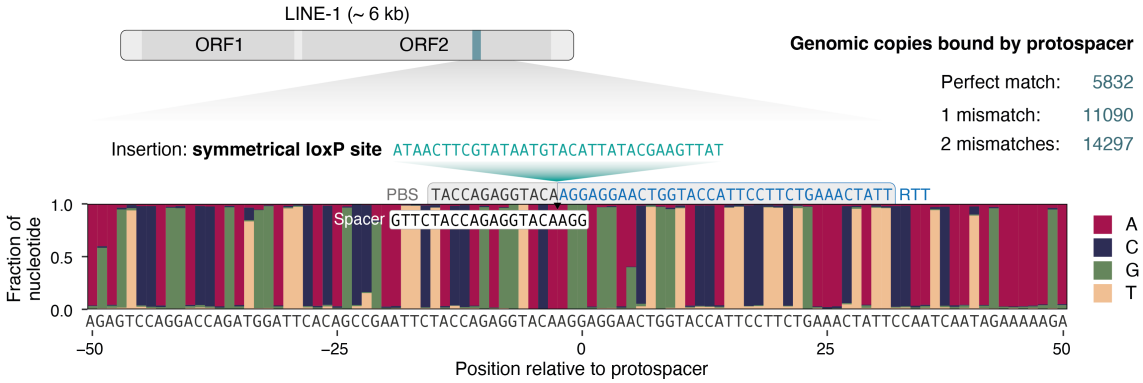

Figure 3.2 A prime editing guide RNA to insert recombinase sequences into thousands of sites across the genome. Schematic of a LINE-1 retrotransposon with two open reading frames (ORF), the target site, and protospacer of the pegRNA in blue. Lower panel: Nucleotide frequencies (y-axis) in a 100 bp window around the nicking site (x-axis) for all 14,297 LINE-1 sequences with 2 or fewer mismatches to the protospacer. The positions and sequences of the primer binding site (PBS) and reverse transcriptase template (RTT) including loxPsym insertion and homology arm are indicated.

The nicking Cas9 in prime editors creates single-strand DNA breaks and I speculated that thousands of simultaneous nicks could trigger the DNA damage response in the edited cells. Indeed, RNA sequencing (RNA-seq) of cells transfected with prime editor and the LINE-1 targeting pegRNA identified the upregulation of pathways associated with the DNA damage response, such as the p53 pathway (Figure 3.3c). Concurrently, there was downregulation in growth-related pathways, including Myc signaling and the oxidative phosphorylation pathway suggesting that cells might be slowing down growth in response to DNA damage. Similarly, Smith et al. observed toxicity when editing LINE-1 elements with nicking, but not with nuclease-dead base editors (Smith et al. 2020). Using nuclease-dead Cas9 is not currently possible for prime editing due to the chemistry of the reaction, which requires a single-strand break for priming.

While a burst of high prime editor activity from transient transfection overwhelms the cell, lower levels of editing and accumulation of loxPsym insertions into LINE-1 elements over a prolonged period may be tolerated. To enable this, I integrated multiple copies of doxycycline-inducible prime editors into HEK293T and HAP1 cells using the PiggyBac transposon system (Wolff et al. 2021) (Using the same cell line presented in Figure 2.2 which can insert a symmetrical loxP site into a single locus (*HEK3*) with up to 80% efficiency after 10 days). I further improved efficiency by using engineered pegRNAs (epegRNAs) that introduce a protective RNA pseudoknot motif to avoid 3'-end degradation and adding the p53-inhibitor piftherin-alpha and basic fibroblast growth factor (Niu et al. 2017; Nelson et al. 2022).

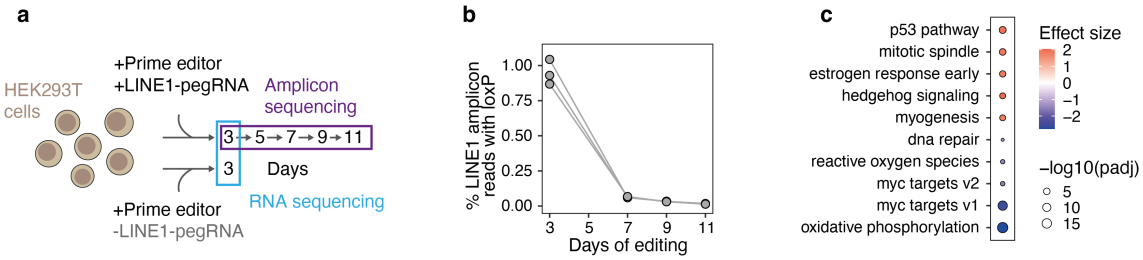

Figure 3.3. Prime editing thousands of sites simultaneously is detrimental to cells. a. Schematic of the experimental protocol for transient LINE-1 targeting pegRNAs. Samples at day 3 were taken for RNA sequencing (c) and samples transfected with both prime editor and pegRNA were analyzed for insertion frequency at days 3-11 (b). **b.** Frequencies of LINE-1 amplicon reads with loxPsym insertions (y-axis) over 11 days (x-axis) following transfection of the prime editing reagents. Markers and lines show one of n = 3 biological replicates. **c.** Top and bottom 5 differentially expressed pathways (by p-value) between cells transfected with prime editor alone or with prime editor and the LINE-1 targeting pegRNA. n = 2 biological replicates.

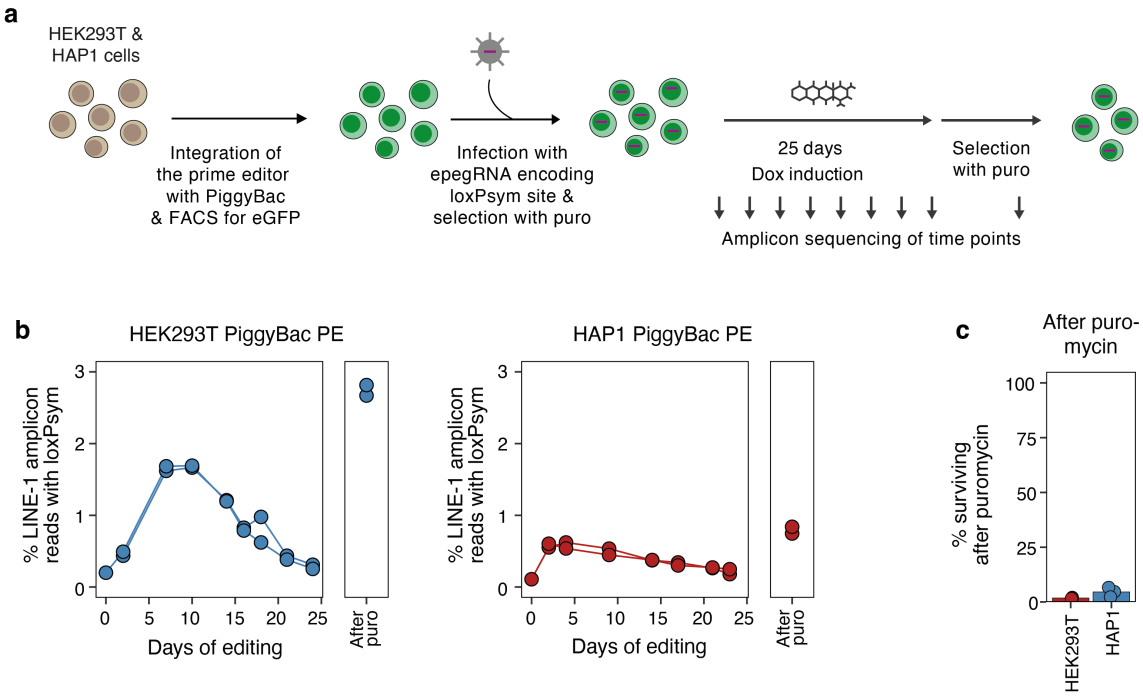

Figure 3.4 Continuous engineering of LINE-1 elements results in high insertion rates. a. Schematic of an editing time course with 25 days of editing and subsequent puromycin selection. Arrows pointing down indicate time points at which samples were taken for amplicon sequencing. **b.** Frequencies of LINE-1 amplicon reads with loxPsym insertions (y-axis) over 25 days of editing and after selection with puromycin at the end of the time course (x-axis) in HAP1 and HEK293T cells with stably integrated prime editor and engineered pegRNAs (panels). Markers and lines show one of N = 2 biological replicates. **c.** The percent of surviving cells after puromycin treatment (y-axis) for HEK293T and HAP1 cells (x-axis) after 25 days of loxPsym site insertion into LINE-1s. The integrated pegRNA vector encodes for puromycin resistance.

Putting these ideas together, I infected the stable, prime editing HEK293T and HAP1 cell lines with lentivirus encoding the LINE-1-targeting epegRNA, selected for lentiviral integration with puromycin. Editing was induced with 1 µM doxycycline and 10 µM piftherin-alpha (Figure 3.4a) and the insertion rates were monitored every 3-4 days. I observed an increase in edited LINE-1 elements over the first 2-5 days, peaking at an average of 1.7% on day 10 for HEK293T cells and 0.62% on day 4 for HAP1 cells followed by a gradual decline (Figure 3.4b). The apparent decline could be caused by cells that have inactivated the prime editor or the epegRNA and outcompete the cells that are still capable of editing. Indeed, selection for active expression from the epegRNA cassette using puromycin in the cell population after 32 days killed more than 90% of all cells (Figure 3.4c) and increased the average fraction of edited LINE-1 elements to 2.7% in HEK293T cells and 0.79% in HAP1 cells (Figure 3.4b) corresponding to the insertion of hundreds of recombinase sites.

I reasoned that single-cell sorting after a period of active editing could separate actively editing cells from cells that silenced the prime editing complexes (Figure 3.5a). After 11 days of editing, I derived and analyzed 16 clones for HEK293T cells and 15 clones for HAP1 cells, and observed editing rates between 0 and 3.3% for HEK293T cells and 0 and 2.1% for HAP1 cells (Figure 3.5b). To further increase the number of loxPsym site insertions, I subjected one HAP1 clone with 1.9% initial editing to another 12 days of doxycycline induction followed by a second round of subcloning which resulted in up to 8% LINE-1 editing (Figure 3.5c).

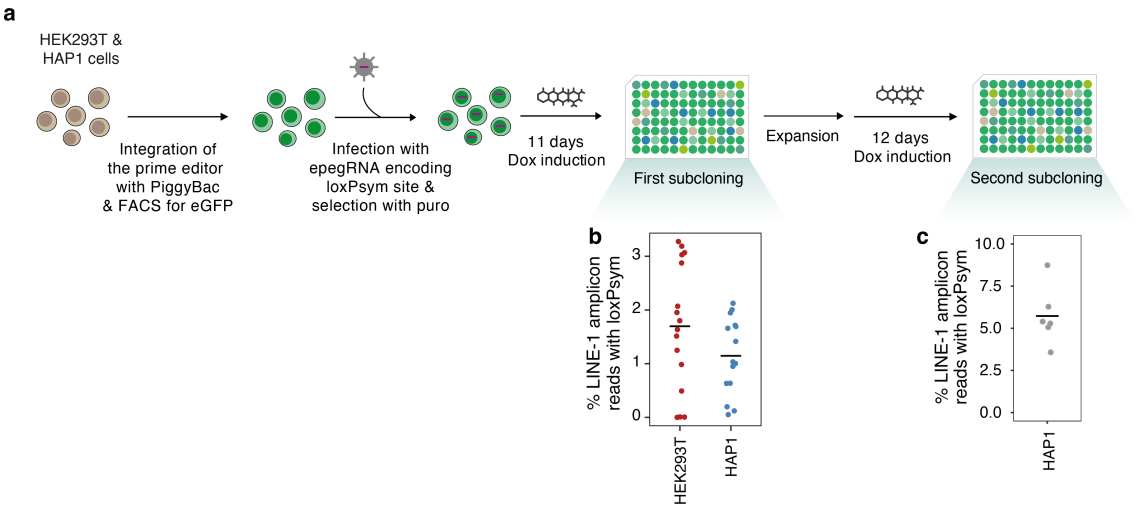

Figure 3.5 Engineering clones with hundreds of loxPsym insertions. a. Schematic of an editing time course with two subcloning steps to enrich cells with hundreds of loxPsym insertions. **b.** Frequencies of LINE-1 amplicon reads with loxPsym insertions (y-axis) for HEK293T or HAP1 clones (markers) after one round of subcloning. **c.** Frequencies of LINE-1 amplicon reads with loxPsym insertions (y-axis) for HAP1 clones (markers) after two rounds of subcloning.

3.2.2 Characterization of edited cells with hundreds of loxP insertions

To precisely map the locations and number of integrated loxPsym sites, I submitted two LINE-1 edited clones - HAP1-loxPsym(301) and HEK293T-loxPsym(638) for whole-genome sequencing using long-read technology (Oxford Nanopore, Figure 3.6). Mapping the 301 clonal insertions in the HAP1 and 638 in the HEK293T clone revealed a broad distribution across the genome with 14.3 sites on average per chromosome (range 1-34) for the HAP1 cells and a strong correlation with the number of LINE-1 elements with perfect matches to the pegRNA (HAP1: R^2=0.83, HEK293T: R^2=0.90, Figure 3.7a). The insertions across the two clones mapped to largely different LINE-1 elements and only 26 were shared (Figure 3.7b). The median allele frequencies of loxPsym site insertions were 1.00 for HAP1 cells and 0.35 for HEK293T-loxPsym(638) cells, consistent with their haploid or mostly triploid genomes (Figure 3.7c).

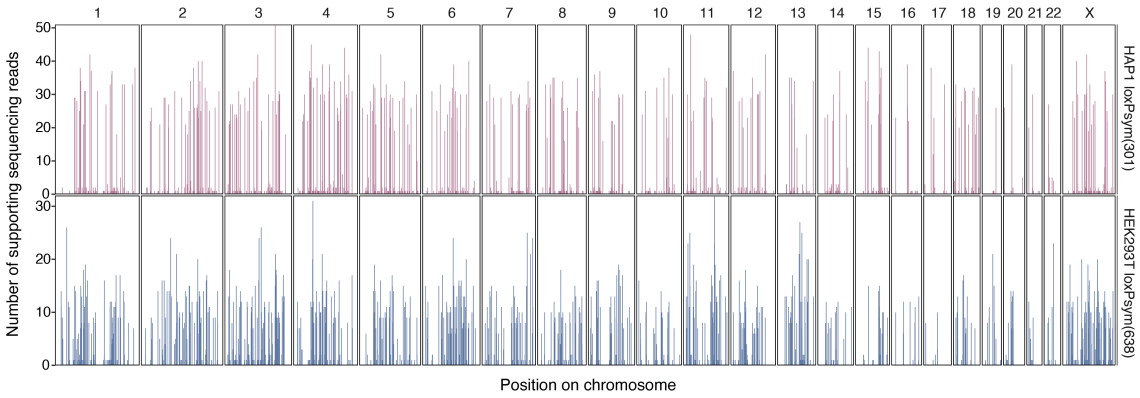

Figure 3.6. Cell lines with hundreds of loxPsym sites distributed across their genomes. Locations of loxPsym site integrations (vertical bars) and the number of supporting sequencing reads (y-axis) across chromosome positions (x-axis) for three clones (panels).

LINE-1 sequences represent sources of near-identical sequences that could be used to understand variation in prime editing efficiencies. To map insertion sites at higher throughput, I devised a Cas9-enrichment-based long-read sequencing method that targets the Cas9 nuclease to LINE-1s with loxPsym insertions (Gilpatrick et al. 2020; McDonald et al. 2021) (Figure 3.8). Hereby, high molecular weight genomic DNA is harvested and dephosphorylated to prevent spurious adaptor ligation. Next, I designed a guide RNA that targets a hybrid sequence containing a loxPsym insertion and a PAM and sequence from the surrounding LINE-1 (Figure 3.8). Sequencing adaptors can only ligate to phosphorylated DNA ends released by Cas9 cleavage of edited LINE-1s. Editing rates of LINE-1s can then be determined by counting sequencing reads.

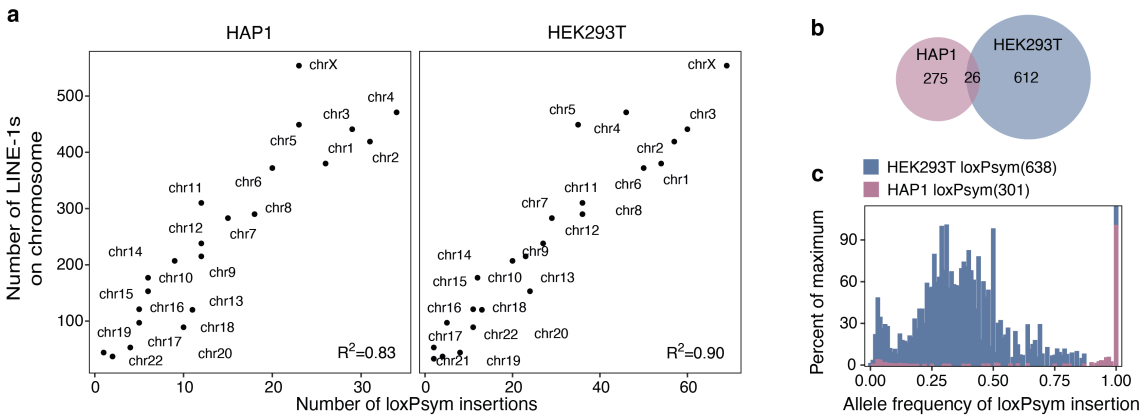

Figure 3.7. Characteristics of LINE-1 insertions. a. The number of loxPsym insertions (x-axis) versus the number of LINE-1 retrotransposons with up to two mismatches to the protospacer (y-axis) in HAP1 and HEK293T cells (panels) for chromosomes (markers). **b.** Overlap of loxPsym sites between the two sequenced clones (colors, circles). The area of the Venn diagram is proportional to the number of loxPsym sites. **c.** The abundance of loxPsym insertions (percent of the allele frequency with maximum abundance, y-axis) at various allele frequencies (x-axis) in the two sequenced clones (colors).

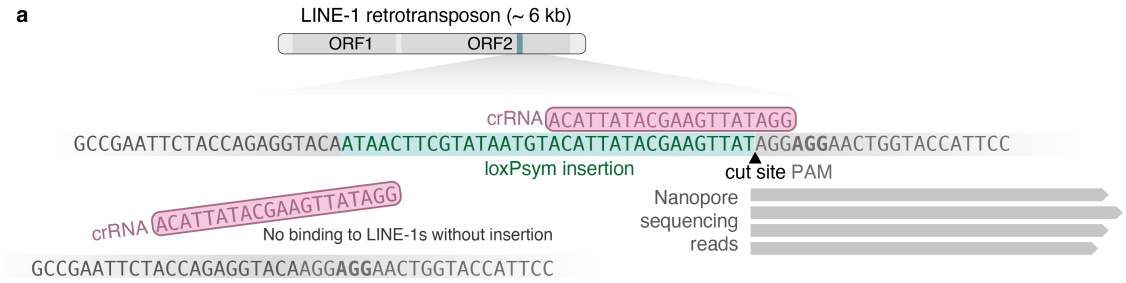

Figure 3.8. A strategy to identify edited LINE-1 elements with high throughput. Schematic of a strategy to map edited LINE-1s at high throughput. The sequence at the top represents a LINE-1 with loxPsym insertion (teal) to which the Cas9-enrichment crRNA can bind (pink).

I observed that LINE-1s with up to two mismatches to the protospacer were still successfully edited at 25 days, but at lower frequencies (27% and 15% in HAP1 and HEK293T cells, Figure 3.9a). In contrast, one and two mismatches in the 34 nt homology arm of the reverse transcriptase template were well tolerated, but editing rates dropped for elements with three or more mismatches (Figure 3.9b).

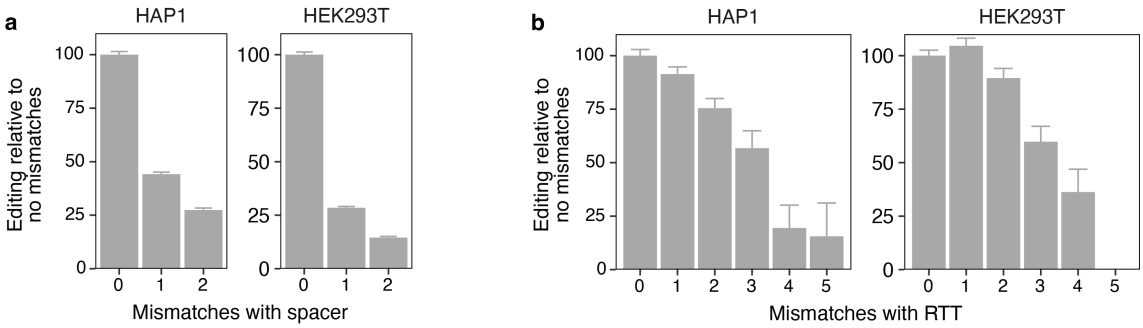

Figure 3.9. The effect of mismatches on prime editing rates. Average sequencing reads from targeted sequencing of edited LINE-1s normalized to no mismatches (y-axis) for LINE-1s with different numbers of mismatches to the pegRNA protospacer (x-axis) across two cell types (panels). Error bars: standard error of the mean of n = 2 biological replicates. (**E**) As (D) but for LINE-1s with mismatches with the reverse transcriptase template (x-axis). The sum of n = 2 biological replicates.

To understand how chromatin affects prime editing, I collected eight publicly available DNase, ATAC-seq, and ChIP-seq data sets in HAP1 and HEK293T cells (Table 3.1) and trained the ChromHMM model with them to obtain chromatin states (Figure 3.10) (Ernst and Kellis 2017). I next correlated the abundance of chromatin states in the LINE-1s and 3 kb of surrounding sequence to the editing rate (Figure 3.11a). Editing events were most enriched in LINE-1s with transcription and enhancer-related chromatin signatures, and depleted in quiescent (no chromatin marks) or repressed/heterochromatic elements (Figure 3.11b-c). This result is consistent with recent work exploring chromatin context for prime editing (Mathis, Allam, Tálas, et al. 2023; Xiaoyi Li et al. 2023).

Table 3.1: Datasets used to determine chromatin states in HAP1 and HEK293T cells

Cell line	Chromatin Mark	Accession number	Accession control
HAP1	H3K4me3	ENCFF461TZF	ENCFF247DSQ
HAP1	H3K27me3	ENCFF708HAB	ENCFF247DSQ
HAP1	H3K9me3	ENCFF528UHF	ENCFF247DSQ
HAP1	K3K36me3	ENCFF216JJJ	ENCFF247DSQ
HAP1	H3K4me1	ENCFF639UYT	ENCFF247DSQ
HAP1	DNase-seq	ENCFF162WTC	ENCFF247DSQ
HAP1	H3K27ac	ENCFF742SZS	ENCFF247DSQ

HEK293T	DNase-seq	ENCFF969MBJ	None
HEK293T	H3K4me1	SRR10981645	None
HEK293T	H3K36me3	SRR5627148	None
HEK293T	H3K9me3	SRR11453034	None
HEK293T	H3K27me3	SRR8937480	None
HEK293T	H3K4me3	SRR8937479	None
HEK293T	H3K27ac	SRR1016003	None

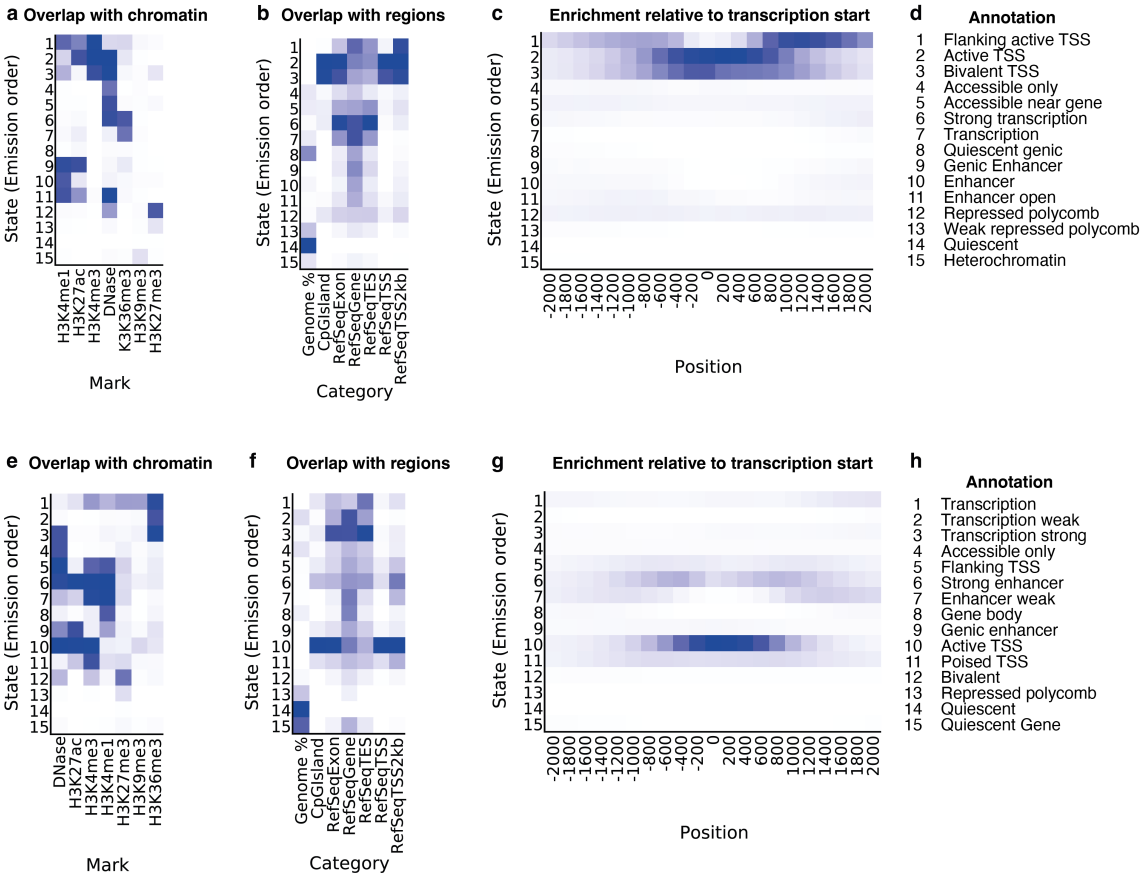

Figure 3.10 Chromatin states in HAP1 and HEK293T cells. a. The overlap of ChromHMM emission states (y-axis) with chromatin marks (x-axis). **b.** As in (a) but for different types of sequence regions in the genome. **c.** The enrichment of ChromHMM emission states (y-axis) relative to varying distances to transcription start sites (x-axis). **d.** Manual annotation of emission states based on (a-c). **e-h**. As in (a-d) but for HEK293T cells.

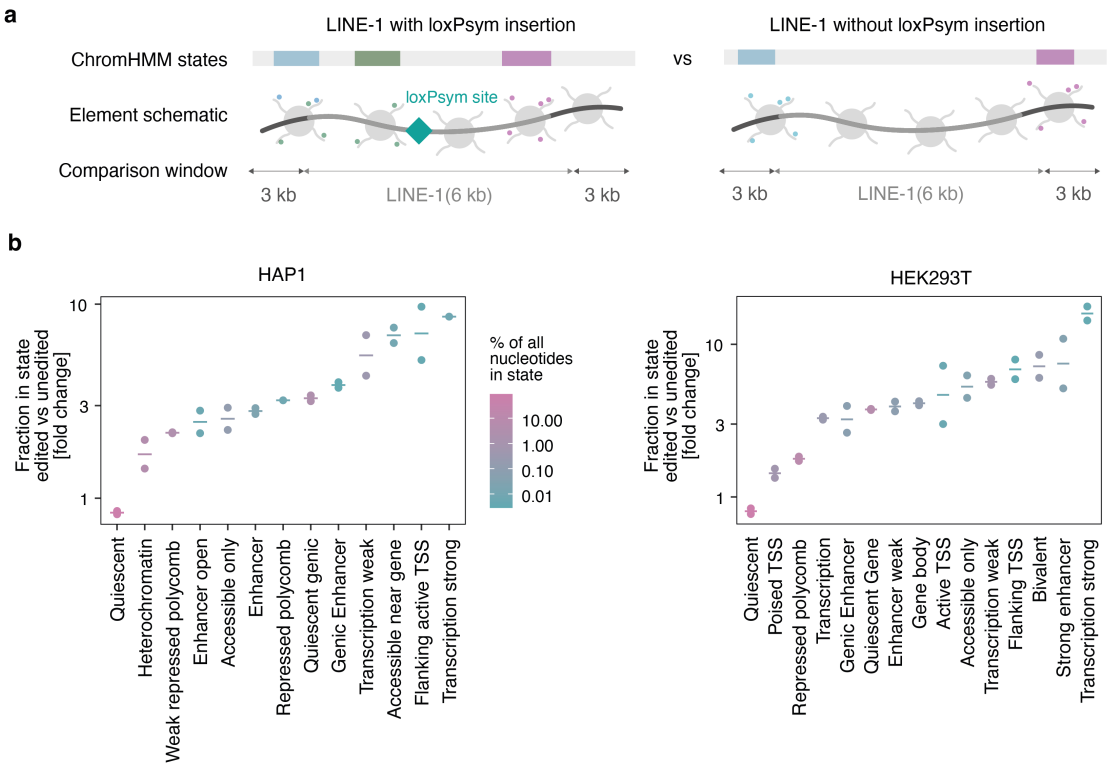

Figure 3.11 LINE-1s in active chromatin are edited more frequently. a. Schematic of the comparison in chromatin composition between edited (n > 1 read) and unedited (n = 0 reads) LINE-1s. **b.** Fraction of edited to non-edited LINE-1s (y-axis) for overall nucleotides in chromatin state (x-axis) in HAP1 and HEK293T cells (panels) colored by the fraction of all nucleotides in the human genome that are in a given state. Cross bars represent averages from n = 2 biological replicates (markers).

3.2.3 Scrambling the human genome

To scramble the genome of clones with hundreds of loxPsym sites, I treated the cells with membrane-permeable TAT-Cre protein for 4 hours, recovered for 16 hours, and used long-read sequencing to map the Cre-induced rearrangements (Figure 3.12a). I developed a computational pipeline to call Cre-induced variants by first running standard variant callers in relaxed settings and then filtering stringently for rearrangements that originate within 300 bp of a pegRNA target site and contain a loxPsym sequence at the breakpoints (Figure 3.12b).

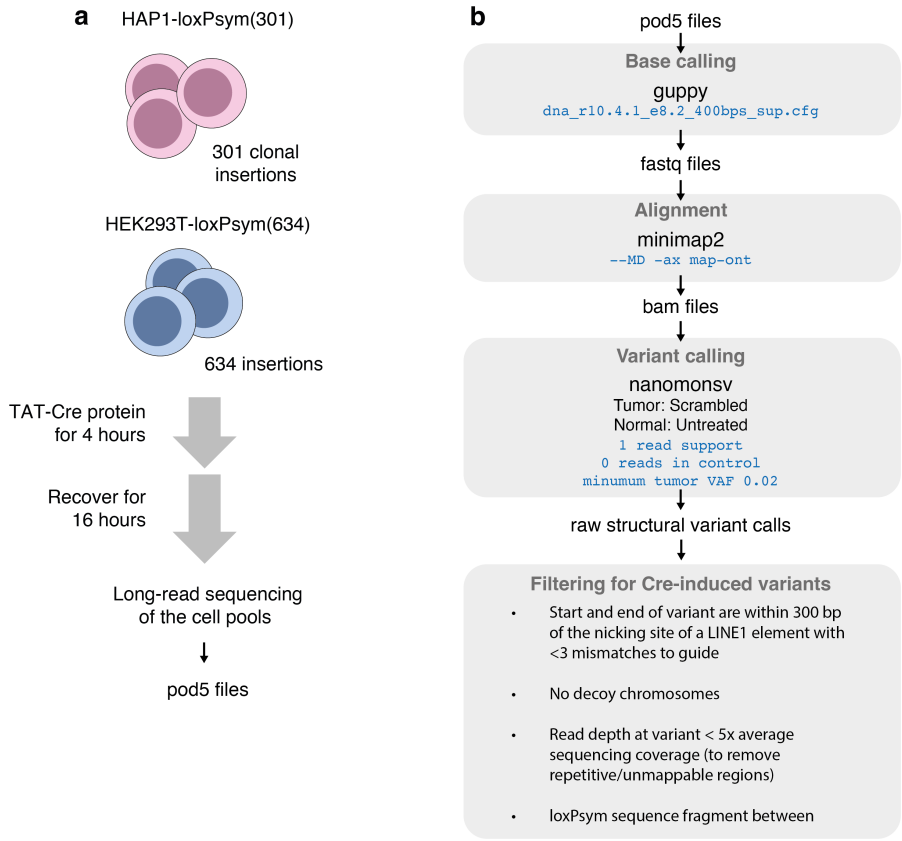

Figure 3.12 Experimental setup and variant calling strategy for Cre-induced rearrangements. a. Schematic of the Cre-induction protocol. Cell lines with loxPsym insertions were treated with membrane-permeable TAT-Cre protein and sequenced with long-read technology. **b.** Schematic outlining the bioinformatics pipeline used to process sequencing data and call Cre-induced structural variants. pod5 files were base called using Guppy and aligned with Minimap2. Nanomonsv was used with relaxed parameters to call raw structural variants which were filtered based on proximity to LINE-1s, absence of decoy chromosomes, read depth thresholds, and the presence of loxPsym sequence fragments between breakpoints.

At this early time point, selection pressures would not have had time to deplete variants that are detrimental to survival or chromosome segregation. Each read with a rearrangement likely represents an independent recombination, since the complexity of the cell pool is much higher than the sequencing coverage, the cells had no time to replicate, and the DNA was not amplified for sequencing. The number of rearrangements scales linearly with the number of loxPsym sites per chromosome, both for variants within the same chromosome (*cis)* (R^2=0.92/0.95 in HAP1 and HEK293T) and across different ones (*trans)* (R^2=0.83/0.77, Figure 3.13). This suggests that increasing the density of loxPsym sites should also increase the number of rearrangements per cell. With 50x and 23x genome coverage, I detected 6,481 and 1,166 rearrangements in HAP1 and HEK293T cells respectively (Figure 3.14). Assuming haploidy in HAP1 and triploidy in HEK293T, Cre treatment induced 131 or 143 rearrangements per cell on average.

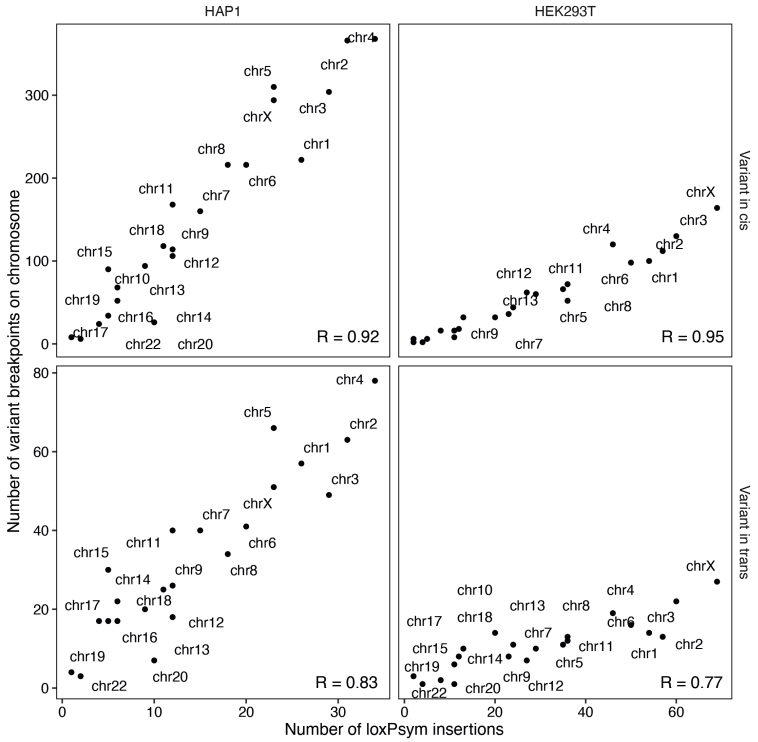

Figure 3.13. The number of rearrangements scales linearly with the number of loxPsym insertions per chromosome. The number of loxPsym insertions (x-axis) versus the number of variant breakpoints on chromosomes (y-axis) in HAP1 and HEK293T cells (panels) for chromosomes (markers).

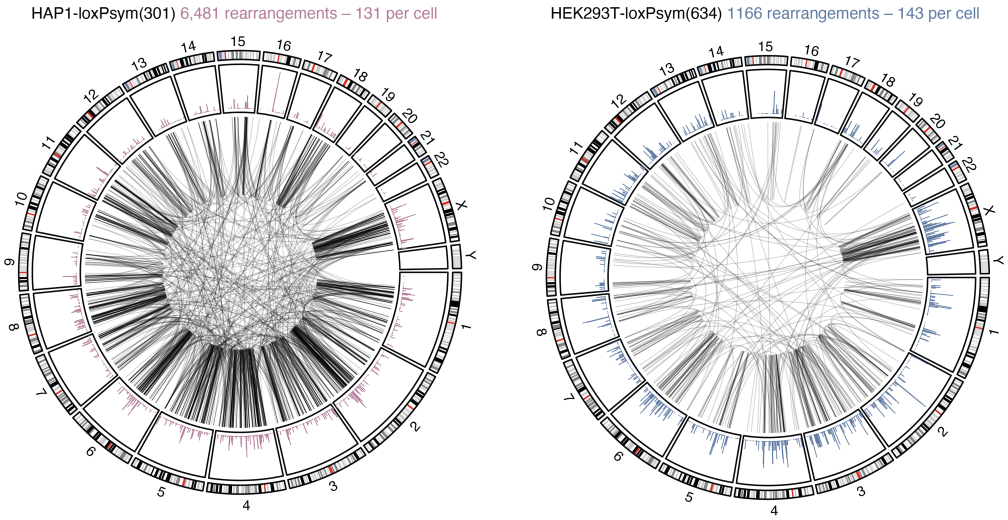

Figure 3.14 Cre recombinase induced thousands of rearrangements. Overview of all structural variants produced in the two clones after 4 hours of scrambling and 16 hours of recovery. From outside to inside: Chromosomes (ideograms), integration locations and rearrangement frequencies for loxPsym sites (bars), Cre induced structural variants (arcs).

Whenever Cre recognizes two loxPsym sites and catalyzes a rearrangement, the outcome can fall into different classes based on the orientation of the DNA ends joined and whether this occurs in *cis* or *trans*. These different events can be distinguished based on the mapping of sequencing reads at breakpoints (Figure 3.15a). If two loxPsym sites are located in *cis*, Cre can produce a pair of inversions (10% or 14% of rearrangements in HAP1 and HEK293T cells respectively, Figure 3.15b), or a pair consisting of an extrachromosomal circular DNA (ecDNA) (HAP1: 11%; HEK293T: 10%) and a deletion (HAP1: 4.5%; HEK293T: 6.6%). In contrast, if the two sites are located on two non-homologous chromosomes in *trans*, the outcome will be a translocation (HAP1: 10%; HEK293T: 11%).

Curiously, the most common class of variants that I observed were fold-backs (HAP1: 68%; HEK293T: 58%) which could arise between homologous chromosomes or two sister chromatids of the same chromosome after replication. The fraction of fold-backs was indeed reduced 3.1-fold, down to 22% of total variants, when HAP1-loxPsym(301) cells were arrested in G1 using the CDK4/6 inhibitor palbociclib (Figure 3.16) before Cre induction (Figure 3.15b). Therefore, cell cycle modulation can alter the types of variants generated. This is important since fold-backs that generate a pair of acentric (no centromere) and dicentric (two centromeres) chromosomes could be catastrophic for cell survival due to chromosome segregation errors during cell division. Overall, the frequency of variant types was similar between the two cell lines despite their different ploidy (Figure 3.15b).

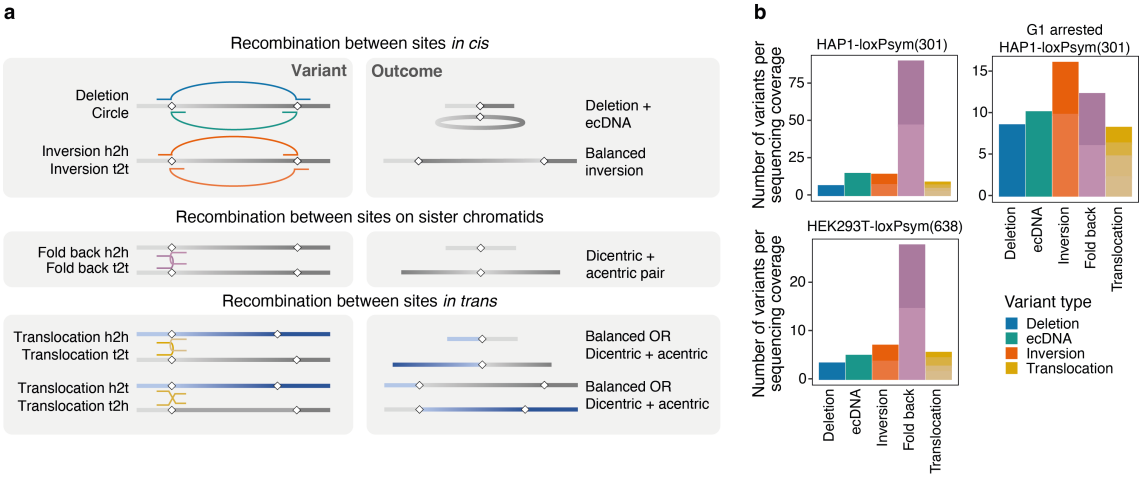

Figure 3.15. Cre induction generates a variety of variant classes. a. Schematic of the classes of variants that can be generated by Cre induction. Gray and blue lines represent chromosome sequence according to the reference genome with diamonds indicating the locations of loxPsym sites. Arcs represent rearrangements that join two DNA segments of varying orientation and location. Color according to variant class. h2h: head to head, t2t tail to head. **b.** The types of structural variants produced (x-axis, color according to a) and their counts (y-axis) and separated by cell line and whether they were inhibited in G1/S during Cre induction (panels).

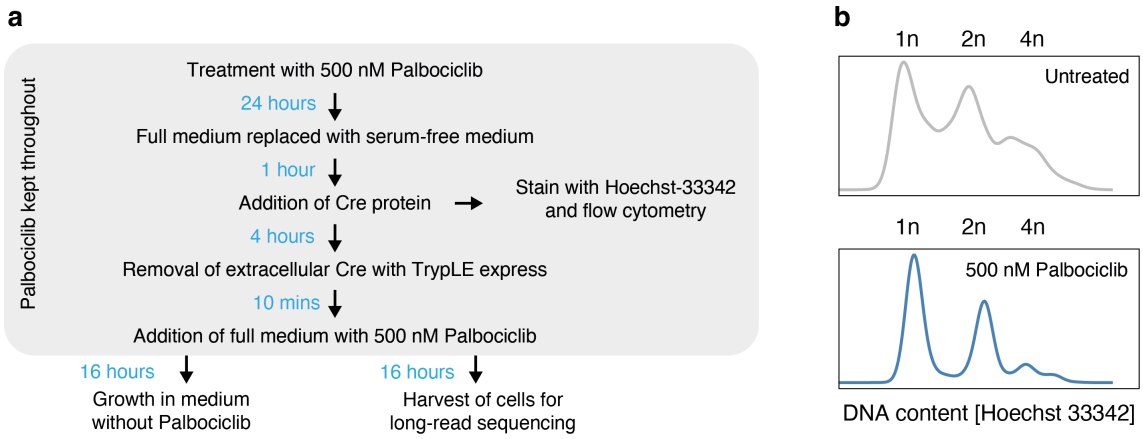

Figure 3.16. Palbociclib arrests cells in G1 before scrambling. a. Schematic illustrating the experimental protocol for treating cells with palbociclib. **b.** Frequencies (y-axis) of DNA content (x-axis) in untreated cells (top panel) and cells treated with 500 nM palbociclib (bottom panel). Likely genome copies (1n, 2n, 4n) are indicated.

The frequency of deletions and inversions decreased with increasing distance between loxPsym sites, with 71% or 63% of variants shorter than 10 Mb for HAP1 and HEK293T cells (Figure 3.17a). The median variant length was 3.0 or 4.9 Mb (HAP1/HEK293T), which in HAP1 cells is 15.7 times smaller than the expectation of 47 Mb assuming an equal probability of recombination for any given pair of loxPsym sites on the same chromosome (Figure 3.17b, middle panel). The observed data could be predicted better by reducing the expected frequency of two loxPsym sites rearranging by 7.5% for each Mb distance between them (exponential decay model, Figure 3.17b, lower panel, Figure 3.17c), consistent with previous observations that the efficiency of Cre-induced rearrangements decays with distance (Zheng et al. 2000).

The exponential decay model of generating rearrangements in cis predicts a dearth of long variants; however, I observe a tail of very long deletions and duplications (Figure 3.17a). This suggests that other factors, such as distance in 3D, may influence the probability of observing a rearrangement. A model that uses 3D distance as assessed by counting Hi-C read pairs fits the data better (Pearson's R = 0.68 (Hi-C, (Lohia, Fox, and Gillis 2022)) vs 0.55 (log-transformed length), Figure 3.17d). A linear model with both linear distance and Hi-C pairs only modestly outperformed one with only Hi-C pairs (R = 0.68 vs 0.67) suggesting that 3D contact frequency between two recognition sites is a better explanation of Cre efficiency. Overall, these results are consistent with a recent study by Zhou et al. profiling 260,000 Cre-induced rearrangements in SCRaMbLEd yeasts (Zhou et al. 2022) and observed that rearrangement frequency was driven by contact frequency and DNA accessibility.

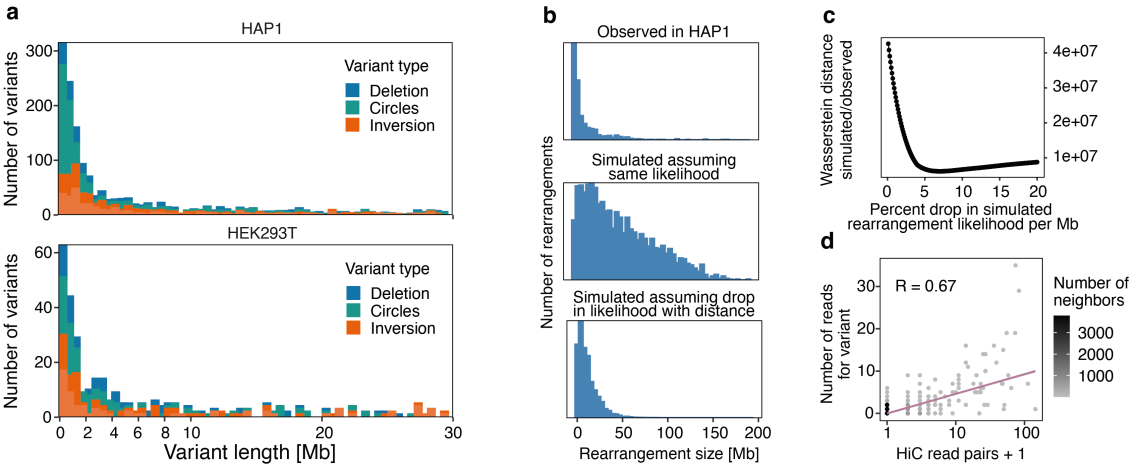

Figure 3.17. Rearrangement frequency decays with increasing distance between loxPsym sites. a. The number of structural variants in *cis* (y-axis) stratified by the distance between the two breakpoints (x-axis). **b.** Histograms contrasting the number of observed rearrangements (upper panel, y-axis) with those generated by simulation assuming equal likelihood of recombination between any two sites on the same chromosome (middle panel) across different rearrangement sizes (x-axis, with 50 bins) in HAP1 cells or assuming a 7.5% decrease in recombination likelihood per mb (bottom panel). **c.** The Wasserstein distance between simulated and observed data (y-axis) for various reductions in rearrangement likelihood per Mb (in increments of 0.1%, x-axis). **d.** Number of supporting sequencing reads (y-axis) for cis-variants in HAP1 cells (markers) with different numbers of read pairs in an independent Hi-C experiment in HAP1 cells (x-axis). Line: linear regression fit.

3.2.4 Selection pressures shape surviving variants

To understand how selection acts on the various types of rearrangements generated after Cre induction, I grew out the HAP1 and HEK293T cells that survived the scrambling process for 13-15 days and long read sequenced the cell pools at 53-164 times whole genome coverage. In addition, I single-cell sorted cells that showed evidence of recombination based on an integrated reporter and sequenced 21 expanded colonies. I observed a 188-fold (HAP1) or 87-fold (HEK293T) depletion of variants in the pools over two weeks of growth (Figure 3.18a), implying strong selection against the majority of rearrangements. After selection, an average of 0.7 rearrangements remained per HAP1 cell (assuming haploidy) or 1.6 per HEK293T cell (assuming triploidy). Similarly, cells that rearranged an integrated Cre activity reporter (and switched from dsRed to tBFP expression upon recombination) depleted from 68% on day 1 to 1% on day 10 suggesting strong selection against cells that took up Cre (Figure 3.18b,c). The depletion was less pronounced for both the reporter and remaining variants in the pool (2.4 on average per cell) if the cells were arrested in the G1 phase with palbociclib before Cre induction (Figure 3.18a,b), consistent with the generation of fewer fold-backs (Figure 3.15b).

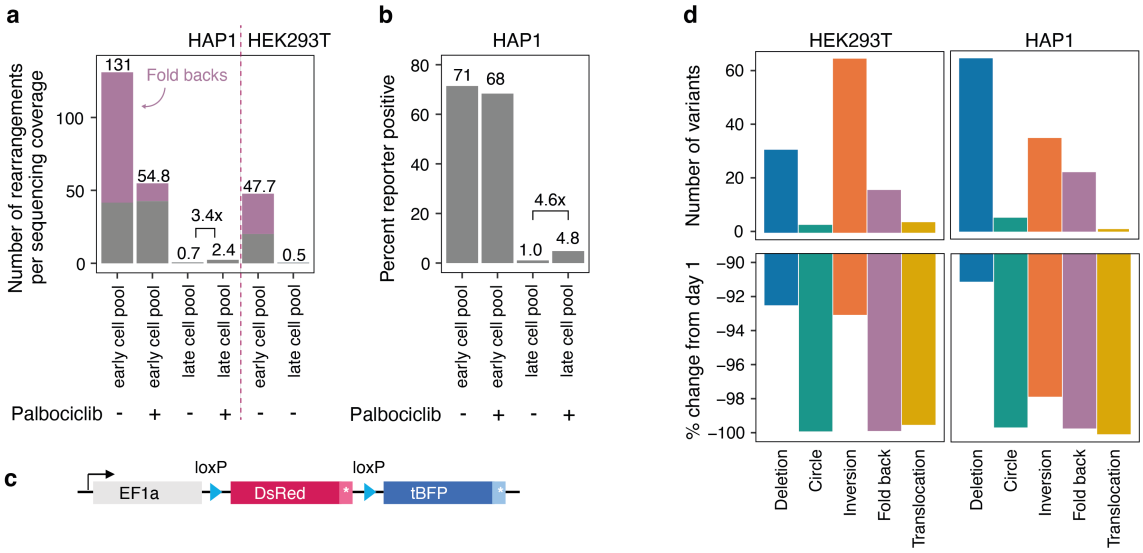

Figure 3.18. The majority of cells with rearrangements do not persist after weeks in culture. **a.** Number of total Cre-induced rearrangements per sequencing coverage (y-axis) for early and late cell pools that were untreated or inhibited in G1 prior to scrambling (x-axis). Bars are colored by variant type (lighter blue: fold-backs, darker blue: all other types). **b.** Percent of cells that were BFP-positive in flow cytometry (y-axis) after rearranging a genome-integrated Cre reporter (schematic bottom) at different time points (x-axis). The cells were either restricted in G1/S with palbociclib or not before scrambling (x-axis). **c.** Schematic of the Cre reporter. **d.** Upper panel: The types of structural variants produced (x-axis) and their counts (y-axis) in HAP1 and HEK293T cells (panels) 13-15 days after Cre induction. Lower panel: The percent difference in variant abundance per sequencing coverage compared to day 1.

The frequency of variant types I observed before (Figure 3.15b) and after scrambling (Figure 3.18d) differs starkly. Because Cre joins up all DNA ends of the two loxPsym sites involved in the recombination, all structural variants are initially balanced. However, some rearrangement products lack centromeres (ecDNAs and acentric chromosomes from fold-backs or translocations), cannot be segregated properly and will be lost during subsequent cell cycles. Indeed, between HAP1 and HEK293T cells, I see a 99.9/99.6% depletion of ecDNAs, a 99.8/99.7% depletion of fold-backs, and a 99.5/100% loss of translocations (Figure 3.18d). However, the depletion of translocations cannot be explained by centromere configuration alone since 50% of translocations should result in two derivative chromosomes with one centromere each. Deletions (HAP1: -93.0%, HEK293T: -97.8%) and inversions (HAP1: -92.4%, HEK293T: -91.0%) were less depleted compared to ecDNAs, fold-backs, or translocations, but fewer than 1 in 10 variants remained.

Deletions might be selected against because they remove essential genes. To test this experimentally, I analyzed data from a genome-wide CRISPR knockout screen in wild-type

HAP1 cells that Elin Madil Peets from our lab performed (Peets et al. 2019). I defined essentiality as the average fold change of gene-targeting guides after two weeks in culture. Essentiality values < 1 indicate a growth defect and values > 1 indicate a growth advantage upon loss. The essentiality values were correlated between replicates (R = 0.57, Figure 3.19a), and essential genes (average essentiality = 0.72) were overlapping but separated from non-essential genes (average = 1.1) and non-targeting controls (average = 2.0, Figure 3.19b).

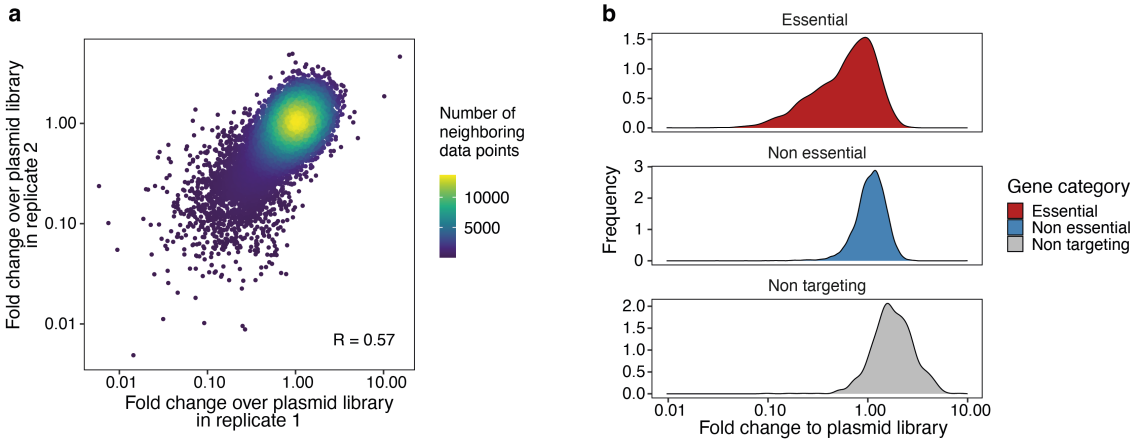

Figure 3.19. Genome-wide CRISPR knockout screen in HAP1 cells. a. Fold change in gene perturbation frequency over the plasmid library between two independent replicates (axes) colored by the number of neighboring data points. **b.** Frequencies (y-axis) of fold change to plasmid library (x-axis) for gene perturbations categorized by gene essentiality (panels and colors).

Equipped with this data set, I could analyze if the essentiality of genes deleted early and late differed. Only 2/17 deletions removed any gene with a >50% growth defect in the HAP1 knockout screen, in contrast to early deletions where 137/185 did (Figure 3.20a). Longer variants have more chance to delete essential DNA and I indeed see a consolidation towards 4.5 to 5.0-fold shorter deletions after weeks in culture for HEK293T and HAP1 cells respectively (median early 3.0/4.2 Mb vs late 0.60/0.93 Mb, Figure 3.20b). Moreover, genes contained within persisting deletions were enriched for ones whose loss conferred a growth advantage in the CRISPR knockout screen (average of genes in the deletions = 1.3, t-test $p=10^{-17}$, Figure 3.20c). A much weaker selection was apparent for inversions (average = 1.08, $p = 5x10^{-5}$, Figure 3.20d).

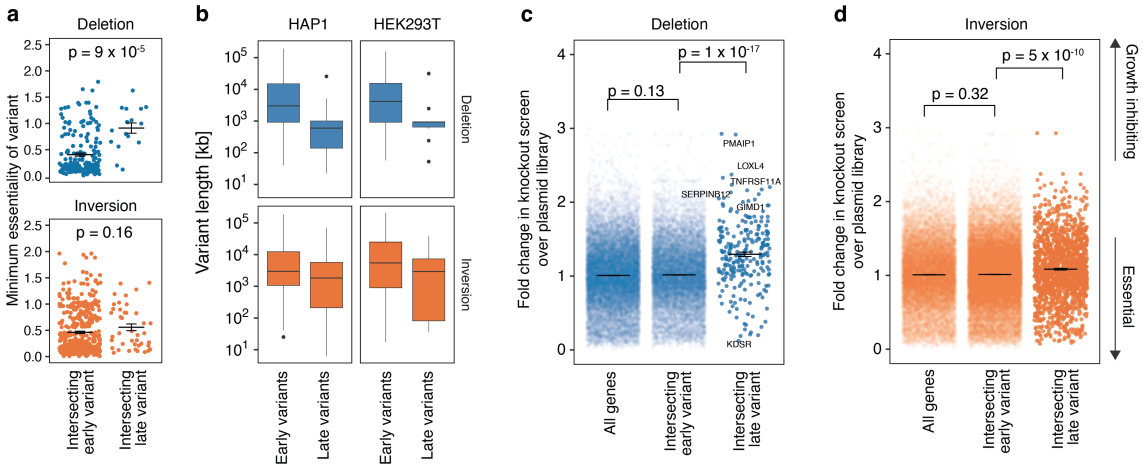

Figure 3.20 Late deletions are depleted in essential genes. a. Median essentiality scores for genes (y-axis) affected by deletion (top) and inversion (bottom) variants (markers), stratified by occurrence in early or late time points (x-axis). Bars indicate the mean and standard error of the mean. **b.** Variant length (y-axis) for early and late variants (x-axis) stratified by cell line and variant type (panels). Box: median and quartiles. Whiskers: Largest or smallest value no further than 1.5 interquartile ranges from the hinge. **c.** Average fold changes of four guides targeting a gene (markers) in a HAP1 CRISPR knockout screen (y-axis) for all genes, genes intersecting early deletions (day 1) or intersecting late (day 13-14, x-axis). **d.** As in (c) but for inversions.

To better understand the genomic features under selection Thomas Vanderstichele compared 822 early and 70 late time point inversions and deletions from HAP1 cells across 47 features spanning chromatin states (Figure 3.10), conservation, mutational constraint, regulatory elements, and sequence features. Late deletions were significantly depleted in constrained regions (gnomAD z-score), gene annotations (coding sequences, exons, start and stop codons, UTRs), and active regulatory elements (transcription start sites and enhancers) (Figure 3.21, Figure 3.22). Conversely, late deletions were enriched for polycomb repressed regions and lamina-associated domains, which are usually heterochromatic. As expected, inversions showed weaker signals of selection across all features but were significantly shorter at later time points and depleted of super-enhancers. Together this data suggests that deletions are under strong selection that can be explained through variant features. However, it should be noted that while some of the feature enrichments will be true imprints of selection, others could show up due to correlation with features under selection (Figure 3.21b). For example, gene annotations are correlated with active chromatin and anticorrelated with lamina-associated domains (Figure 3.21b).

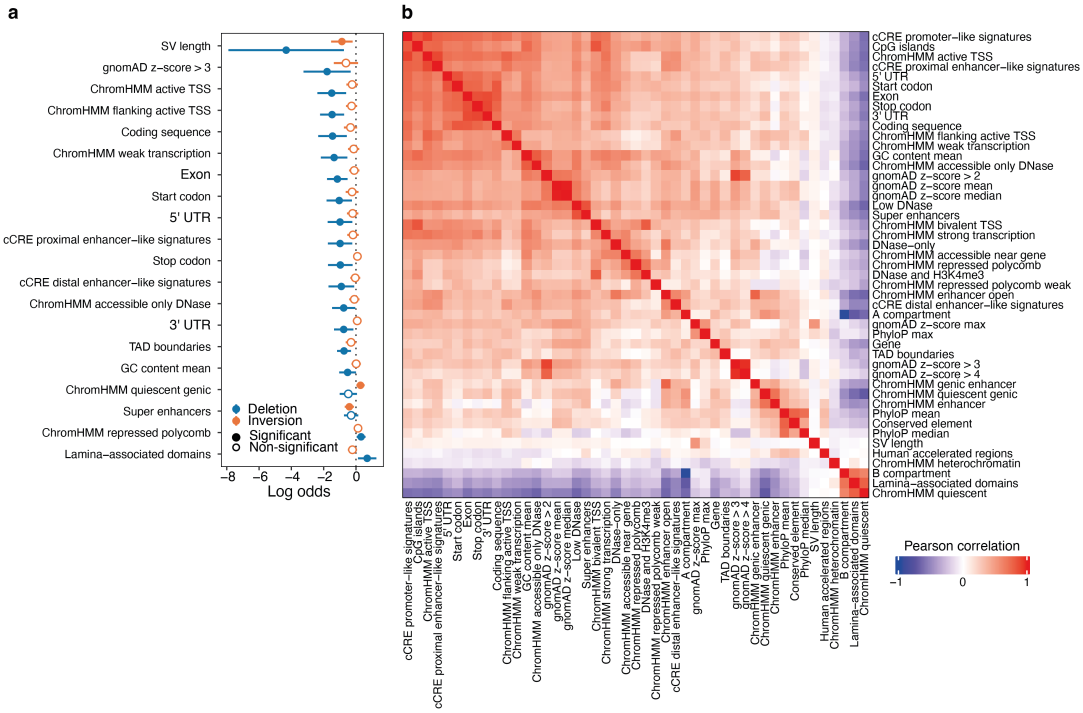

Figure 3.21 Surviving variants have a set of features that is distinct from initially generated variants. a. Log odds (x-axis) for features (y-axis) between early variants and surviving variants colored by variant type. Markers represent the log odd and whiskers 95% confidence intervals. Unless indicated otherwise, the fraction of each variant covered by the respective feature is used as the statistic. Only significant features are shown. **b.** Heatmap of correlations (colored by Pearson's R) between various genomic and epigenomic features (x and y axes) and their frequency of occurrence within deletion and inversion structural variants.

Many of the remaining deletions in HAP1 cells were millions of base pairs long (Table 3.1), highlighting that even in haploid cells substantial amounts of DNA are dispensable for growth under cell culture conditions. One 4.3 Mb variant on chromosome 18 deleted 19 expressed (TPM > 0.1) genes with an average essentiality of 1.43. This variant removed both a selectively essential as well as a strongly growth-suppressing gene: 3-ketodihydrosphingosine reductase (*KDSR*) had a relative fitness of 0.2 in HAP1 and was selectively essential in the cancer dependency map (183/1095 cell lines (Meyers et al. 2017)). Conversely, *PMAIP1* is a pro-apoptotic gene and its loss was highly growth-promoting in the HAP1 screen (relative fitness = 2.9) and 1035/1095 cancer dependency map cell lines. Curiously, the loxPsym site marking the end for this deletion was the starting point for another, independent deletion that covered another 5.1 Mb on chromosome 18 (Table 3.2) but did not contain any essential genes in the CRISPR screen (fold change range of 11 deleted genes: 1.0-2.0). Since the sequencing was done from a pool of cells it is possible that some of the variants observed did not happen in isolation or occurred in cells that previously reverted to diploidy (14% of the cell population at the time of sequencing). This is likely the case for one 25.6 Mb variant that deleted 231 genes on chromosome X including many essential ones.

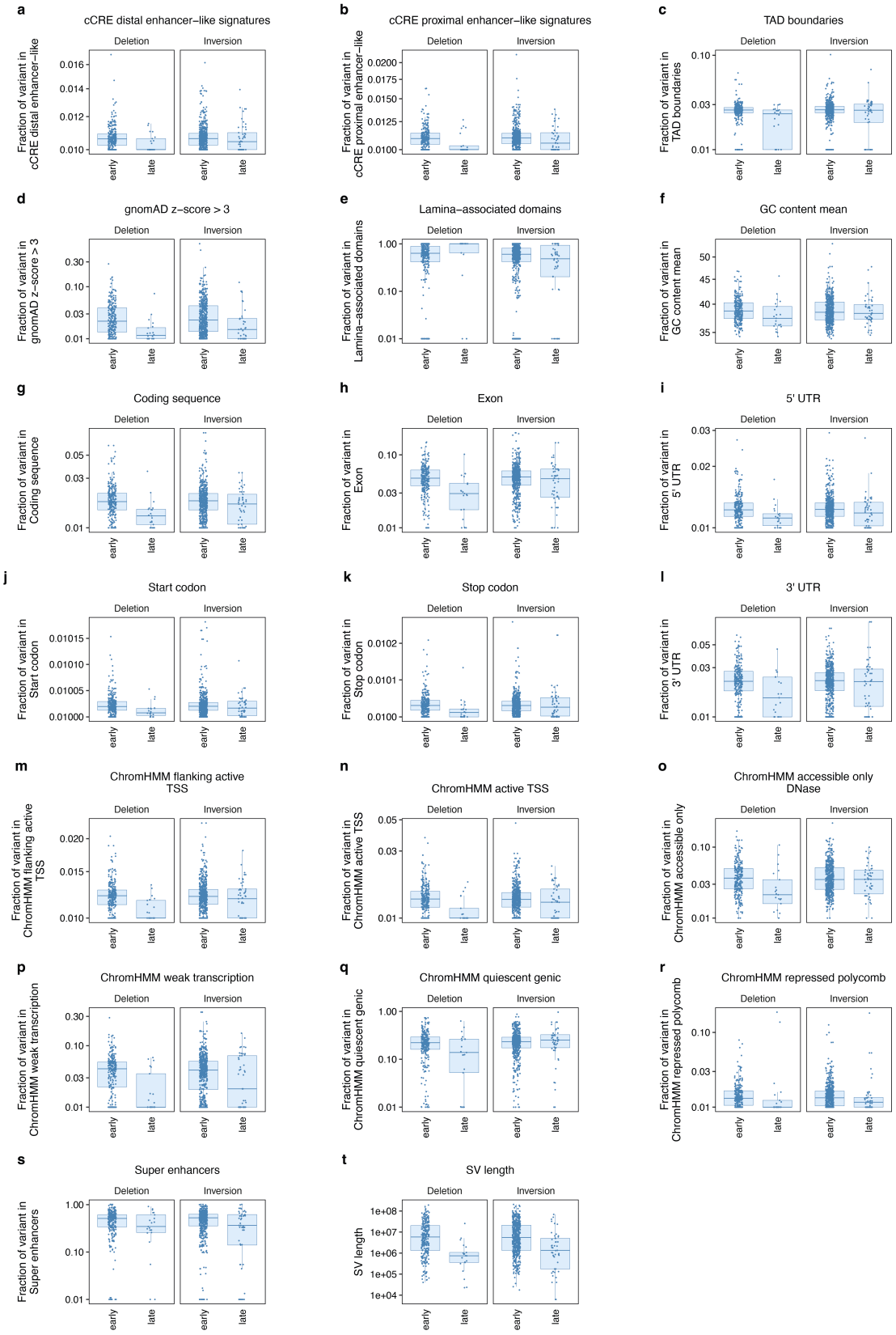

Figure 3.22 A variant-by-variant view on features of surviving variants. a-t. Feature level (y-axis) for early and late time point deletions (x-axis) in HAP1 for each significant feature (from Figure 3.21a). Markers represent individual structural variants. Box: median and quartiles. Whiskers: Largest or smallest value no further than 1.5 interquartile ranges from the hinge.

Table 3.2 Megabase-sized surviving variants in a haploid cell

Chr	Start [Mb]	End [Mb]	Variant size [Mb]	N genes affected	Mean essentiality
ChrX	129.9	155.4	25.6	231	1.29
Chr18	59.4	63.7	4.3	19	1.43
Chr18	63.7	68.8	4.1	14	1.36
Chr4	21.3	23.6	2.4	4	1.55
Chr8	34.5	36.5	1.9	2	1.63

Together, these data demonstrate that continued survival and growth in cell culture consolidated an initially mixed pool of variants towards shorter deletions and inversions that affect fewer and preferentially growth inhibiting genes, avoided enhancer-like elements, and highly constrained DNA.

3.2.4 Scrambled clones

Clones that survived the scrambling process represent a unique collection of novel human cell lines that only differ in specific structural variants. I sought to leverage this resource to understand how such variants affect gene expression and ultimately cell fitness by generating shallow-depth whole genome sequencing for 21 clones and RNA sequencing for three clones with interesting genetic architectures.

The first investigated HAP1 clone had three Cre-induced variants (Figure 3.23a), a 23 kb deletion that did not overlap with any genes, a 599 kb deletion that deleted the *EPHA7* receptor tyrosine kinase gene (Figure 3.23b), and a 40 kb inversion that affected the last three exons of *TMEM38B*, a cation channel involved in calcium homeostasis (Figure 3.23c). *EPHA7* and *TMEM38B* were indeed the top and 21st most downregulated genes (5000 and 24-fold respectively, Figure 3.23d). For the inversion and deletions, none of the genes within 3 Mb of the structural variant were significantly dysregulated (Figure 3.23e,f). To understand if the three variants collectively affecting 662 kb of DNA would affect the fitness of the clone, I competed it against parental cells and observed no difference in growth speed (Figure 3.23a, Clone 1).

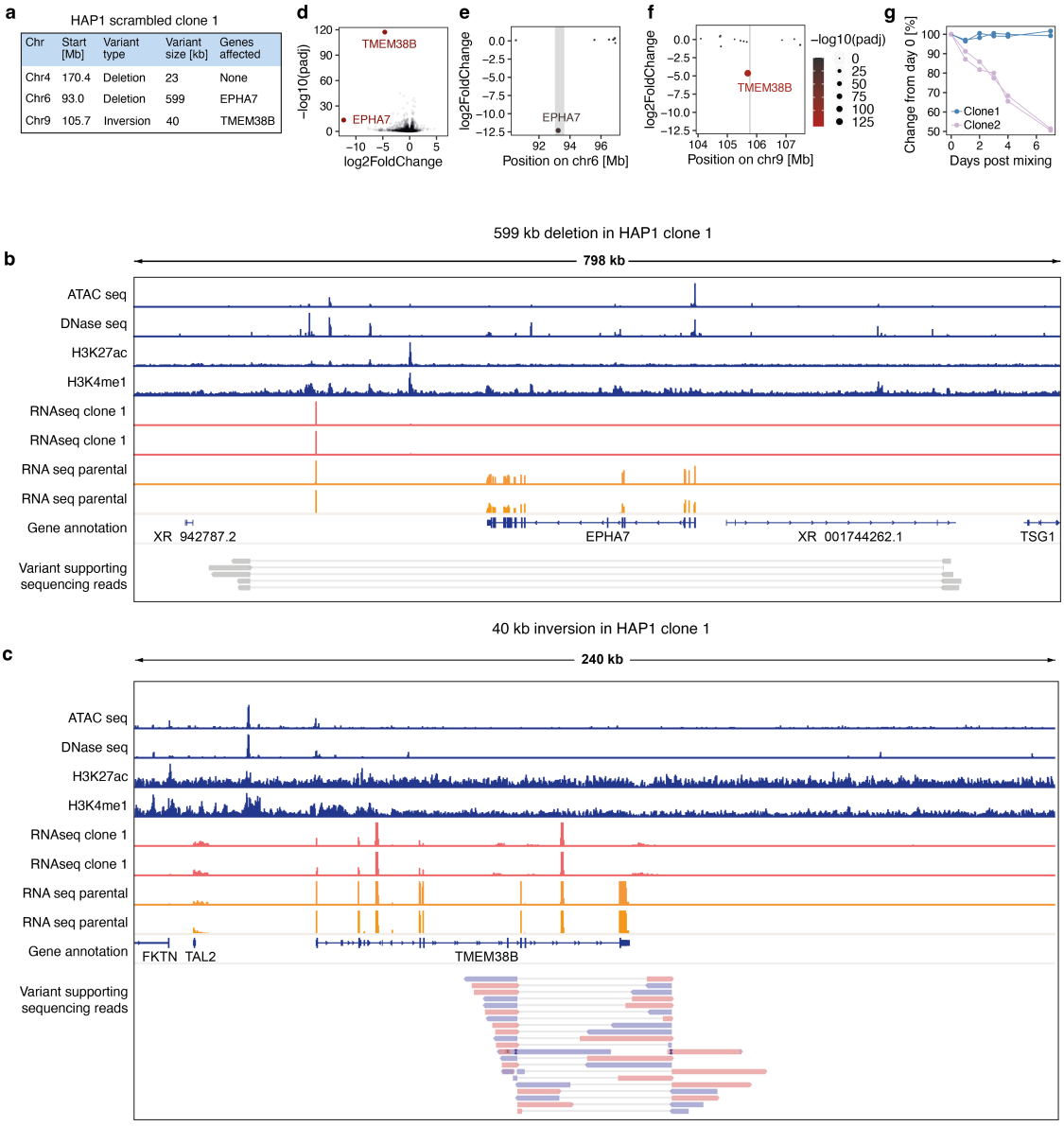

Figure 3.23 Genotype to phenotype map of the first scrambled HAP1 clone. a. A table with variants found in the HAP1 scrambled clone 1. **b.** Integrated genomic profile displaying a 240 kb region encompassing a 40 kb inversion. Tracks from top to bottom represent ATAC-seq and DNase-seq data indicating open chromatin regions; histone modification marks H3K27ac and H3K4me1 associated with active regulatory elements; RNA-seq data from HAP1 clone 1 and parental lines; gene annotation; and variant supporting sequencing reads, with alignments depicting the inversion breakpoints. **c.** As (b) but for a 798 kb deletion in clone 1. **d.** Significance of differential gene expression (y-axis) and expression fold change (x-axis) between clone 1 and parental HAP1 cells (DEseq2) for genes with at least 10 reads (markers) from n = 2 biological replicates. **e.** Changes in gene expression (y-axis) between the scrambled clone and parentals for genes (markers) located +/- 3 Mb of a 599 kb deletion (x-axis, deletion shaded gray). Marker size and color according to significance. n = 2 biological replicates. **f.** As in (e) but around a 40 kb inversion. **g.** Changes of clone abundance post mixing with parental cells (y-axis) for two scrambled HAP1 clones (colors) with two biological replicates (lines) over 7 days (x-axis).

The second HAP1 clone I examined had one 745 kb deletion that affected two genes, *OLA1* and *SP3* which were the top and second most downregulated genes in this cell line (Figure 3.24a,b). The 10 Mb surrounding the 742 kb deletion are gene-rich and contain 68 expressed genes. However, none of the surrounding genes were significantly misregulated (Figure 2.24c). Intriguingly, the variant almost exactly excises a single TAD containing *OLA1* and *SP3* (Figure 3.24d). The regulatory elements that were within the TAD (Figure 3.24e) might not be able to influence genes outside the TAD, and consequently, their deletion would not affect neighboring gene expression. *OLA1* is weakly essential (HAP1 knockout screen fold change: 0.64) and many genes were significantly differentially expressed in this cell line (652 up, 172 down; adjusted p-value < 0.01, Fold change < 0.5 or > 2). Growth-promoting pathways (Myc target genes 1 and 2) were downregulated and the cell line had a growth defect compared to parental cells (Figure 3.23g, Clone 2).

Lastly, one HEK293T scrambled clone had a 683 kb deletion on chromosome 3 and a fold-back on the minor allele of the triploid chromosome 14 (Figure 3.25a-c). The fold-back should have resulted in a dicentric/acentric pair of isochromosomes. The dicentric derivative chromosome persisted, which raised the copy number for 24.5 Mb of the underlying sequence from three to four (Figure 3.25d) and changed the single nucleotide polymorphism B allele frequency from 0.33/0.66 to 0.5/0.5 (Figure 3.25e). The acentric derivative chromosome was lost, resulting in a decrease of copy number for 76.5 Mb of sequence from 3 to 2 and loss of heterozygosity (Figure 3.25e). The genes encoded on the affected sequences followed the expected trend (a median increase of 34% for duplicated sequences and a decrease of 27% for lost sequences). In addition, the clone lost entire copies of chromosomes 1, 2, 4, 7, 10, 11, and 20 resulting in an average karyotype of 61 chromosomes compared to 68 chromosomes of the parental cell line (Figure 3.26). These pervasive chromosome losses could have been caused by missegregation after scramble events that created dicentric-acentric pairs. This clone highlights how scramble in cells with higher ploidy can create isochromosomes and aneuploidies, thus forming a useful resource to study how aneuploidy affects growth and gene expression and how dicentric chromosomes are resolved.

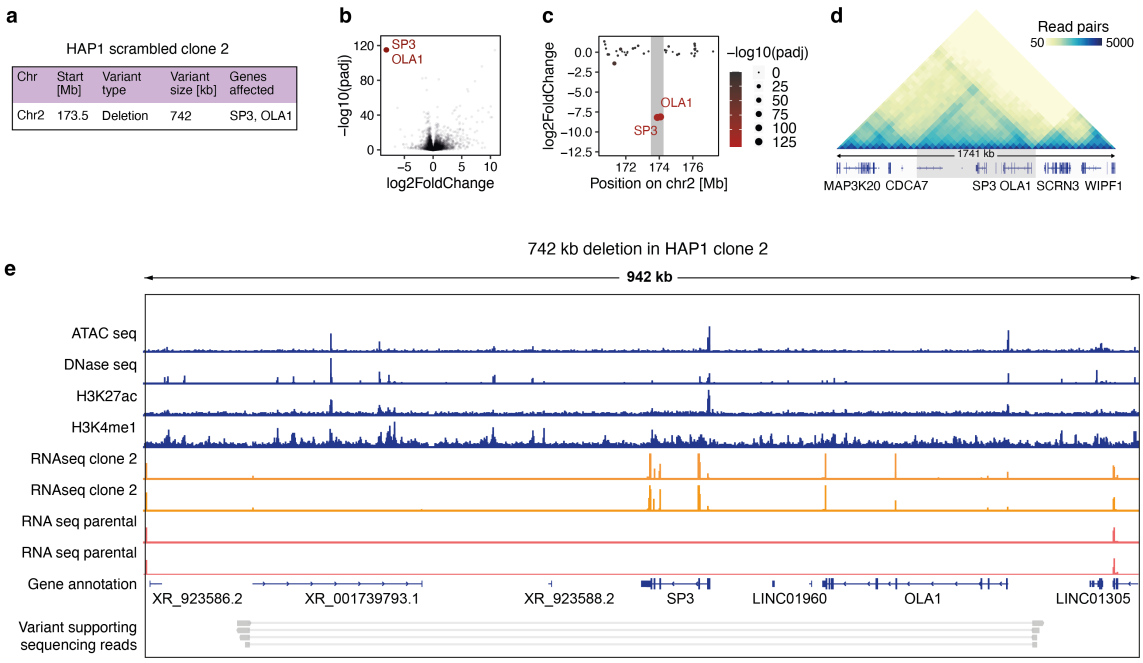

Figure 3.24 Genotype to phenotype map of the second scrambled HAP1 clone. a. Table of variants in HAP1 clone 2. **b.** Significance of differential gene expression (y-axis) and expression fold change (x-axis) between clone 2 and parental HAP1 cells (DEseq2) for genes with at least 10 reads (markers) from n = 2 biological replicates. **c.** Changes in gene expression (y-axis) between the scrambled clone and parentals for genes (markers) located +/- 3 Mb of a 742 kb deletion (x-axis, deletion shaded gray). Marker size and color according to significance. n = 2 biological replicates. **d.** Triangle heatmap showing the frequency of chromatin interaction read pairs (color gradient) within a 1741 kb segment (deletion shaded gray) and gene annotations below the heatmap. **e.** Integrated genomic profile displaying a 942 kb region encompassing a 742 kb deletion. Tracks from top to bottom represent ATAC-seq and DNase-seq data indicating open chromatin regions; histone modification marks H3K27ac and H3K4me1 associated with active regulatory elements; RNA-seq data from HAP1 clone 2 and parental lines; gene annotation; and variant supporting sequencing reads, with alignments depicting the deletion breakpoints.

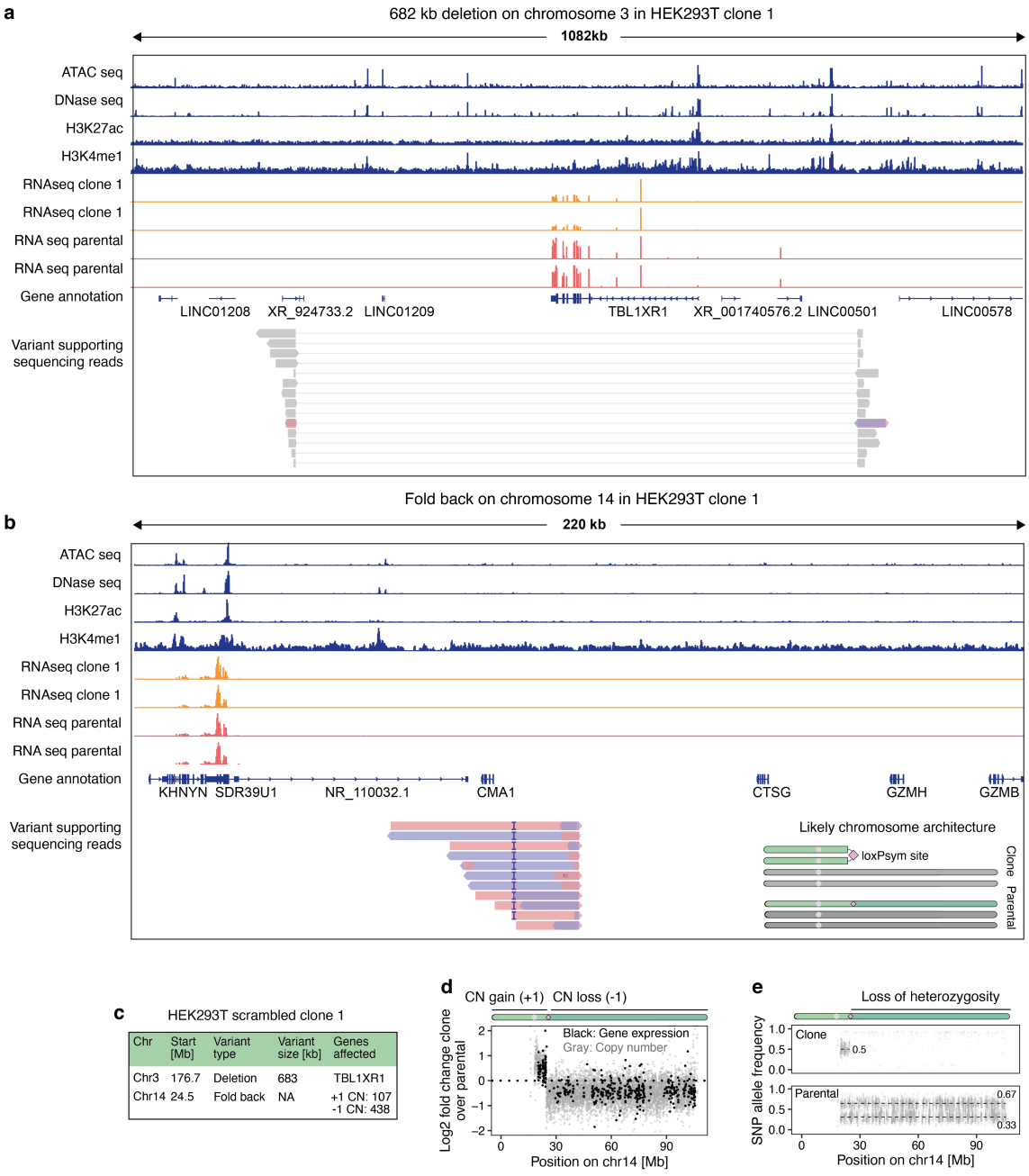

Figure 3.25. Genotype to phenotype map of a scrambled HEK293T clone with a fold-back chromosome. a. Integrated genomic profile displaying a 1082 kb region encompassing a 692 kb deletion. Tracks from top to bottom represent ATAC-seq and DNase-seq data indicating open chromatin regions; histone modification marks H3K27ac and H3K4me1 associated with active regulatory elements; RNA-seq data from HEK293T clone 1 and parental lines; gene annotation; and variant supporting sequencing reads, with alignments depicting the inversion breakpoints. **b.** As (a) but for a fold-back on chr 14. Schematic of a fold-back on chr 14 is shown in the bottom right corner. **c.** A table with variants found in the HEK293T scrambled clone 1. **d.** Log2-fold changes between clone and parental (y-axis) for mean whole-genome sequencing read depth in 50 kb windows (gray markers) or gene expression (black markers) along chromosome 14 (x-axis). n = 2 biological replicates. **e.** B allele frequencies (y-axis) for heterozygous SNPs along chr 14 (x-axis) for clone and parental (panels). Dashed lines at 0.5 and 0.33.

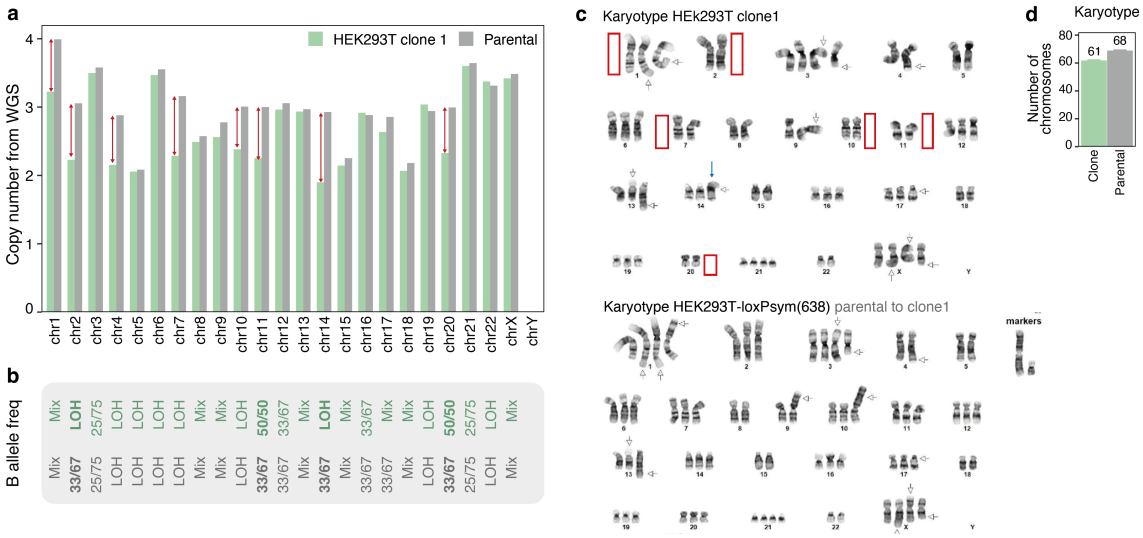

Figure 3.26 Karyotype of a scrambled HEK293T clone. a. Median copy number (y-axis, normalized to chromosomes 12 and 19 which are triploid in parental and clone) per chromosome (x-axis) for parental HEK293T-loxPsym(638) cells and a scrambled clone (colors). Red arrows indicate chromosomes with copy number differences > 0.5. **b.** Most common B allele frequencies (BAF) for the scrambled clone and parental per chromosome. 'Mix' for chromosomes where no single BAF made up at least 50% of the chromosome. LOH: Loss of heterozygosity. **c.** Representative G-band karyotype of parental and the scrambled clone (n = 23 metaphase spreads). Red boxes indicate chromosomes that are absent in the scrambled clone and arrows indicate translocations (the blue one points to the abnormal chromosome 14 in the scrambled clone). **d.** Number of chromosomes (y-axis) assessed by karyotyping of 24 metaphase spreads of parental or clone1 cells (x-axis). Bars and numbers indicate average, error bars the standard error of mean.

3.3 Discussion

In this chapter, I leveraged prime editing of repetitive elements to engineer the incorporation of thousands of Cre-recombinase sites into the human genome, creating a substrate to induce thousands of distinct recombinations at will. The variants affect all chromosomes and encompass deletions, inversions, ecDNAs, fold-backs, and translocations. I found that several megabase-scale deletions are viable in haploid cells for growth in culture. More broadly, I defined the characteristics of tolerated structural variation by comparing sequence, gene annotation, regulatory, mutational constraint, and conservation features between generated and surviving variants. Finally, I characterized clones with several large Cre-induced rearrangements, including an isochromosome, and found that variants strongly affected gene expression when they modified copy numbers but did not influence the expression of nearby genes.

Prime editing is currently the most versatile genome editing technology because it enables direct search and replace operations on DNA. Here, I used it to insert thousands of sequences into the same genome, obtaining engineered cell lines with the highest number of novel sequence insertions that I am aware of. The integrated recombinase sites made it possible to scramble a mammalian genome for the first time, a starting point of forays into the vast space of potential genomes for biotechnology and basic research. Manipulation of genomes at this scale was previously only achievable through *de novo* genome synthesis in a genome 0.3% the size of ours (Richardson et al. 2017; Y. Zhao et al. 2023; Schindler et al. 2023).

Our strategy to scramble the human genome to generate thousands of distinct structural variants for study gives an inroad to illuminate the extensive but understudied non-coding genome. In one such experiment, I detected more than 200 deletions and 500 inversions that together span 4 or 9 gigabases of sequence, several times the size of the entire human genome. So far, most of our knowledge on structural variation comes from heavily pre-selected germline variants in the human population (1000 Genomes Project Consortium et al. 2010; Sudmant et al. 2015; R. L. Collins et al. 2020; Abel et al. 2020; Vanderstichele et al. 2023) or somatic variants in cancers (Y. Li et al. 2020; Cosenza, Rodriguez-Martin, and Korbel 2022) that represent the small fraction of all variants that are compatible with survival. In contrast, Cre-induced rearrangements at early time points have not yet entered the selective ratchet, and the vast majority of the initially diverse landscape of genomic alterations did not persist over two weeks in culture. Still, the remaining deletions and inversions together affected over 3% and 12% of the genome, respectively - more than the protein-coding sequences combined. The advantage of the strategy presented here is that the persisting variants can be directly compared against the baseline of all generated ones to illuminate the selection pressures acting on the genome. For example, I saw that the remaining variants were enriched in quiescent and heterochromatic regions and avoided coding regions as well as gene regulatory elements.

Alongside deletions, inversions, translocations, and fold-backs, Cre recombination also created hundreds of ecDNAs. Most of these ecDNAs lack centromeres and cannot be actively separated upon cell division. Accordingly, they did not persist in the context of the transformed human cell lines used as models here. Yet, ecDNAs are common and associated with poorer prognosis in human cancers where their uneven distribution across daughter cells can boost copy numbers into the hundreds, amplifying underlying oncogenes such as *EGFR*, *MDM2*, and *MYC* (H. Kim et al. 2020; Yi et al. 2022). A recent study used a similar Cre-lox system to engineer ecDNAs in primary cells and demonstrated that ecDNAs containing *MDM2* would accumulate over several weeks and help the primary cells overcome p53-dependent senescence (Pradella et al. 2023). Genome-

wide scramble in more primary cell types could identify regions that are positively selected for when amplified as ecDNA, and shed light on the genesis, propagation, and biology of ecDNAs.

I engineered, scrambled, and characterized the effects of structural variants and aneuploidies on the survival and gene expression of two cell lines. While insights from these cell lines will be highly valuable for bioproduction (HEK293T) and basic research (HAP1), both are transformed cells with abnormal karyotypes. Regulatory sequences are highly context-specific (Boix et al. 2021) and sequences that can be deleted in one cell line might be essential in another context. To generalize towards more contexts, my collaborator Raphael Feirrera from the Church lab (Harvard Medical School) attempted to expand the method to stem cells (fibroblasts and iPSCs). While the transformed cells described in this chapter were able to tolerate highly multiplexed prime editing, non-cancerous cells exhibited substantial cytotoxicity. This suggests that non-cancerous cells may possess distinctive DNA repair mechanisms or heightened stress responses upon extensive nicking, consistent with previous work showing that catalytically inactive Cas9 is required for successful editing in iPSCs using the same sgRNA (Smith et al. 2020). Notably, HAP1 and HEK293T cells harbor multiple mutations known to inhibit apoptosis and promote cell survival (Carette et al. 2011; DuBridge et al. 1987) and future endeavors in stem and primary cells would require strategies to prevent apoptosis in response to extensive genomic nicking. Nevertheless, the HAP1 model system has been instrumental for the classification of pathogenicity in protein variants (Radford et al. 2022; Buckley et al. 2023; Findlay et al. 2018) and findings on the properties of tolerated structural changes should also generalize more broadly.

The size, number, and diversity of variants that can be created is a function of the number of integrated loxPsym sequences. In the HAP1-loxPsym(301) clone, recombinase sites are separated by a median of 7 Mb, expected to contain over 40 genes, several of which are essential. This is in contrast to the synthetic yeast chromosomes where loxPsym sites are only separated by a few hundred base pairs (J. S. Dymond et al. 2011). Several improvements to prime editing and pegRNA design have been published in the meantime (Doman et al. 2023; Mathis, Allam, Kissling, et al. 2023; G. Yu et al. 2023; Mathis, Allam, Tálas, et al. 2023) which should all make it easier to insert a higher number of sequences with prime editing of repetitive elements. Here I focused on LINE-1 retrotransposons, but other repetitive elements (Alu sequences, endogenous retrovirus, and various types of microsatellites) could also be targeted (Zou et al. 2022; Niu et al. 2017). An approach incorporating multiple types of repetitive elements, each with a unique genomic distribution, could enhance the number of rearrangement anchors and achieve a more diverse landscape of rearrangements.

Unlike evolution through point mutations, structural changes generated by scrambling can simultaneously affect tens to hundreds of genes and megabases of non-coding regions. By exploring a much larger mutational space genome scrambling enables a more comprehensive assessment of a phenotypic landscape. I envision that these properties open up exciting possibilities in two major applications.

First, scrambling can generate novel cell lines with evolved properties. By coupling scrambling as a source of sequence diversity to phenotypic selection, cellular properties could be optimized and the resulting evolutionary paths analyzed for a better understanding of genotype-phenotype landscapes. These types of experiments have already been successful in yeast strains with synthetic chromosomes (W. Liu et al. 2018; Gowers et al. 2020; Kang et al. 2022; Ma et al. 2019). In mammalian cells, scrambling could shed light on mechanisms of drug resistance, and growth under adverse conditions (e.g. in minimal medium), or help in biomanufacturing. For example, the HEK293 derivative cell lines are used to produce the ChAdOx1 nCoV-19 vaccine (Michalik et al. 2022), as well as a wide range of proteins (Tan et al. 2021).

Second, the random generation of large deletions opens up the exciting possibility of assaying the essentiality of sequences genome-wide. With an increasing diversity and density of recombinase site positions across a population of cells, the generation of 10,000s of deletions spanning the genome many times over becomes feasible. By measuring variant dropout, genome-wide maps of essentiality could be built. Such essentiality maps, like the maps being generated from saturation genome editing studies (Fowler et al. 2023), will be invaluable for the interpretation of pathogenic variants and provide a deeper understanding of genome structure and function.

3.4 Methods

Cell lines

HEK293T cells were acquired from AMS Biotechnology (AMS.EP-CL-0005) and the HAP1 Δ*MLH1* cell line was purchased from Horizon Discovery (HZGHC000343c022). HEK293T cells were cultured in DMEM (Invitrogen) and HAP1 cells in IMDM (Invitrogen), both supplemented with 10% FCS (Invitrogen), 2 mM glutamine (Invitrogen), 100 U/ml penicillin and 100 mg/ml streptomycin (Invitrogen) at 37 °C and 5% CO2.

Compounds

Palbociclib (S1116 Selleck), Paclitaxel (Cambridge Bioscience), Doxycycline (Selleckchem), Pifithrin-alpha (Merck Life Science UK Limited, P4359-5MG) were dissolved in DMSO to generate stock solutions of 10 mM. Pyromycine (ThermoFisher) was dissolved in water to a concentration of 10 mM and used at 2 µg/ml.

pegRNA cloning

Regular pegRNAs for transient transfection were cloned using golden gate assembly as described in (Anzalone et al. 2019). Briefly, for each pegRNA, forward and reverse oligonucleotides were ordered for the spacer, scaffold, and 3'-extensions (Integrated DNA Technologies or Merck). The pegRNA acceptor plasmid (Addgene #132777) was linearized with BsaI, oligonucleotides hybridized, and the scaffold phosphorylated. The components were assembled using a golden gate reaction (with BsaI and T4 ligase) and transformed into XL10 gold ultracompetent bacteria (Agilent). Correct constructs were confirmed by Sanger sequencing. Lentiviral engineered pegRNA plasmids were cloned the same way but using a lentiviral epegRNA acceptor construct as the starting point (Table 3.3) which was linearized with BsmBI instead of BsaI. In addition, an optimized scaffold was used (cr772, gtttaagagctaagctggaaacagcatagcaagtttaaataaggctagtccgtt atcaactcgaaagagtggcaccgagtcggtg).

Inserting loxP sites into LINE-1 retrotransposons by transient transfection

One day before transfection, 500,000 HEK293T cells were plated into each well of a six-well plate and two cell pellets with 2 million cells each were frozen for RNA extraction. 2 µg of PE2 plasmid (Table 3.3) and 500 ng of pegRNA plasmid (Table 3.3) were transfected per well using Lipofectamine LTX (Invitrogen, using 2.5 µl plus reagent and 7.5 µl LTX solution). After 3 days, cells from 3 wells were frozen for RNA extraction. The remaining 3 wells were split 1:5 and replated in a 6-well plate. The remaining cells from the split were frozen for amplicon sequencing. The splits and cell pellet freezings were repeated on days 7, 9, and 11.

Table 3.3. Plasmids used in this chapter

Name	Description	Benchling link	Reference
pPEG-LINE-1-loxP	Insertion of loxP sequence into LINE-1s by tansient transfection	https://benchling.com/s/seq-yDBpUYN5dw8wM0ueqCMO?m=slm-xWV9ro7ehnq2cLCuSaE9	Chapter 3
pCMV-PE2	Prime editor 2 plasmid for transient transfection	https://benchling.com/s/seq-hgB6PmTgl7x0mhNbRgUG?m=slm-WbcOCQNuzi3g7MwLfu4Y	Anzalone et al., Nature, 2019
pPB-TREG3G-PE2-rtTA3G-P2A-eGFP	Piggybac vector with doxycyclin-inducible prime editor	https://benchling.com/s/seq-rCcJG0pk2TUvOSVljikI?m=slm-2LxVK7M5LvfREDcBRfgX	Chapter 2
pCMV-hyPBase	PiggyBac transpoase	Not available	Yusa et al, Proc Natl Acad Sci USA, 2011
pLenti-PEG-HEK3-loxP	Lentiviral guide RNA vector to insert a loxP sequence into the *HEK3* locus	https://benchling.com/s/seq-hcqrsiKZ655IuGHrVd8Q?m=slm-hGVxYWhbjf6QFKFLYWaO	Chapter 2
pLenti-ePEG-LINE-1-loxPsym	Lentiviral guide RNA vector to insert the loxPsym sequence into LINE-1.	https://benchling.com/s/seq-ZxDgLLb1ixt6phIbSLSC?m=slm-PZgJaReU5l3K5Z8jh37e	Chapter 3
pLent-ePEG-acceptor	Acceptor vector for lentiviral epegRNA cloning	https://benchling.com/s/seq-hMFGqUh7AKlqmrRfEwAb?m=slm-XVvkjqPa52w1U7sYZB9E	Chapter 3

Generating HAP1 Δ*MLH1* and HEK293T cell lines that stably express prime editor

HAP1 and HEK293T cell lines expressing prime editors were generated by cotransfecting pCMV-hyPBase (Yusa et al. 2011) and pPB-TREG3G-PE2-rtTA3G-P2A-eGFP (Table 3.3) (Koeppel et al. 2023). First, 500,000 HAP1 Δ*MLH1* and HEK293T cells were each seeded into one well of a six-well plate one day before transfection. For each transfection, 3 µg of each plasmid was mixed with 6 µl of Plus reagent and 7.5 µl of Lipofectamine LTX (Invitrogen) reagent, incubated for 30 min, and then added to the cells. At two weeks post-transfection, cells were sorted into single clones based on eGFP expression.

Table 3.4. Oligonucleotides used in this chapter

ID	Name	Sequence	Function
P1	HEK3_F	ATGTGGGCTGCCTAGAAAGG	Amplification of the HEK3 locus to assess editing rate of prime editing clones
P2	HEK3_R	CCCAGCCAAACTTGTCAACC	Amplification of the HEK3 locus to assess editing rate of prime editing clones
P3	LINE-1_F_S1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAAAGAGTCCAGGACCAGATGGAT	Amplification sequencing of LINE-1s to assess loxPsym insertion rates (1 nt stagger)
P4	LINE-1_F_S7	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGACGACAAAGAGTCCAGGACCAGATGGAT	Amplification sequencing of LINE-1s to assess loxPsym insertion rates (7 nt stagger)

P5	LINE-1_R	GAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTCCCGGCTTTGGTATCAGAATG	Amplification sequencing of LINE-1s to assess loxPsym insertion rates
P6	PE1.0	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T	Indexing primer for NGS
P7	Rev primer	CAAGCAGAAGACGGCATACGAGATN10GAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC T	Indexing primer for NGS

Lentivirus production

Lentivirus was produced as outlined in section 2.4 (page 58).

Flow cytometry

Samples were run on the CytoFLEX Flow Cytometer (Beckman) for analysis. The data was acquired with the CytExpert software and analyzed with FlowJo V10 or CytoExploreR. Events were first gated for cells based on forward and side scatter. Next, singlets were distinguished from doublets based on the width and height of the side scatter light. Finally, cells were analyzed for their respective fluorescence channels. Sensitivity was set so that the mean fluorescence intensity of the negative population was around 10^1 - 10^3. Cells were sorted for single clones on either Sony MA900, Sony SH800S, or MoFlo XDP. I had assistance from the flow cytometry core facility for sorts on MoFlo XDP.

LINE-1 editing time course

500,000 HAP1 and HEK293T cells stably expressing PiggyBac-integrated prime editor and pLenti-ePEG-LINE-1-loxPsym were seeded into 6 well plates and editing was induced with 1 µM doxycyclin in the presence of 10 µM PFT-alpha. The cells were split every 2 or 4 days and reseeded at approximately 30% confluency. The remaining cells were harvested for sgRNA extraction and analysis. Cell culture work for this experiment was done by Gareth Girling, a technical specialist from our laboratory.

Amplicon sequencing of LINE-1s

Genomic DNA was either extracted from cell pellets using the DNeasy Blood & Tissue kit (Qiagen) with the addition of 50 µg RNaseA in the pellet resuspension step or cell pellets were prepared by direct lysis using home-made quick extract buffer (1 mM $CaCl_2$, 3 mM $MgCl_2$, 1 mM EDTA, 1% Triton X-100, 10 mM Tris pH 7.5) with freshly added proteinase K (0.2 mg/ml) followed by 15 min incubation at 65°C and 20 min incubation at 95°C. For library preparation, 3 µl of genomic DNA (~ 30 ng) or 3 µl of direct lysis extract were used as templates in 50 µl PCR reactions using KAPA HiFi HotStart ReadyMix (Roche), primers P3/4 and P5 (Table 3.4). This

first PCR was run for 16-19 cycles (3min 95°C, 18x(20sec 98°C, 15sec 66°C, 30sec 72°C), 5min 72°C) and then purified with Agencourt AMPure XP beads in 1:1 ratio (beads to PCR reaction volume). Sequencing adaptors and barcodes were added with a second round of PCR using the KAPA HiFi HotStart ReadyMix (Roche), primers P6 and P7 (Table 3.4), and 1 µl of the purified first PCR product as template. Amplicons were purified with Agencourt AMPure XP beads in 1:1 ratio (beads to PCR reaction volume) and quantified using the Nanodrop 2000. The amplicons were pooled together and sequenced on the Illumina MiSeq 2500 (500 cycles, 250 paired-end).

Scramble time course with TAT-Cre protein

Without Cre reporter: HAP1 cells with 301 monoclonal loxPsym site integrations (HAP1-loxPsym(301)) were plated in 6 well plates with 1.5 million cells per well in 1.5 ml of IMDM medium supplemented with FBS. 5 µM or 2 µM of TAT-Cre protein (Cambridge Biochemistry Department) was added to the cells. The TAT-Cre-containing medium was removed 4 hours after induction and replaced with 2 ml of fresh medium. Cells were trypsinized and seeded into T150 flasks after 16 hours and harvested and split at the indicated time points.

With Cre reporter: HAP1 cells with 301 monoclonal and HEK293T cells with 638 monoallelic loxPsym site integrations and expressing the tBFP-Cre reporter construct (HAP1-loxPsym(301)R and HEK293T-loxPsym-638R) were plated in 6 well plates with 1.5 million cells per well in 1.5 ml of IMDM medium without FBS. 1 µM TAT-Cre protein was added to the cells 30 minutes after seeding and removed 4 hours later. 2 ml of fresh medium containing FBS was added to the cells. The cells were trypsinized and seeded into T150 flasks after 16 hours. The cells were harvested and split at the indicated time points and the fraction of BFP-positive cells was monitored by flow cytometry. HEK293T-loxPsym-638R and HAP1-loxPsym(301)R cells were sorted for BFP-positive cells at day 13 and day 15 respectively.

Cell cycle inhibition

500,000 HAP1-loxPsym(301) cells were plated in each well of a 6-well plate. One day later, 3 wells were treated with 500 nM palbociclib for 24 hours. To confirm cell cycle inhibition, one well of cells was harvested, stained for DNA (Hoechst-33342), and analyzed using flow cytometry. For the remaining wells, the medium was replaced with IMDM without FBS. After 1 hour, 1 µM TAT-Cre protein was added to the cells and removed 4 hours later. Palbociclib was kept in the media throughout the process. To remove extracellular Cre, cells were incubated in TrypLE Express (Theromfisher) for 10 minutes. Fresh media containing Palbociclib was added to the cells. 16 hours later, the cells were trypsinized. 2 million cells were harvested for long-read sequencing and the rest were plated in larger flasks and grown in full media without Palbociclib to restart the cell cycle.

High molecular weight DNA extraction

Between 2 and 4 million cells were harvested by centrifugation and high molecular weight (HMW) genomic DNA was extracted using the Monarch High Molecular Weight DNA Extraction kit (New England BioLabs) and agitation speeds of 1,500-2,000 rpm.

Long read whole-genome sequencing

HMW DNA was sheared to 20 kb using Covaris G-Tubes (for shallow-depth sequencing with MinION) or using the Megaruptor 3 (Diagenode; for high-depth sequencing with PromethION). Nanopore sequencing libraries were prepared using the Ligation Sequencing Kit V14 (Oxford Nanopore) and sequenced on the MinION Mk1B using R10.4.1 flow cells (FLO-MIN114) or on the PromethION. Base calling was performed using the 'Super high accuracy model' (dna_r10.4.1_e8.2_400bps_sup.cfg) of the guppy basecaller (versions 6.3.8 or 6.4.6).

Identification of insertion sites for loxPsym sites

Long-read whole genome sequencing fastq files were aligned to the hg38 reference genome using minimap2 (H. Li 2016, 2018). The resulting sam files were sorted and compressed into bam files using samtools (H. Li et al. 2009). Structural variants were called with relaxed settings using Sniffles2 (Smolka et al. 2023) with the --non-germline --phase --output-rnames --tandem-repeats --minsupport 1 --minsvlen 25 parameters. Custom R scripts were used to identify loxPsym insertion sites and call Cre-induced rearrangements. For the identification of insertion sites, R::agrep with the max.distance = 0.2 option was used to find variants that contain 'TAACTTCGTATAATGTACATTATACGAAGTTA' in the ALT sequence and subsequently filtered for insertions < 50 nt. Clonal insertions in HAP1 were defined as sites with >5 supporting reads and an allele frequency of > 0.5. Clonal insertions in HEK293T were defined as sites with at least 3 supporting reads and an allele frequency of > 0.1. Insertion sites are deposited as Data S1.

Identification of Cre-induced variants

Raw structural variants were called using nanomonsv (Shiraishi et al. 2023). The unscrambled parentals were considered as control and the scrambled cell lines as tumor. One supporting read was required for the tumor and 0 for the control and a panel of normals (Shiraishi et al. 2023). The raw structural variants were filtered to (1) contain fragments of a loxPsym site at the break junction ("TAACTTCGTAT | ATACGAAGTTA | AATGTACATTAT | GTATAATGTAC") (2) Start and end within 300 bp of a nicking site of the LINE-1 targeting pegRNA (allowing 3 mismatches to the protospacer). (3) Do not have coverages higher than 5x the average sequencing

depth (regions with higher depth are usually ambiguously mapped). (4) Are located on chromosomes 1-22 + X, Y. A schematic of the variant calling process is shown in Figure 3.11.

Further genome analysis.

Coverages were estimated using samtools bedcov, considering reads with at least quality 10 and binning in 10 kb windows. For plotting, depth was calculated using mosdepth (0.3.3 (Pedersen and Quinlan 2018)) and 50 kb binning windows. B allele frequencies were calculated using amber in tumor-only mode (Cameron et al. 2019). The 4DNFI1E6NJQJ HiC data set was used for HAP1 cells and visualized using plotgardner (Kramer et al. 2022).

Simulations of contact frequency and 3D distance

A matrix of all possible recombinations was generated by simulating rearrangements for all pairs of loxPsym sites on each chromosome. In the naive model, each variant was equally likely. The observed rearrangements were then compared to all possible ones. For the exponential decay model, the simulated rearrangements were adjusted to decrease in frequency of occurrence with increased distance (between 20% and 0.1% per Mb in 0.1% increments). The distributions resulting from simulations were compared to the observed data and their Wasserstein distance was calculated. Pearson correlations between the number of sequencing reads (0 if no rearrangement between a given pair was observed) and the logarithms of variant length or Hi-C read pair numbers (4D nucleome, 4DNFI1E6NJQJ) were estimated using a linear model.

Cas9-enrichment long-read sequencing of edited LINE-1s

Cas9-enrichment was performed according to the Cas9 sequencing kit protocol (CAS9106 Protocolv109, Oxford Nanopore Technologies). Briefly, HMW DNA was sheared to 20 kb using Covaris G-Tubes, and 5 µg was dephosphorylated. ALT-R tracrRNA and crRNA targeting LINE-1s with a loxPsym site insertion (ACATTATACGAAGTTATAGG) were ordered from Integrated DNA Technologies (IDT) and complexed with HifiCas9 V3 (Integrated DNA Technologies) to form RNPs. Dephosphorylated DNA was treated with Cas9 RNPs for 1 hour at 37˚C. Sequencing adapters were ligated to the cut DNA. The ligation step was extended to 1 hour at room temperature. The libraries were then purified using Ampure XP beads (Agilent), washed with Long fragment buffer, and eluted in elution buffer for 1 hour. The libraries were sequenced with the MinION Mk1B using R9.4.1 flow cells (FLO-MIN106). Base calling was performed using the 'Super high accuracy model' (dna_r10.4.1_e8.2_400bps_sup.cfg) of the guppy basecaller (versions 6.3.8 or 6.4.6). The resulting read files were filtered with seqkit to only contain reads that cover LINE-1s by matching the following sequence ("AGGAGGAACTGGTACCATTCCTTCTGAAACTATT") while allowing up to 3

mismatches. The filtered read file was aligned to the hg38 reference genome (Dec. 2013 GRCh38/hg38) using minimap2. The number of reads covering each LINE-1s with up to two mismatches from the protospacer (Data S3) was then determined using bedtools multicov (Quinlan and Hall 2010) requiring a mapping quality score of at least 5.

Chromatin states and epigenetic analyses

The following publicly available datasets were collected: HEK293T: DNase-seq ENCFF969MBJ, H3K4me1-ChIP-seq SRR10981645, H3K36me3-ChIP-seq SRR5627148, H3K9me3-ChIP-seq SRR11453034, H3K27me3-ChIP-seq SRR8937480, H3K4me3-ChIP-seq SRR8937479, and H3K27ac-ChIP-seq SRR1016003. HAP1: DNase-seq ENCFF162WTC, H3K4me1-ChIP-seq ENCFF639UYT, H3K36me3-ChIP-seq ENCFF216JJJ, H3K9me3-ChIP-seq ENCFF528UHF, H3K27me3-ChIP-seq ENCFF708HAB, H3K4me3-ChIP-seq ENCFF461TZF, and H3K27ac-ChIP-seq ENCFF742SZS (Table 3.1) (ENCODE Project Consortium 2012; J. Zhang et al. 2020). For data sets from ENCODE, hg38-aligned bam files were downloaded and for data sets from the sequence read archive, fastq files were downloaded and aligned to the hg38 genome build using bwa-mem. The resulting bam files were binarized using ChromHMM and a 15-state model was learned (Figure 3.9). In addition, the GSM4625025 for CTCF-ChIP-seq in HAP1 and ENCSR000DTW for CTCF-ChIP-seq in HEK293T data sets were visualized in IGV.

For the comparison of epigenetic states in edited (> 1 read in the Cas9 sequencing experiment) and non-edited LINE-1s (0 reads in the Cas9 sequencing experiment), the fraction of nucleotides in each ChromHMM state was calculated across a sequence window consisting of the LINE-1s and 3 kb of flanking sequence (to mitigate possible misalignment of short read data in repetitive LINE-1s). States that were found in less than 10 edited LINE-1s were filtered out. Enrichments (risk ratios) were calculated by dividing the fraction in each state across edited and non-edited LINE-1s.

RNA sequencing

RNA was extracted from 2-5 million flash-frozen cells using the RNeasy plus mini kit (Qiagen). Libraries were prepared using the New England BioLabsNext® Ultra™ II Directional RNA Library Prep Kit for Illumina (New Englands Biolabs), multiplexed, and sequenced on two Illumina-HTP Novaseq 6000 lanes using 150 bp paired end reads. The median insert size was 280 bp (quartiles 231, 355). Between 58 and 191 million reads were generated per sample. Salmon (2.0.0 (Patro et al. 2017)) was used to quantify transcripts against a salmon index built for the hg38 human reference genome. Further analysis was done using custom R scripts. Transcripts were collapsed to gene level using tximport (1.22.0) (Data S5). Normalization (rlog) and differential expression analysis was performed using DESeq2 (1.34.0). The ensembl v110 release

was used for gene structure annotation (imported into R using biomaRt (2.50.3). For all analyses involving genes, the gene lists were filtered to contain only "protein_coding", "lncRNA", "snRNA", "snoRNA", "miRNA", "rRNA", "ribozyme" genes with expression base means > 20. Gene set enrichment analysis was done using fgsea (1.20.0) and the h.all.v2022.1.Hs.symbols.gmt gene set (https://data.broadinstitute.org/gsea-msigdb/msigdb/release/2022.1.Hs/)

Features of structural variants

To understand the selection effects acting on structural variants Thomas Vanderstichele, a PhD student from the Davenport lab, conducted an enrichment analysis across 47 features comparing variants that I observed at early and late time points. PhyloP conservation scores were downloaded from the University of Santa Cruz genome browser (Pollard et al. 2010) and each structural variant was annotated with the mean, median, and max across all overlapping bases. The fraction of conserved elements overlapping each structural variant was computed using GERP++ conserved elements (Davydov et al. 2010). The fraction of each structural variant overlapping human accelerated regions was computed using annotations from Girskis et al. (Girskis et al. 2021). Constraint z-scores passing all quality control checks for coding and non-coding regions were downloaded from gnomAD (S. Chen et al. 2022). Structural variants were annotated with the maximum, median, and mean gnomAD z-score across all overlapping 1 kb windows as well as the fraction of windows with a z-score greater than 2, 3, and 4. Gene features (exon, 5' UTR, 3' UTR, stop codon, start codon, CDS, and gene) were computed as the fraction overlapping each structural variant using GENCODE release v44 (Harrow et al. 2012). The fraction of each structural variant overlapping super-enhancers was calculated using annotations from SEdb 2.0 (Y. Wang et al. 2023). The fractional overlap of each structural variant over TAD boundaries, A/B compartments, and lamina-associated domains from HAP1 cells was calculated using data from the 4D nucleome project (Sanborn et al. 2015; Reiff et al. 2022; van Schaik et al. 2020). The fraction of the structural variant overlapping CpG islands was computed using annotations from the University of Santa Cruz genome browser (Gardiner-Garden and Frommer 1987). The mean GC content across each SV was calculated using data from the UCSC genome browser.

Candidate cis-regulatory elements in HAP1 cells were downloaded from ENCODE and for each variant, the fractional overlap over each type of element was computed (ENCODE Project Consortium et al. 2020). The fractional overlap over chromatin states was computed using the HAP1 chromHMM annotation described previously. Finally, the length of the structural variant was included as a feature. Thomas tested for enrichment for deletions and inversions separately, normalizing each feature within each variant class. Thomas performed logistic regression modeling whether the structural variant was observed at an early or late time point as a function

of each feature individually. 95% confidence intervals for the fitted parameters were calculated using the standard normal distribution.

HAP1 essentiality screen

This screen was performed by Elin Madi Peets. The screening procedures are described in more detail in (Peets et al. 2019). The screen in HAP1 cells is briefly described here. 20M Hap1-Cas+ cells were transduced in two biological replicates using a spinfection in 6 well plates with lentivirus aiming for a multiplicity of infection of 0.3. The spinfection was carried out at room temperature for 30 minutes spinning at 1000g and the cells were supplemented with 8 µg/ml polybrene (hexadimethrine bromide, Sigma). 2 µg/ml puromycin was added to the cells 24 hours post-infection. 2 µg/ml puromycin was kept in the media throughout the screen. Cells were grown for 14 days and time points were taken at days 3, 7, 10, and 14 post-infection. For each time point, at least 24 million cells were harvested. To derive essentiality scores, the fold changes of guide abundance in sequencing reads between day 14 and the plasmid library were calculated for six independent guides per gene and collapsed to gene level by taking the average. Finally, the two biological replicates were merged by taking the average gene-level fold change (Data S6).

Isolation of scrambled clones

13 days after scrambling (see section Scramble time course with TAT-Cre protein), Cre reporter expressing cells were single-cell sorted for tBFP expression (indicating successful Cre recombination) and haploidy (based on cell size in forward and side scatter) into 96-well plates containing 200 µl of growth medium per well. Cells were transferred to flasks after visible colonies were formed. High molecular weight DNA was extracted as above. Clones were sequenced at 3-5x whole genome coverage either on the minion or on the PromethION after pooling the DNA from various clones. I had help from Gareth Girling, a technical specialist in our lab, for culturing of clones.

Competition assay between scrambled clones and parental cells

Scrambled clones (tBFP-positive from the rearranged Cre reporter) were mixed with parental cells that do not express the Cre-reporter (tBFP-negative) in a 50:50 ratio and grown for 7 days. The cell ratio of parental to scrambled cells was determined on the indicated days by flow cytometry for tBFP expression.

G-band karyotyping

Complex G-band karyotyping was performed and analyzed by Karyologic, Inc. from samples of cryopreserved cells.

Software

Genomics: bwa (0.7.17), Benchling (accessed between 2019-2023), bedtools (v2.29.0), ChromHMM (1.24), guppy (6.3.8 and 6.4.6), minimap2 (2.22), mosdepth (0.3.3), sniffles2 (2.0.7), samtools (1.14), salmon, seqkit (v.2.0.0), IGV (2.16.2), R (4.1.3), nanomonsv (0.6.0).

Flow cytometry: FlowJo (v10), CytoExploreR (1.1.0). CytExpert.

R packages: biomaRt (2.50.3), DESeq2 (1.34.0), fgsea (1.20.0), ggrepel (0.9.3), ggpointdensity (0.1.0), plotgardnerer (1.4.2), plyranges (1.14.0), Repitools (1.40.0), ShortRead (1.46.0), spgs (1.0-3), StructalVariantAnnotation (1.10.1), tidyverse (1.3.2), VariantAnnotation (1.40.0), viridis (0.6.2), tximport (1.22.0)

4

Randomizing gene regulatory regions using prime editing

The genome scramble strategy is powerful and unique because we can generate thousands of structural variants in one experiment that collectively cover gigabases of sequence. However, the events are random, distributed across the entire genome, and do not have enough granularity to gain a detailed understanding of individual regulatory regions. We can use the same basic ingredients of prime editing to insert multiple recombinase recognition sequences, recombinases to shuffle sequences, and long-read sequencing to understand the resulting variants to devise a complementary strategy to illuminate regulatory regions at high granularity.

Enhancer clusters recruit transcription factors and activate genes from afar. However, it is not well understood how individual enhancers within a cluster interact and how their spacing and relative orientations drive gene expression. To answer these questions, I used prime editing to tile an enhancer cluster in the *OTX2* homeobox gene with symmetrical loxP sites. Induction of Cre recombinase in the engineered cell lines resulted in stochastic inversions and deletions within the enhancer cluster. I then used Oxford nanopore long-read sequencing with Cas9-based enrichment and methyl-transferase treatment of chromatin to simultaneously map new architectures of the enhancer cluster and detect their associated methylation levels and chromatin accessibility. In what follows, I discovered and characterized novel enhancer architectures with differential *OTX2* expression. These findings demonstrate the feasibility of highly efficient, multiplexed prime insertions, and will shed light on the grammar of gene expression for regulatory elements.

I had help from Mélanie Gouley and Valentin Reberning, two excellent master's and bachelor's students whom I supervised. I generated the majority of data, analyzed all the data, and made all the plots in this chapter; Mélanie and Valentin designed and cloned constructs for five pegRNAs encoding loxPsym insertions; Mélanie helped with the generation and characterization of the prime editing HAP1 Δ*MLH1* cell line and performed a proof of concept nanopore sequencing experiment (data is not included here); Valentin helped with the isolation of clones with two and three loxPsym insertions, clones with deletions after scrambling of cell lines with three loxPsym sites, and *OTX2* qPCRs. All these experiments were done under my direct supervision.

4.1 Introduction

To ensure genes are expressed in the appropriate dose at the right time and place, a diverse set of regulatory elements act in concert to modulate the expression of their target genes. Mutations that disrupt this process can cause disease and developmental defects (Banerji, Olson, and Schaffner 1983; Gillies et al. 1983; Kioussis et al. 1983; Lettice et al. 2003). The potential of individual regulatory elements to drive gene expression can be measured accurately and at a massive scale (Melnikov et al. 2012). However, it is unclear how elements interact with each other, and how their embedding in the 3D genome, spacing from each other, and relative orientations come together to drive gene expression.

This lack of understanding is rooted in the strategies that are currently used to study gene regulatory regions. MPRAs are highly scalable but do not capture the endogenous context (Melnikov et al. 2012; Patwardhan et al. 2012). Genomic manipulation of regulatory elements retains the context, but the types of edits that are possible are usually limited to deleting, inactivating, or activating elements (Lopes, Korkmaz, and Agami 2016). Thus, genetic manipulation offers limited clues to questions about the relevance of orientation, distance to the target gene, or interactions between elements. Synthetic reconstitution attempts to design and synthesize many variations of a large gene locus, bring them into cells, and subsequently test their ability to recapitulate the behavior of the endogenous locus (Pinglay et al. 2022; Brosh et al. 2023; Ordoñez et al. 2023; Blayney et al. 2023). Reconstitution is a great strategy to dissect interactions and context of elements but building loci from scratch is still difficult and requires infrastructure and expertise for a series of model organisms (yeasts, bacteria, mammalian cells).

Ideally, a strategy would create hundreds of regulatory architectures of a locus right in the endogenous context (Figure 4.1). The combination of prime editing, recombinases, and scalable phenotyping should fit this requirement. Prime editing is a genome engineering technology that enables the precise insertions of short sequences (such as the loxPsym site) at programmable target sites without requiring an external DNA donor (Anzalone et al. 2019, 2022; Yarnall et al. 2022; Koeppel et al. 2023). Once multiple loxPsym sites are integrated into a regulatory region using prime editing, Cre recombinase induction will stochastically delete and invert DNA sequence between any possible pair of sites.

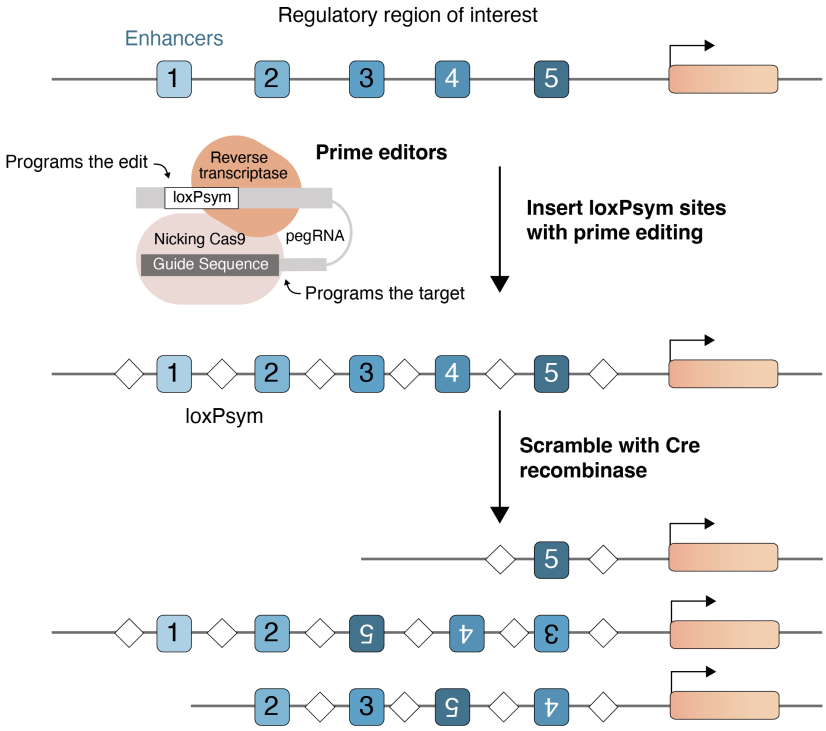

Figure 4.1 Combining prime editing with Cre recombinase to scramble regulatory regions. Schematic of a regulatory region of interest with five enhancers (blue boxes) and a gene of interest (brown box). Prime editors (schematic) can be used to insert loxPsym sites (white diamonds) and Cre recombinase will create diverse enhancer architectures.

Long-read sequencing of native DNA molecules can be used to simultaneously map the stochastic rearrangements generated by Cre as well as CpG methylation of single DNA molecules (Simpson et al. 2017; Rand et al. 2017) without the bias of PCR amplification. An additional layer of information on chromatin accessibility from the same DNA molecules can be gained by first treating the nuclei with a bacterial N^6-adenine methyltransferase (m6A) which preferentially methylates adenines within accessible chromatin (Stergachis et al. 2020; I. Lee et al. 2020). Since human DNA is almost completely devoid of m6A modifications, it can be used as a proxy for chromatin accessibility.

Sequencing larger loci of interest (>10 kb) while preserving full-length sequencing reads and native DNA modifications remains a major challenge. Nanopore Cas9-targeted sequencing enables this at moderate throughput (Gilpatrick et al. 2020). Here, genomic DNA is first dephosphorylated and incubated with Cas9 and cr::trRNAs to cut out the site of interest *in vitro*, releasing DNA ends that are accessible for sequencing adaptor ligation. Combining Cas9 enrichment with single-molecule DNA modification calling makes it possible to reconstruct the novel regulatory architecture and its associated level of DNA accessibility and CpG methylation, all from single molecules in pools of cells harboring diverse genetic variants.

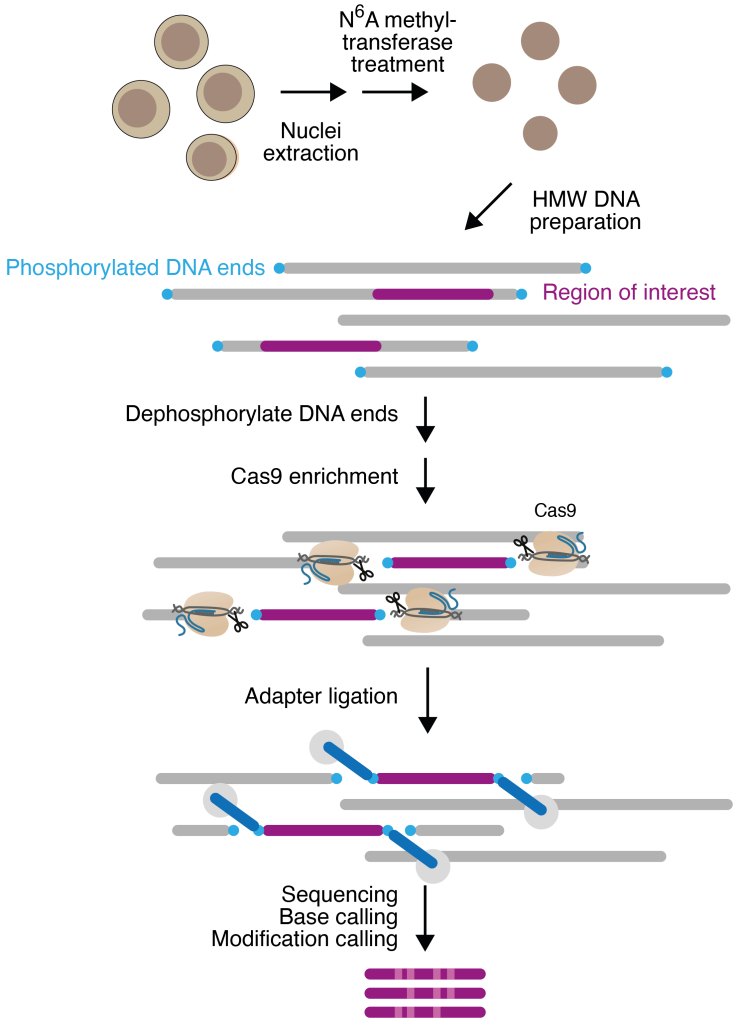

Figure 4.2 Targeted sequencing. Schematic of a Cas9-enrichment protocol that combines accessibility information. First nuclei are extracted and treated with a N⁶A methyltransferase. High molecular weight DNA is extracted, dephosphorylated, and cut with Cas9. Only freshly cut sites will be phosphorylated and can ligate with sequencing adaptors.

Here, I tiled a 70 kb regulatory region of the *OTX2* gene with six loxPsym sites and created tens of novel architectures, which I could link to levels of *OTX2* expression. For three deletions, I could further distinguish differences in CpG methylation and accessibility. Mapping rearrangements to associated expression changes highlighted three enhancers in an enhancer cluster which contributed to 50% of *OTX2* expression. Deleting remaining enhancers while simultaneously moving the cluster closer to the transcription start site was compatible with strong *OTX2* expression.

4.2 Results

The *OTX2* homeobox transcription factor combined features that made it a good target for regulatory randomization. (1) It is well expressed in haploid HAP1 cells (TPM = 151, Figure 4.3a) but not essential. (2) *OTX2* expression varies widely across cell lines (Barretina et al. 2012) (Figure 4.3b) and tissues (Lonsdale et al. 2013) (High expression in the cerebellum, Figure 4.3c), suggesting regulatory complexity. (3) *OTX2* is a developmental transcription factor that needs to be turned on at the right place and time to ensure faithful forebrain development (Beby and Lamonerie 2013), and its loss is embryonically lethal in mice (Acampora et al. 1995), further suggesting its functional importance and tight regulation. (4) *OTX2* does not share a TAD with any other protein-coding genes (Figure 4.4a), and the nearest protein-coding genes are 149 kb (5', *TMEM260*) or 384 kb (3', *EXOC5*) away (Table 4.1). (5) The enhancers are clustered, making it possible to read out scrambled architectures using single, long sequencing reads (discussed below).

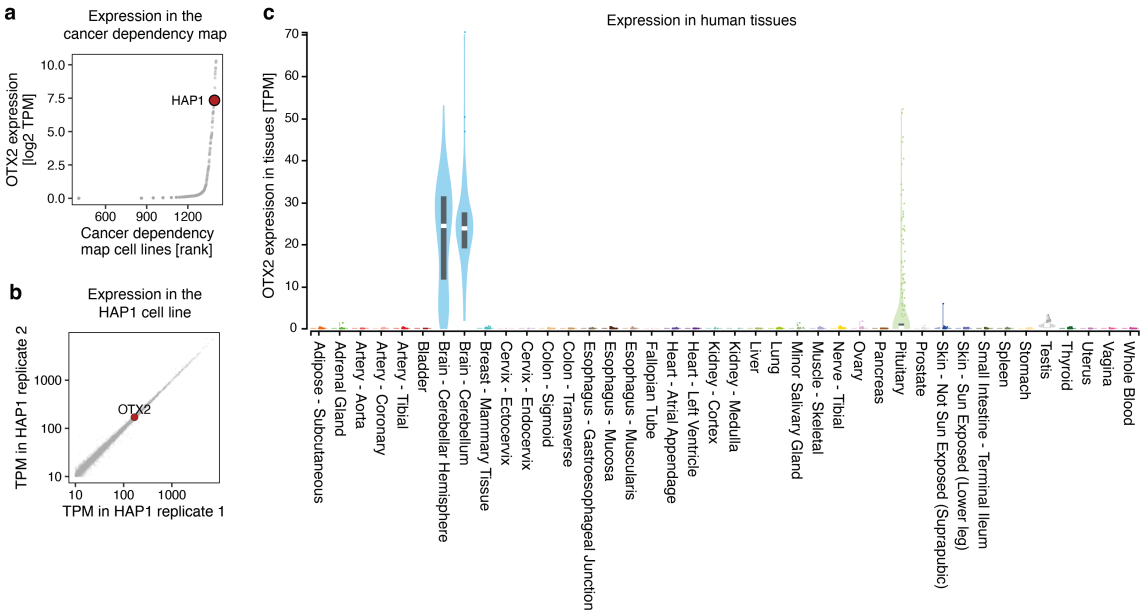

Figure 4.3. *OTX2* expression is highly variable across cell lines and tissues. a. *OTX2* expression (x-axis) in 1406 cancer cell lines (y-axis, sorted by rank). Gray markers represent cell lines with the HAP1 cell line indicated in red. **b.** Gene expression in two replicates of the prime editing HAP1 *ΔMLH1* cell line (x- and y-axes). Markers represent genes with *OTX2* indicated in red. **c.** *OTX2* expression (y-axis) in human tissues (Genotype-Tissue Expression project, axis), colored by tissue type. Density plots show the distribution of *OTX2* expression, and boxplots indicate the median and 25th and 75th percentiles.

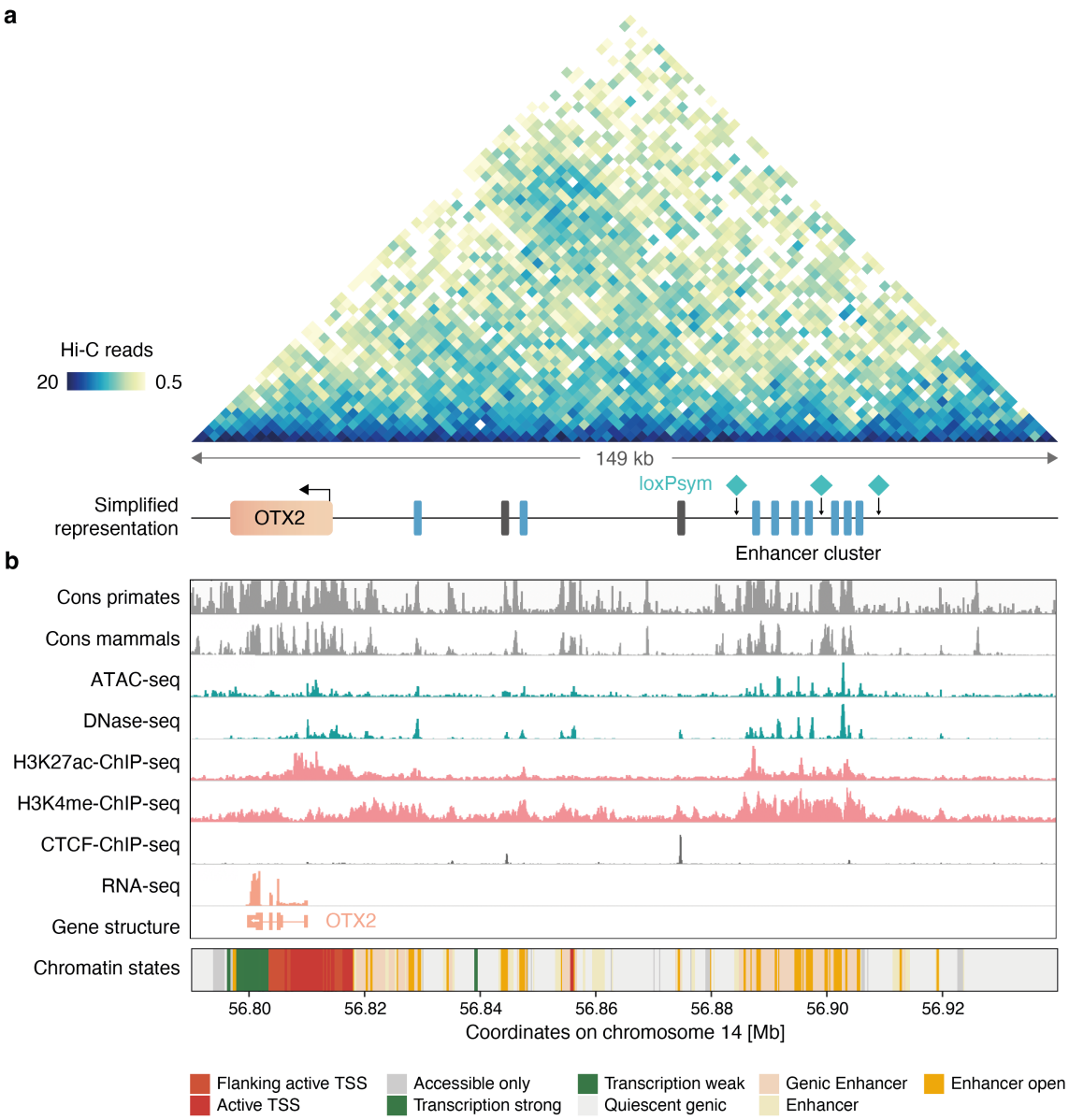

Figure 4.4. A complex set of regulatory elements controls *OTX2* expression in HAP1 cells. **a.** Top: Triangle heatmap showing the frequency of chromatin interaction read pairs (color gradient) within a 149 kb segment encompassing OTX2 and several regulatory elements (bottom). **b.** An integrated genomic profile displaying the region in (a). Tracks from top to bottom represent ATAC-seq and DNase-seq data indicating open chromatin regions; histone modification marks H3K27ac and H3K4me1 associated with active regulatory elements; ChIP-seq data from CTCF; RNA-seq data from prime editing HAP1 Δ*MLH1* cells; Gene structure; Chromatin states (colors) obtained from ChromHMM.

To map the regulatory architecture of *OTX2*, I aligned publicly available HAP1 chromatin accessibility, histone modification, and CCCTC-binding factor (CTCF) binding data sets (Table 3.1) to the *OTX2* locus as well as constraint scores from phylogenetic analysis of 240 placental mammals (Christmas et al. 2023). To condense the information, I trained a genome-

wide chromatin state model with the chromatin accessibility and modification data (Ernst and Kellis 2017) (Figure 3.9). Together these data indicate that in HAP1 cells, several regulatory elements cluster together in a 20 kb region 69 kb from the *OTX2* transcription start site. In addition, several individual regulatory elements are scattered between the cluster and *OTX2* gene (Figure 4.4b, Table 4.1). I hypothesized that the cluster regulates the expression of *OTX2* based on physical contact as measured by Hi-C (Figure 4.2a) and chromatin state patterns in 833 biosamples (Boix et al. 2021). The *OTX2* gene, two intermediate enhancers, and the cluster were in the polycomb repressed state in the majority of biosamples, except for pluripotent and embryonic stem cells as well as retina cells where the *OTX2* gene as well as enhancers were in an active chromatin state (Figure 4.5).

Table 4.1 Positions of selected features on chromosome 14 around the *OTX2* gene

Feature	Start	End	Explanation
TMEM260	56,488,354	56,650,606	Nearest protein coding gene in 5' direction
OTX2	56,799,905	56,816,693	*OTX2* gene
Enhancer	56,827,800	56,830,200	
loxPsym insertion	56,840,403	56,840,404	Added in the second round of experiments
CTCF	56,843,800	56,845,000	
Enhancer	56,847,000	56,848,400	
loxPsym insertion	56,851,953	56,851,954	Added in the second round of experiments
loxPsym insertion	56,864,718	56,864,719	Added in the second round of experiments
CTCF	56,874,400	56,874,800	
loxPsym insertion	56,884,634	56,884,635	3' of the cluster
First enhancer in cluster	56,885,800	56,886,200	Start of a 20 kb enhancer cluster
loxPsym insertion	56,898,376	56,898,377	Within the cluster
Last enhancer in cluster	56,905,400	56,906,000	End of a 20 kb enhancer cluster
loxPsym insertion	56,907,608	56,907,609	5' of the cluster
EXOC5	57,200,507	57,268,905	Nearest protein-coding gene in 3' direction

In the table, enhancers are defined as 'Enhancer open' chromHMM states. If more than one state were separated by less than 800 bp, they were considered as one enhancer.

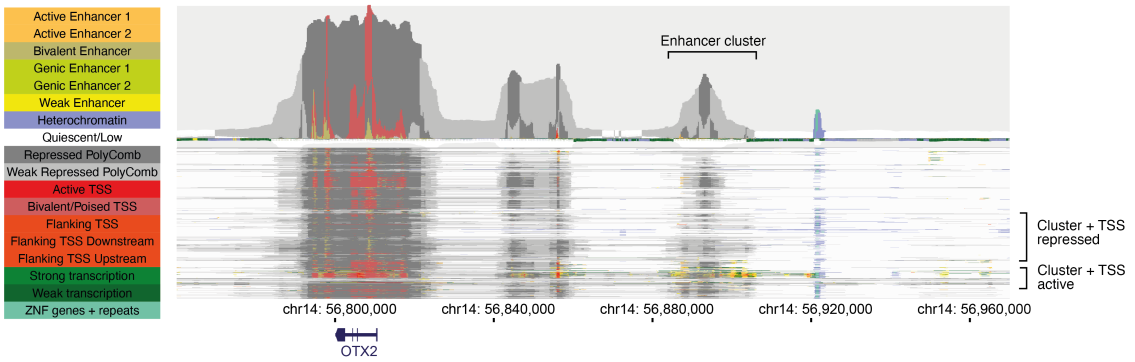

Figure 4.5. Chromatin states for the *OTX2* gene and nearby enhancers in 833 biosamples. Chromatin states (colors, according to annotation on left) in 833 biosamples (y-axis, lines) for a region of chromosome 14 (chr14:56790000-56940000, x-axis). Aggregated states are shown on the top. The *OTX2* gene as well as the positions of the enhancer cluster are annotated.

4.2.1 Engineering the *OTX2* locus

To understand how the 20 kb cluster regulates *OTX2*, I engineered three loxPsym sites and a fluorescent reporter into the same haploid HAP1 cell line with genomically integrated doxycycline-inducible prime editor that was also used for the prime insertion screen in chapter two and genome scrambling in chapter three (Figure 2.2, Figure 4.6a).

First, I chose three locations to place loxPsym sites – two at either end of the cluster and a third in the middle – avoiding peaks in ATAC-seq, DNase-seq, H3K27ac ChIP-seq and H3K4me1 ChIP-seq (Figure 4.4a). The sites were inserted by inducing prime editor expression with doxycycline, co-transfecting HAP1 cells with three pegRNAs encoding for loxPsym sequences, and measuring efficiencies by capillary electrophoresis of target site amplicons (Figure 4.6a). The 5' and internal loxPsym sites were integrated at 32% and 67% efficiency while there was no apparent integration for the 3' site. The edited population was subcloned and a cell with precise clonal integrations in both targets was derived as evidenced by Sanger sequencing (Figure 4.6b). Using a new design, I achieved 32% editing for the 3' site and isolated a cell line with clonal integration of three loxPsym sequences (Figure 4.6a,b).

Next, to map regulatory architectures to *OTX2* expressions, I tagged the endogenous *OTX2* gene with the T2A self-cleaving peptide and mScarlet. To achieve this, HAP1 cells with three clonal loxPsym sites were transfected with an HDR donor and a Cas9-gRNA ribonuclear complex. The 7.7% mScarlet-positive cells were single-cell sorted and three clones with correct and on-target T2A-mScarlet integration were isolated (Figure 4.7).

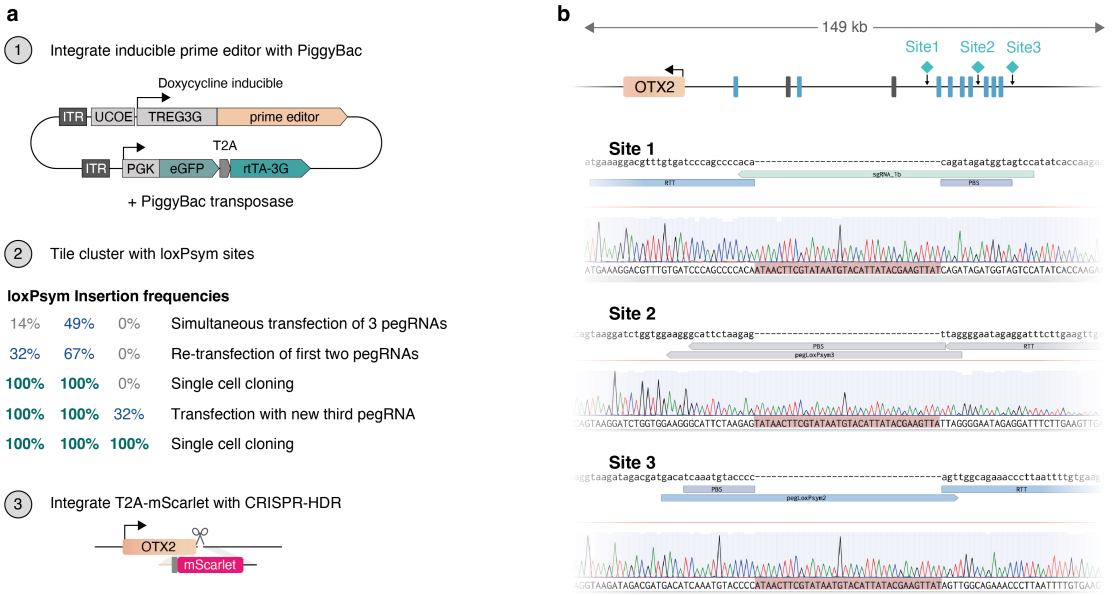

Figure 4.6. Engineering the *OTX2* locus. a. Schematic showing the steps of engineering the *OTX2* regulatory region. (1) Doxycycline-inducible prime editor construct that was integrated into HAP1 *ΔMLH1* cells using the PiggyBac transposon. (2) Tiling of the cluster with loxPsym sites. The percentages shown are estimations of insertion efficiency by capillary electrophoresis. Efficiencies deemed sufficient for subcloning are in blue text. (3) Schematic of T2A-mScarlet knock-in. **b.** Sanger sequencing traces for loxPsym site integrations. The top sequence represents the hg38 reference sequence and the second sequence is the one determined by sequencing. The relative locations of sites are shown in the schematic at the top.

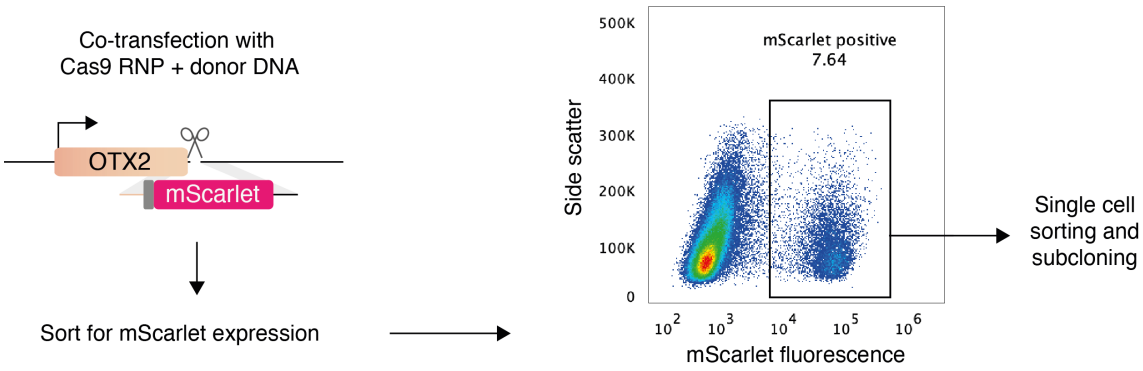

Figure 4.7. Efficient tagging of *OTX2* with a fluorophore. Schematic of the steps for deriving a cell clone where *OTX2* was tagged with *T2A-mScarlet*. Left side: Schematic of *T2A-mScarlet* knock-in. Scissors represent cutting by Cas9. Right side: Side scatter (y-axis) and mScarlet fluorescence (x-axis) for single cells as assessed by flow cytometry. Cells intersecting with the mScarlet-positive gate were single-cell sorted and subcloned.

4.2.2 Cre induction randomizes the *OTX2* enhancer cluster

With the engineered cell line in place, I could test the ability of Cre recombinase expression to induce random structural changes across the enhancer cluster after treatment with 2 µM of membrane-permeable TAT-Cre protein for 6 hours. I sequenced the enhancer architecture using Cas9-enriched nanopore sequencing after two weeks of recovery in culture (Figure 4.8a) and could detect wild-type reads, all expected simple variants (deletions and inversions of the enhancer 1-4 (E1234), enhancers 5-7 (E567), or the entire locus (E1234567), as well as complex reads that have more than one variant (e.g. a deletion and an inversion, Figure 4.8a). Full deletions of the locus were the most frequent (73% of variants). They also result in the shortest reads and could therefore be preferentially enriched by Cas9 sequencing. To test this, Valentin Rebernig derived 24 clones from the scrambled population and found 12 full deletions by junction PCR (50%, Figure 4.8c), confirming their high abundance. The wild-type architecture was the next most frequent in the nanopore reads (17%). None of the remaining variants contributed to more than 3% of all reads. This frequency distribution could be caused by the instability of intermediate states. The two remaining loxPsym sites from a partial deletion could, for example, continue to rearrange towards the full deletion. The same would be true for any inversions.

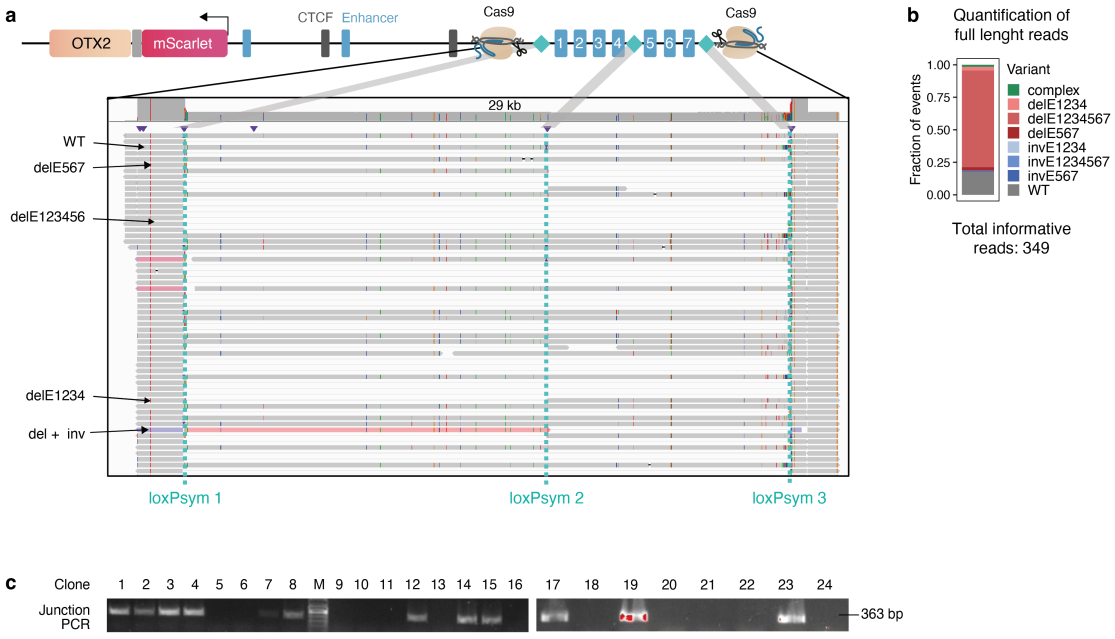

Figure 4.8. Cre recombinase induction randomizes the *OTX2* enhancer cluster. a. Sequencing reads (gray block arrows) in 29 kb of Cas9-enriched sequence aligned to the hg38 reference genome. Reads with inversions are shaded in pink and purple. Primary and supplementary alignments are connected by thin gray lines. Selected reads are annotated for variants. The positions of three loxPsym sites are indicated in teal. Relative locations are shown in the schematic at the top. **b.** Fraction of sequencing reads covering the entire locus (y-axis) corresponding to each rearrangement (colors). **c.** Gel electrophoresis image of a junction PCR that only yields a 363 bp band if the entire locus was deleted.

After two weeks in culture, the rearranged, but not untreated, cells contained populations with weaker mScarlet signal (Figure 4.9a), suggesting that some novel enhancer cluster architectures are less able to drive *OTX2* expression. To separate regulatory configurations by their enhancing potential, I sorted the mixed cell pool into four populations based on mScarlet expression (Figure 4.9b). Reassuringly, the *OTX2* expression levels remained stable across the four bins after seven additional days in culture (Figure 4.9c). I harvested the cells and used Cas9 enrichment to sequence the enhancer architectures in each sorting bin.

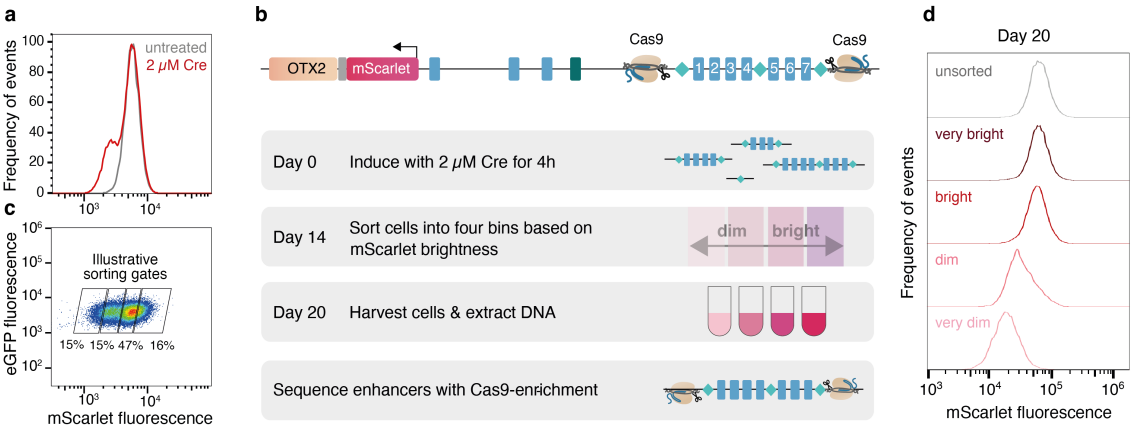

Figure 4.9. A strategy to separate enhancer variants in a pooled sorting screen. a. The frequency of events (y-axis, normalized to mode) with varying mScarlet fluorescence (x-axis) stratified by Cre treatment (colors). **b.** Schematics of the locus and steps of the screen. **c.** eGFP (y-axis) and mScarlet fluorescence (x-axis) for single cells (markers, colored by density). Illustrative sorting gates are drawn. Percentages of cells enriched in each gate are indicated. **d.** Frequency of events (y-axis, normalized by mode) with varying mScarlet fluorescence (x-axis) stratified by sorting gates (panels and colors).

Sequencing resulted in 59-605 informative reads for each sorting gate (Figure 4.10a). The types of rearrangements found across the four sorting bins differed markedly (Figure 4.10a-c). Wild-type reads were associated with higher *OTX2* expression and made up 66% of all reads in the very bright gate compared to only 0.66% in the very dim gate, consistent with the observation that scrambling the cluster reduced overall *OTX2* expression (Figure 4.9a). Conversely, deletions of the entire cluster (delE1234567) were associated with lower *OTX2* expression (91% in the very dim gate vs 6.8% in the very bright one).

The partial deletions showed an unexpected pattern. Loss of the 5' half of the cluster (delE567) reduced *OTX2* expression similarly to loss of the entire cluster while loss of the 3' half of the cluster (delE1234) behaved similarly to the wild-type sequence (Figure 4.10c). The 3' accessible

regions might facilitate the action of 5' enhancers. However, a deletion would both remove the 3' accessible regions while simultaneously moving the 5' enhancers closer to the transcription start site. The combined effect might be indistinguishable from the wild type. Inversion events, in contrast, could provide a glimpse into distance-dependent effects. Unfortunately, inversions were too rare (only 12-14 across all sorting gates) to conclusively interpret their effect. Together, the distribution of deletion variants across sorting gates demonstrated that the enhancer cluster I tiled with loxPsym sites drives *OTX2* expression predominantly through three enhancers towards the 5' end. Moreover, the results demonstrated that sorting and sequencing cell populations with diverse enhancer architectures based on the expression of a fluorescent reporter could resolve variants based on their ability to drive target gene expression.

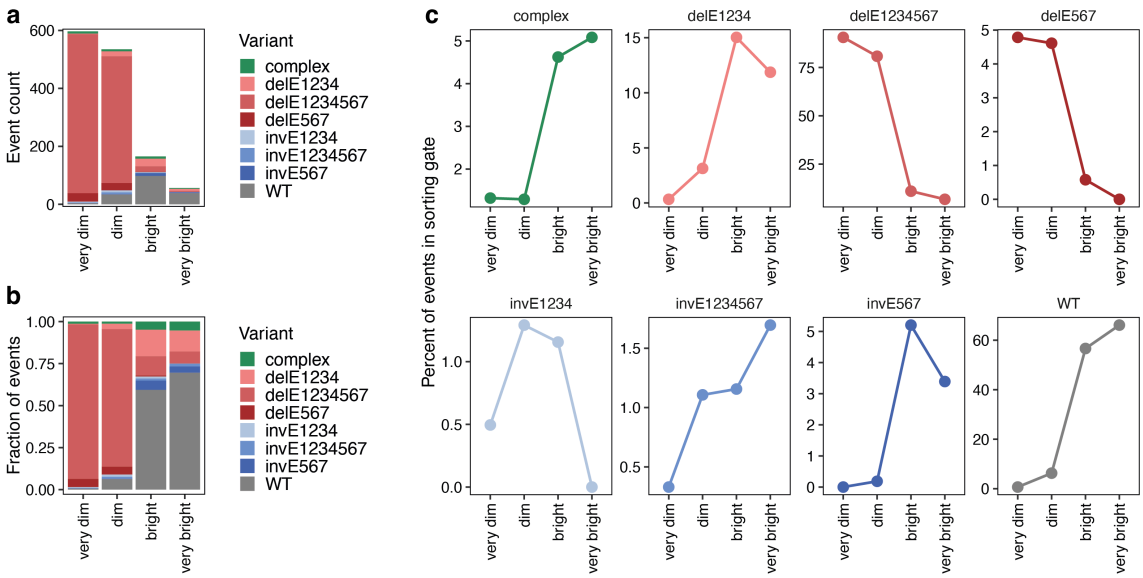

Figure 4.10. Enhancers at the 5' end of the cluster are required for high *OTX2* expression. **a.** Count (x-axis) of regulatory architectures (colors) for the four sorting gates (x-axis). **b.** As (a) but for the fraction of events. **c.** Fraction of events in sorting gate (y-axis) for the four sorting gates (x-axis) for eight regulatory architectures (panels and colors).

To measure *OTX2* expression from novel enhancer architectures, I isolated two clones with deletions of the entire cluster (Figure 4.8c) and three with deletions of enhancers 1-4, and estimated mRNA abundance by quantitative PCR (qPCR) and RNA sequencing. Deletion of the entire enhancer cluster resulted in an *OTX2* downregulation of around 50%, seen both in qPCR (Figure 4.11a) and RNA sequencing (Figure 4.11b), with no effect on the two closest genes *EXOC5* and *TMEM260* (Figure 4.11c). In contrast, the deletion of enhancers 1 to 4 did not result in *OTX2* downregulation (Figure 4.11a), in agreement with the results from population sequencing after sorting for mScarlet (Figure 4.10).

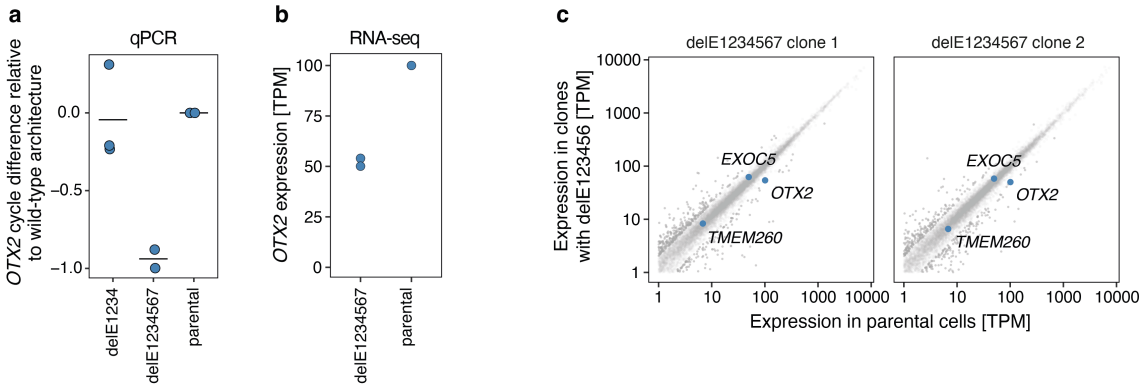

Figure 4.11 Loss of the cluster reduced *OTX2* expression by 50%. a. qPCR results for unscrambled parentals, three clones with the delE1234 variant, and two clones with the delE123456 variant (x-axis). The y-axis shows the difference in the qPCR cycle threshold until the half-maximal signal compared to the *TBD* housekeeping gene (ΔCT) normalized to the wild-type architecture. Markers are averages of two *OTX2* amplicons in a single clone. Bars represent the averages of clones. **b.** *OTX2* expression determined by RNA-seq (y-axis) for clones (markers) with different enhancer architectures (x-axis). **c.** Gene expression in clones with the E123456 deletion (y-axis) compared to expression in parental cells (x-axis). Markers represent genes and are colored more strongly if the ratio to wild-type is > 2 or < 0.5. *OTX2* and its two closest genes are indicated in blue.

4.2.3 Randomizing an expanded *OTX2* regulatory region

The remaining 50% expression could be driven by the promoter and additional enhancers outside the cluster (Figure 4.4). To understand the contributions of these elements and elucidate the consequences of changing the relative locations of regulatory elements, I inserted three additional loxPsym sequences (Table 4.1) and derived an engineered cell line with six total loxPsym insertions that tile 67 kb of non-coding DNA (Figure 4.12a). I confirmed successful locus engineering by *de-novo* assembly from Cas9-enriched nanopore sequencing reads (discussed below, Figure 4.12b).

To maximize the diversity of possible rearrangements, I titrated Cre expression. First, Wadia et al. observed that TAT-Cre protein is more effective in cell culture media without serum (Wadia, Stan, and Dowdy 2004), which I confirmed by measuring the rearrangement of a Cre activity reporter (Figure 3.17c, 4.13a). Next, I dosed cells with all six loxPsym integrations with various concentrations of TAT-Cre protein and measured mScarlet fluorescence after 14 days (Figure 4.13b). Cells treated with 500 nM of TAT-Cre protein had the most diverse expression levels, and I sorted them into four expression bins (as in Figure 4.9b) for long-read sequencing.

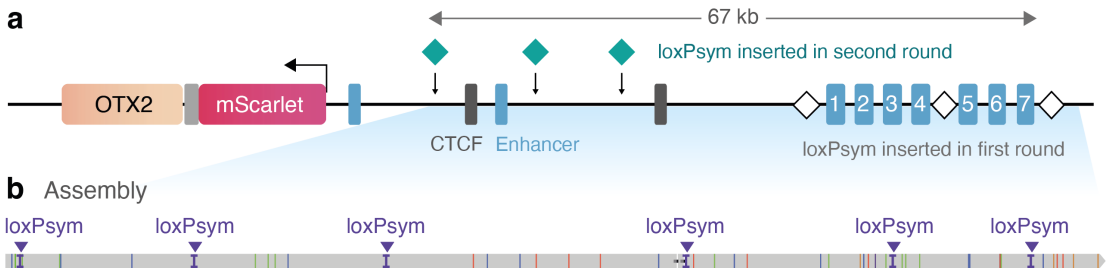

Figure 4.12. Schematic of the *OTX2* locus with six loxPsym site integrations. a. Schematic of the *OTX2* regulatory region with loxPsym sites integrated in the first round indicated with white diamonds and loxPsym sites integrated in the second round indicated in teal and with arrows. **b.** An assembly of nanopore reads from Cas9 enriched sequencing of the engineered cell lines aligned to the indicated region of the hg38 reference genome. LoxPsym insertion sites are indicated in purple. Other colorful lines indicate single-base substitutions.

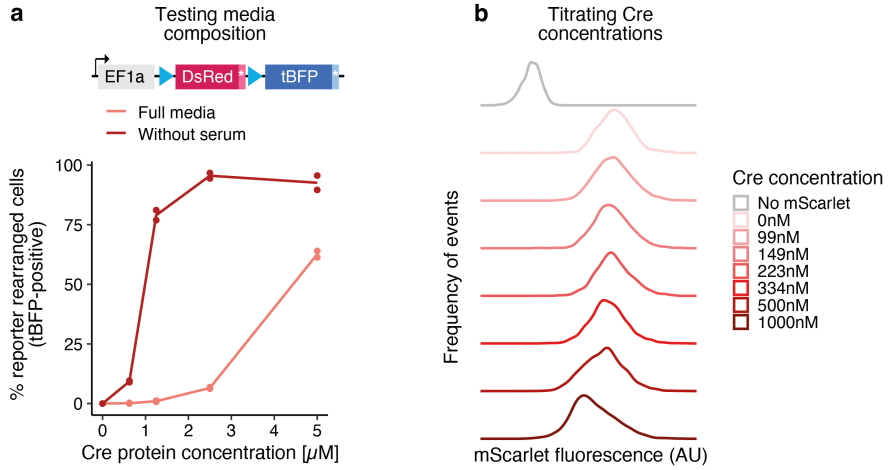

Figure 4.13. Finding optimal Cre concentrations for regulatory randomization. a. Percent of tBFP-positive cells (indicating the rearrangement of the Cre reporter, y-axis) for varying TAT-Cre protein concentrations (x-axis) for cells that were treated in full media or serum-free media (colors). **b.** Frequency of events (y-axis, normalized by mode) with varying mScarlet fluorescence (x-axis) stratified by Cre concentration (panels and colors).

With six loxPsym sites, I needed to enrich a 70 kb region with nanopore sequencing which was less efficient compared to the previous 24 kb region. Nevertheless, I obtained a maximum coverage of 52-157 reads across the four sequencing gates, enough to resolve six architectures with more than five reads across all sequencing gates (Figure 4.14a,b).

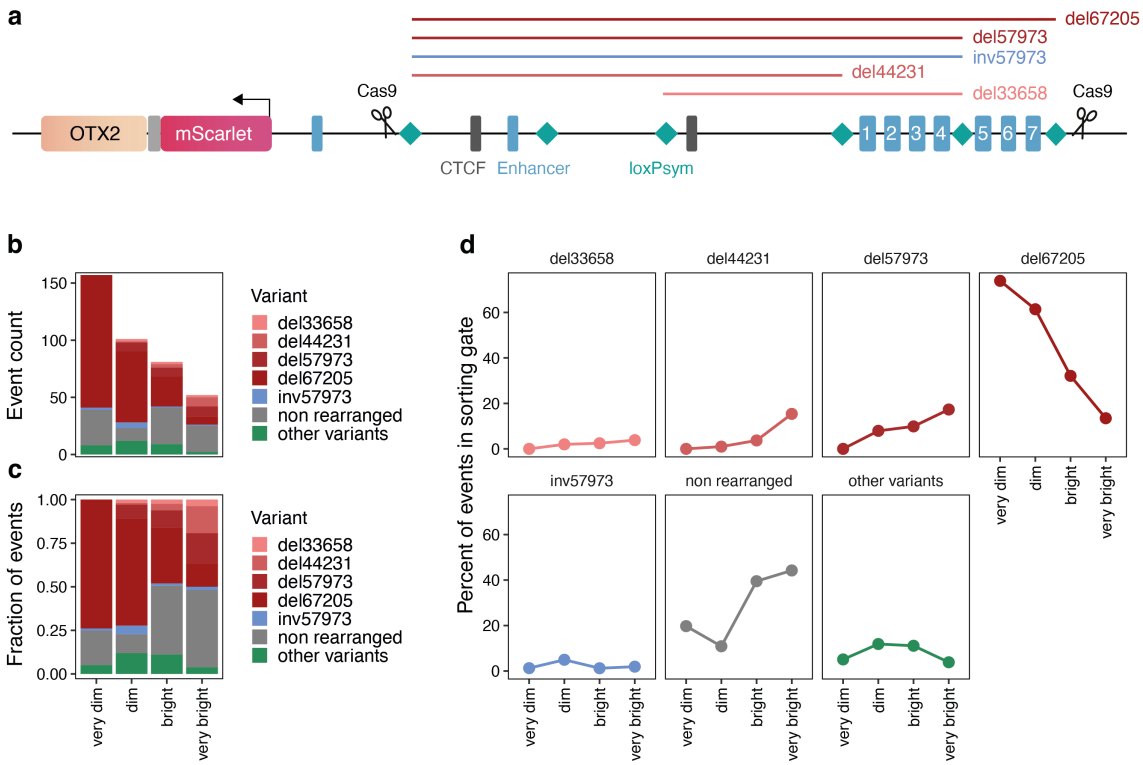

Figure 4.14. Architectures that move the enhancer cluster closer to the transcription start site are associated with high *OTX2-mScarlet* expression. a. Schematic of the *OTX2* regulatory region with selected Cre-induced variants shown as lines. Numbers correspond to variant lengths in bp. **b.** Count (x-axis) of regulatory architectures (colors) for the four sorting gates (x-axis). **c.** As (a) but for the fraction of events. **d.** Fraction of events in sorting gate (y-axis) for the four sorting gates (x-axis) for eight regulatory architectures (panels and colors).

As with the experiment with three loxPsym sequences, the deletion between the two outermost sites (del67205) was most abundant and strongly associated with low *OTX2-mScarlet* expression (74% in the very dim vs 13% in the very bright gate, Figure 4.14c,d). Three additional deletions were resolved that all moved the enhancer cluster closer to the TSS while removing the intermediate enhancers and CTCF binding sites. All three of these variants showed association with higher *OTX2* expression. This was most pronounced for a 44231 bp deletion that moves the entire cluster towards the TSS (15% in the very bright vs 0% in the very dim gate), and a 57973 bp deletion that moves the three important 5' enhancers of the cluster (E5, E6, E7, 17% in the very bright vs 0% in the very dim gate, Figure 4.14c,d).

4.2.4 DNA modification and accessibility changes

Native nanopore sequencing reads can not only resolve variants but also detect CpG methylation. In addition, I developed a quick and practical protocol inspired by (Stergachis et al. 2020) to assess DNA accessibility (proxied by m6A methylation) by treatment with a commercially available N^6-adenine methyltransferase (EcoGII, in contrast to the original protocol that purified their methyltransferases). This treatment added a single 30-minute incubation step to a high molecular weight DNA extraction protocol with a standard kit (described in more detail in section 4.4).

To understand if these additional layers of information faithfully recapitulate the molecular biology of the *OTX2* regulatory region, I combined all sequencing reads from previous experiments and computed average modification frequencies (Figure 4.12). Indeed, the accessibility profile from m6A modification closely resembled the pattern of a publicly available HAP1 DNase-sequencing data set (Figure 4.15). Moreover, CpG methylation followed an inverse pattern with high average methylation levels (70%) except for CpG dinucleotides located within accessible DNA regions (Figure 4.15).

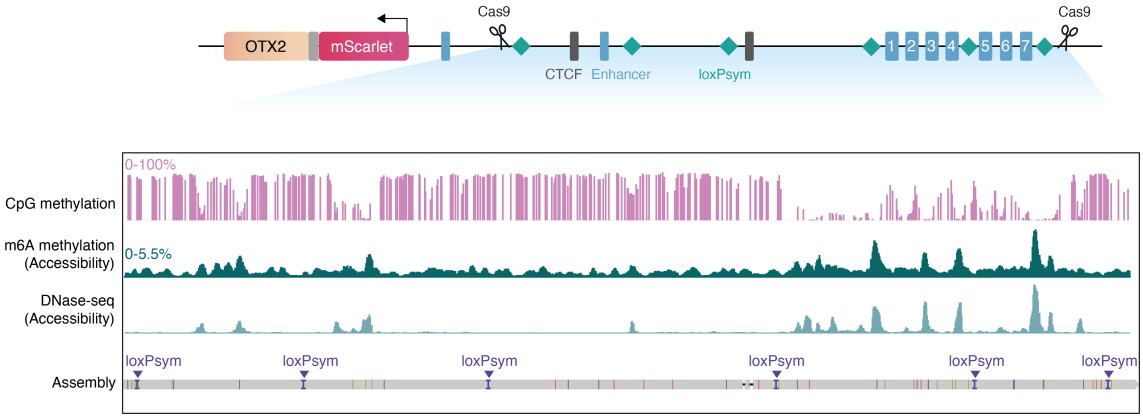

Figure 4.15. Base modification patterns add complementary information. Schematic of the *OTX2* regulatory region with the Cas9-enrichment endpoints shown as scissors and expanded in plots below. From top to bottom, all aligned to hg38: averaged percent CpG methylation from nanopore sequencing reads; Averaged m6A methylation from nanopore reads; publicly available DNase-seq data; Assembly of the engineered locus from nanopore reads. LoxPsym insertion sites are indicated in purple.

Combining variant calls with readouts of CpG methylation and accessibility should paint a fuller picture of how loss, inversion, and repositioning of enhancers affect expression levels. I first compared CpG methylation for partial deletions of the 20 kb enhancer cluster. Upon loss of the E567 area, 84/226 cytosines increased methylation by an average of +6.0% while only 27/226 cytosines modestly decreased methylation by -1.8% (Figure 4.16). The regions of increased methylation overlapped accessible regions (particularly regions 1, 3, and 4, Figure 4.16). More CpG methylation across the remaining enhancers suggested that deletion of the E567 enhancers, led to less engagement of DNA binding factors across the E1234 area, in line with lower *OTX2* expression in cells with this variant (Figure 4.11). In contrast, deletion of the E1234 area was not associated with lower *OTX2* expression (Figure 4.11), and no consistent trend was observed for methylation of the remaining E567 area (36/92 up by an average of +1.9%, 54/92 down by an average of -3.7%, Figure 4.16).

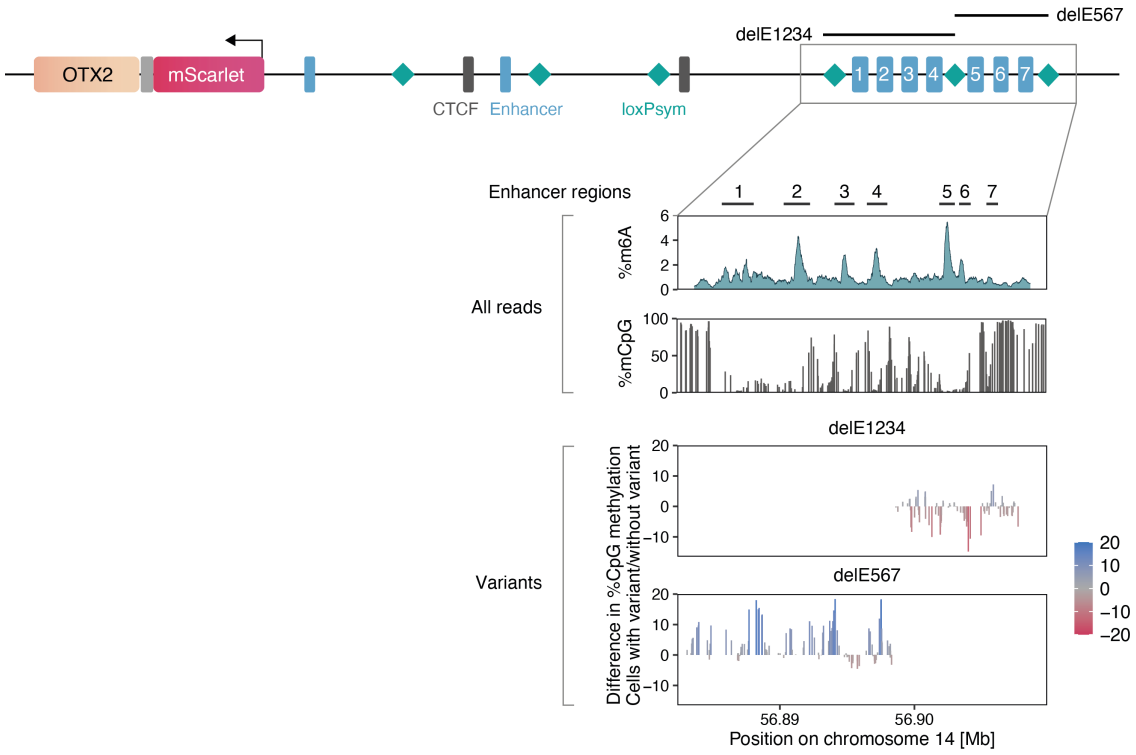

Figure 4.16. Loss of the E567 region increases methylation in the remaining cluster. Schematic of the *OTX2* regulatory region with selected Cre-induced variants shown as lines. Data of various types aligned to the 20 kb enhancer locus (x-axis) are shown in more detail. Plots from top to bottom: Annotation of enhancer regions; averaged percent CpG methylation from all nanopore sequencing reads covering the region; Averaged m6A methylation from all nanopore sequencing reads covering the region; publicly available DNase-seq data; Differences in CpG methylation for reads containing the variants to reads without the variants (color by difference) for the E1234 and E567 deletions.

Next, I compared CpG methylation and accessibility for a 57973 bp variant that moved the 3' area of the 20 kb enhancer cluster closer to the transcription start site. The effects on CpG methylation were overall mixed, slightly increasing at the edges and around regions 5 and 6 (Figure 4.17a). Conversely, accessibility increased in the regulatory regions (from a maximum of 6.2% m6A methylation to 8.4% for region 5, Figure 4.17b) and decreased at their edges. However, these results were only derived from an average of 10 reads per CpG site, and higher-depth sequencing might be necessary to derive conclusive answers on how accessibility and CpG methylation change with novel enhancer architectures.

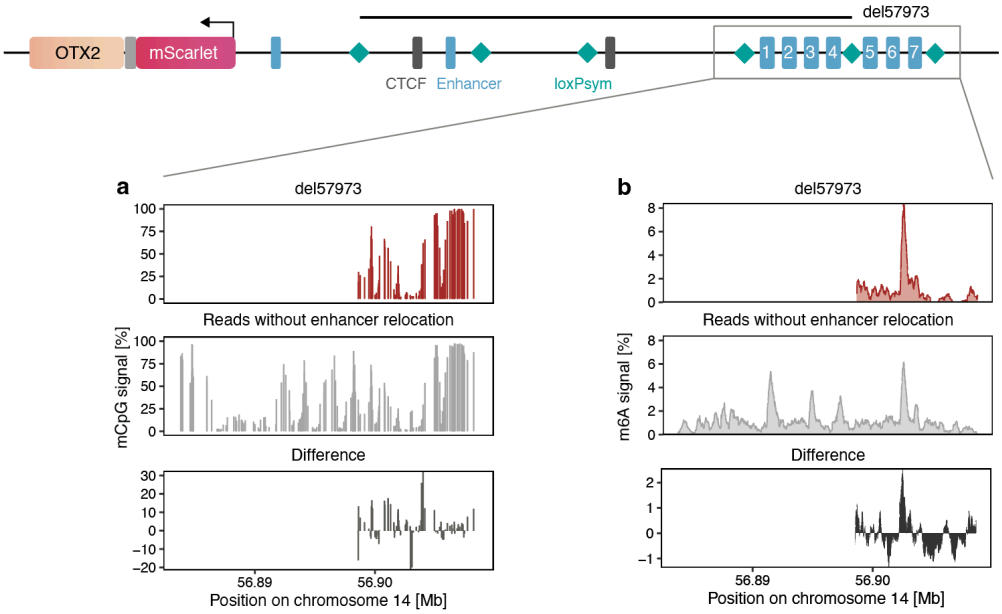

Figure 4.17. Relocation of the enhancer cluster and deletion of intermediate regulatory elements has subtle effects on methylation and accessibility. Schematic of the *OTX2* regulatory region with selected Cre-induced variants shown as lines. Data of various types aligned to the 20 kb enhancer locus (x-axis) are shown in more detail. **a.** Plots from top to bottom: Averaged percent CpG methylation from nanopore sequencing reads containing the del57973 variant; Averaged CpG methylation from nanopore sequencing reads without the variant; Differences in CpG methylation for reads containing the variant to reads without the variant. **b.** As in (a) but for adenine methylation (accessibility) instead of CpG methylation.

4.3 Discussion

In this chapter, I engineered recombinase recognition sequences into the regulatory region of the developmental transcription factor *OTX2* and randomized the region with Cre recombinase. I characterized the resulting variants and associated accessibility as well as CpG methylation using long-read sequencing, linking architectures to gene expression.

At the core of this work was complex genome engineering. Specifically, I blended transposons, prime editing, recombinases, Cas9 cutting, and homology-directed repair to randomly integrate 14 copies of a doxycycline-inducible prime editor, tag an endogenous gene with a fluorophore, incorporate six recombinase recognition sequences, and create tens of novel enhancer architectures. Beyond the single example shown here, this work highlights how genome writing technologies have advanced to a point where sophisticated re-design of complex genomes is becoming feasible through multiplexed editing.

Several *OTX2* enhancers have been discovered and characterized in humans and mice that act throughout different tissues, and developmental stages (Emerson and Cepko 2011; Kurokawa et al. 2014; Kaufman et al. 2021; Bhansali, Cvekl, and Liu 2020). I focused on an evolutionary conserved enhancer cluster that has, to my knowledge, not been described before, and demonstrated its contribution to ~50% of *OTX2* expression in HAP1 cells. In the cell line setting, deleting intermediate enhancers and CTCF binding sites while relocating an enhancer cluster closer to the transcription start site created a simpler architecture with a similar expression to the original one. However, these intermediate elements might add a layer of more fine-grained control to gene regulation during development by adding additional knobs to tune expression and allowing varying degrees of interactions between the cluster and transcription start site in response to external and internal stimuli. Indeed, the presence of facilitating regulatory elements in complex regulatory regions is increasingly appreciated (Blayney et al. 2023; Lin et al. 2022; Brosh et al. 2023; H. Thomas et al. 2023).

Haploid HAP1 cells lack confounding alleles and are highly amenable to complex genome engineering and subcloning. However, they lack the biological context in which these enhancers would usually be active. For example, *OTX2* is active in the adult retina, brain tissues, and throughout the early embryo (Beby and Lamonerie 2013; Acampora et al. 1995), neither of which HAP1 cells represent. Instead, HAP1 cells likely express *OTX2* because they contain high levels of stem cell factors (Yamanaka factors) after retroviral reprogramming attempts (Carette et al. 2011; Takahashi et al. 2007). To understand regulatory regions in more native biological contexts, enhancer randomization with prime editing and recombinases could be attempted in haploid stem

cells (Cui et al. 2020). Differentiation of such stem cells would also illuminate how modified enhancer architectures would respond to a changing cell state. Alternatively, the locus of interest could be first 'haplodized' (Erwood et al. 2022) in a relevant cell line. Nevertheless, DNA sequence, epigenetic memory, and the quantity of available transcription factors together determine how genes are expressed, and I believe HAP1 cells are a great model to derive a mechanistic understanding of how these layers are linked.

The other two low-hanging fruits for optimization are the diversity of Cre-induced outcomes, and strategies to shift away from the dominance of full locus deletions and wild-type architectures as well as improving the sequencing throughput. I will focus on strategies to scale in the final, fifth chapter of this thesis.

In summary, I have developed a novel strategy to study gene regulatory regions in their endogenous context by first preparing recombination anchors with multiplexed prime editing, subsequent diversification with Cre, and finally pooled readout and phenotyping by sequencing. The method employed here should combine well with other methods to study regulatory architectures. For example, clones with enhancer deletions will retain a loxPsym site at the center of the deletion which represents a novel landing pad for the re-integration of regulatory sequence libraries flanked with loxPsym sites whose ability to activate *OTX2-mScarlet* expression could be evaluated by cell sorting and sequencing, akin to an MPRA. Further, the cell lines contain endogenous copies of prime editor which could be leveraged to screen libraries of single nucleotide substitutions in the individual regulatory elements (successfully implemented in (Martyn et al. 2023)), possibly in conjunction with enhancer scrambling, opening exciting new paths to study gene regulation.

4.4 Methods

Cell culture

HAP1 Δ*MLH1* cells were cultured in IMDM (Invitrogen) with 10% FBS (Invitrogen) and glutamine and penicillin-streptomycin (Invitrogen) at 37 °C and 5% CO2.

HAP1 expression data in cancer cell lines and tissues

Cancer cell line gene expression data were downloaded from the Cancer Cell Line Encyclopedia Data portal (https://sites.broadinstitute.org/ccle/datasets; 22Q2 release (Barretina et al. 2012)). Expression data in human tissues were obtained and plotted through the Genotype-Tissue Expression data portal (https://www.gtexportal.org/home/gene/OTX2; GTEx Analysis Release V8 (dbGaP Accession phs000424.v8.p2; accessed December 2023; (Lonsdale et al. 2013)).

Chromatin states and epigenetic analyses

The publicly available HAP1 datasets shown in Table 3.1 were collected and analyzed as outlined in section 3.4 (page 101). In addition, the GSM4625025 for CTCF-ChIP-seq in HAP1 and ENCSR000DTW for CTCF-ChIP-seq in HEK293T data sets were visualized in IGV. The 4DNFI1E6NJQJ HiC data set was used for HAP1 cells and visualized using plotgardner.

Prime editing cell lines

Generation of a stable prime editing *MLH1* knockout HAP1 cell line was described in chapter 2.4 (page 54).

pegRNA design

DeepPrime and PRIDICT were not yet published at the time of pegRNA design. DeepPE was available but not trained on insertions. Therefore, I used Cas9-optimized tools for pegRNA design. First, suitable protospacers were identified using CHOPCHOP (Labun et al. 2019; Montague et al. 2014) (Settings: hg38 assembly, CRISPR/Cas9, knockout). Spacers were filtered for ones that had a predicted efficiency score > 40 and no other targets in the genome with 1 or 2 mismatches. The first nucleotide of the spacer was adjusted to be a G for better U6 expression. The pegRNAs were constructed by matching spacers with the cr772 scaffold (Jost et al. 2020) gtttaagagctaagctggaaacagcatagcaagtttaaataaggctagtccgttatcaactcgaaagagtggcaccgagtcggtgc and endowed with a 3' extension containing a 13 nt PBS, a 34 nt loxPsym site, and 30 nt homology to the target. pegRNA designs are shown in Table 4.2.

Table 4.2. pegRNAs used in chapter 4

Protospacer	Scaffold	Extension	Purpose
GGACTACCAT CTATCTGTGT	cr772	tgaaaggacgtttgtgatcccagccccacaATAACTTCG TATAATGTACATTATACGAAGTTATcagatagatggta	Insertion of loxPsym insertion site chr14: 56,884,634.
GACATCAAAT GTACCCCAGT	cr772	acttcacaaaattaagggtttctgccaactATAACTTCG TATAATGTACATTATACGAAGTTATggggtacatttga	Insertion of loxPsym at chr14: 56,898,376.
GAGGGCATTC TAAGAGTTAG	cr772	ttcaacttcaagaaatcctctattcccctaATAACTTCG TATAATGTACATTATACGAAGTTATactcttagaatgc	Insertion of loxPsym at chr14: 56,907,608.
GCTAGAGTAG GGCAGCTACA	cr772	cagcaatgactcctacctgtggatccctgtATAACTTCG TATAATGTACATTATACGAAGTTATagctgccctactc	Insertion of loxPsym at chr14: 56,840,403.
GAATTTAGAG CCCTACGAGG	cr772	gctgacagagacaggcggtaggatccacctATAACTTCG TATAATGTACATTATACGAAGTTATcgtagggctctaa	Insertion of loxPsym at chr14: 56,851,953.
GATTGGTCCA GAATGCCCAT	cr772	atctctcaggctacgccaatcctacctatgATAACTTCG TATAATGTACATTATACGAAGTTATggcattctggacc	Insertion of loxPsym at chr14: 56,864,718.

cr772: gtttaagagctaagctggaaacagcatagcaagtttaaataaggctagtccgttatcaactcgaaagagtggcaccgagtcggtg. All cloned in an epegRNA acceptor vector (Addgene #174038).

epegRNA cloning

Engineered pegRNAs were cloned using golden gate assembly as described in (Anzalone et al. 2019; Nelson et al. 2022). Briefly, for each pegRNA, forward and reverse oligonucleotides were ordered for spacer, scaffold, and 3'-extensions (Integrated DNA Technologies or Merck). An engineered pegRNA acceptor plasmid (Addgene #174038) was linearized with BsaI, the oligonucleotides were hybridized, and the scaffold phosphorylated. The components were assembled using a golden gate reaction (with BsaI and T4 ligase) and transformed into XL10 gold ultracompetent bacteria (Agilent). Plasmids were isolated using the Miniprep kit (Qiagen) and correct assembly was confirmed by Sanger sequencing of the pegRNA.

Inserting loxPsym sites into the OTX2 enhancer cluster.

One day before transfection, 500,000 HAP1 cells were plated into each well of a six-well plate. One hour before transfection, prime editor expression was induced with 1 µM doxycycline. 500 ng of each pegRNA plasmid (in most cases three constructs were co-delivered at once) and 500 ng of a plasmid encoding for BFP and puromycin resistance were transfected per well using Lipofectamine LTX (Invitrogen, using 2.5 µl plus reagent and 7.5 µl LTX solution). For the continuous integration of loxPsym sites 4-6, 2 µg of PE2 plasmid were additionally co-transfected to counteract the silencing of genome-integrated prime editors. One day after transfection, 2 µg/ml puromycin was added to the cells. Three days post-transfection, puromycin, and dead cells were removed and replenished with fresh media. Once cells were confluent, 1-2 million cells were harvested for DNA preparation and the remaining cells were cryopreserved. I had help with transfections from Mélanie Gouley and clone derivation from Valentin Rebernig.

Assaying integration efficiencies.

DNA was extracted from cell pellets using the DNeasy Blood & Tissue kit (Qiagen) following the manufacturer's instructions with one modification: 3 µl of RNase A (New England BioLabs) was added to a mixture of 180 µl of phosphate buffered saline and 20 µl of proteinase K during the resuspension of cell pellets to remove RNAs. To amplify the respective insertion sites, 10-100 ng of extracted DNA were used as input (P1-12, Table 4.3) for 30 cycles of PCR using Q5 High-Fidelity PCR master mix (New England BioLabs). The PCR products were then purified using either the Qiaquick PCR Purification Kit (Qiagen) or Monarch PCR & DNA purification kit (New England BioLabs) and resolved through capillary gel electrophoresis on a DNA high sensitivity chip (Agilent, Bioanalyzer). The insertion efficiencies were estimated based on the molar ratio of the higher molecular weight DNA band (with loxPsym insertion) to the total of the unedited and edited DNA bands.

Table 4.3. Sequences of oligonucleotides used in chapter 4

ID	Name	Sequence	Purpose
P1	*OTX2*_loxPsym_1_F	TGGTTGTTGGAGGTGGGTGGGG	Amplification of loxPsym insertion site chr14: 56,884,634.
P2	*OTX2*_loxPsym_1_R	GGATGGCGTATGAGCGGGATGC	Amplification of loxPsym insertion site chr14: 56,884,634.
P3	*OTX2*_loxPsym_2_F	TGCCTCCTCCTCTCATGAAACCT	Amplification of loxPsym insertion site chr14: 56,898,376.
P4	*OTX2*_loxPsym_2_R	GCAAAACGGCTCAGACAACCCCA	Amplification of loxPsym insertion site chr14: 56,898,376.
P5	*OTX2*_loxPsym_3_F	AGACAATGTCCCTGCCCTCAAG	Amplification of loxPsym insertion site chr14: 56,907,608.
P6	*OTX2*_loxPsym_3_R	ACCAGCATTGCTTGGAAGTGTT	Amplification of loxPsym insertion site chr14: 56,907,608.
P7	*OTX2*_loxPsym_4_F	TCTCCTTCCACTCTGATTGCTCT	Amplification of loxPsym insertion site chr14: 56,840,403.
P8	*OTX2*_loxPsym_4_R	AGCATAGAAAGTGGCTGGAGCT	Amplification of loxPsym insertion site chr14: 56,840,403.
P9	*OTX2*_loxPsym_5_F	TAGTCGCAGTTACTTGGGAGGC	Amplification of loxPsym insertion site chr14: 56,851,953.
P10	*OTX2*_loxPsym_5_R	TTAGCACAAGGGCCAGAAATGC	Amplification of loxPsym insertion site chr14: 56,851,953.
P11	*OTX2*_loxPsym_6_F	AGTACTCACTTGGCAACCAGCT	Amplification of loxPsym insertion site chr14: 56,864,718.
P12	*OTX2*_loxPsym_6_R	CTCCTGGCAAGGGAAGGAAAGA	Amplification of loxPsym insertion site chr14: 56,864,718.
P13	*OTX2*_mScarlet_F	GACCTTCCCTCCCTTCCTTCAC	Amplification of T2A-mScarlet in the context of OTX2
P14	*OTX2*_mScarlet_R	CTTCCTACTTTGGGGGCATGGA	Amplification of T2A-mScarlet in the context of OTX2
P15	*OTX2*_qPCR_1_F	CCAGGAGGCAGTTTGGTCCTTA	Amplicon 1 for *OTX2* qPCR
P16	*OTX2*_qPCR_1_R	CTTCCTACTTTGGGGGCATGGA	Amplicon 1 for *OTX2* qPCR
P17	*OTX2*_qPCR_2_F	CCACCTCCTCTCGCATGAAGAT	Amplicon 2 for *OTX2* qPCR
P18	*OTX2*_qPCR_2_R	ATGGGCTGAGTCTGACCACTTC	Amplicon 2 for *OTX2* qPCR
P19	*TBP*_qPCR_1_F	GCACCACTCCACTGTATCCC	Amplicon for *TBP* qPCR
P20	*TBP*_qPCR_1_R	TATATTCGGCGTTTCGGGCA	Amplicon for *TBP* qPCR

DNA oligonucleotides were designed using Benchling (https://www.benchling.com/) and ordered from IDT or Sigma Aldrich with standard desalting for purification.

Tagging of OTX2 with T2A-mScarlet

One day before transfection, HAP1 cells with 3 loxPsym integrations were seeded into a 24-well plate with 100,000 cells per well. 3 µl of 100 µM Alt-R CRISPR tracrRNA (Integrated DNA Technologies) were mixed with 3 µl of *OTX2*-targeting 100 µM Alt-R CRISPR crRNA (Table 4.4) and heated at 95˚C for 5 minutes and then cooled down to room temperature at 0.1˚C per second. 0.7 µl of the complex was mixed with 0.5 µl Cas9 (30 pmol) and 0.3 µl sterile phosphate buffered saline to form the RNP. For transfection, 0.5 µl of RNP, 500 ng of Alt-R HDR donor block (Integrated DNA Technologies), and 2.5 µl of Cas9 plus reagent were mixed in one tube, and 1.5 µl CRISPRMAX reagent (Integrated DNA Technologies) was mixed with 25 µl of Opti-Mem in the other tube. The mixtures were combined, incubated at room temperature for 10 minutes, and added to the cells. Seven days post-transfection, cells were analyzed by flow cytometry and mScarlet-positive cells were single-cell sorted. DNA was extracted from expanded colonies (see assaying integration efficiency section) and correct mScarlet insertion was confirmed by amplicon PCR (P13-14, Table 4.3) and Sanger sequencing of the PCR product.

Table 4.4. crRNAs used in chapter 4

crRNA	Purpose
ctacaggtcttcacaaaacc	Tagging of *OTX2* with T2A-mScarlet
gttgaccatgttgaaccagg	Cas9-enrichment of the 20 kb enhancer cluster. 5' probe 1.
gttgaccatgttgaaccagg	Cas9-enrichment of the 20 kb enhancer cluster. 5' probe 2.
tggtccaagtaaacaaacgg	Cas9-enrichment of the 20 kb enhancer cluster. 3' probe 1.
gttccataccaagcaagcag	Cas9-enrichment of the 20 kb enhancer cluster. 3' probe 2.
gttccataccaagcaagcag	Cas9-enrichment of the 70 kb region with 6 loxPsym sites. 3' probe 2.
gcagaaacacgaagtaacat	Cas9-enrichment of the 70 kb region with 6 loxPsym sites. 5' probe 1.
agggtgcccattatagtggt	Cas9-enrichment of the 70 kb region with 6 loxPsym sites. 5' probe 2.
cactcctcaaatgcactacc	Cas9-enrichment of the 70 kb region with 6 loxPsym sites. 3' probe 2.

crRNAs were ordered from IDT as Alt-R CRISPR crRNAs.

Junction PCRs of scrambled clones

100,000-500,000 cells were collected by centrifugation in 96 well plates. The media was removed and the pellets resuspended in home-made quick extract buffer (1 mM $CaCl_2$, 3 mM $MgCl_2$, 1 mM EDTA, 1% Triton X-100, 10 mM Tris pH 7.5) with freshly added proteinase K (0.2 mg/ml) followed by 15 min incubation at 65°C and 20 min incubation at 95°C. 2 µl of cell extract were used in 50 µl of PCR reaction with primers binding outside the deletion breakpoints (P1 + P6, Table 4.3, 18 µl H_2O, 2 µl cell extract, 2.5 µl forward primer, 2.5 µl reverse primer, and 25 µl 2x Q5 master mix). The DNA was amplified for 30-35 cycles and visualized by gel electrophoresis (E-gel™ Power Snap Electrophoresis, Invitrogen). I had help with the derivation of clones and junction PCRs from Valentin Rebernig.

Flow cytometry

Samples were analyzed and sorted as described in section 3.4 (page 97)

Deriving a clone with six loxPsym insertions

The steps described above were combined in the following sequence. First, prime editing HAP1 Δ*MLH1* cells were induced with doxycycline and co-transfected with three pegRNAs encoding the first three sites. The pool of cells was subsequently re-induced with doxycycline and re-transfected with the two epegRNAs and single-cell sorted to derive a clone with two loxPsym integrations. Doxycycline was induced in this cell line and it was transfected with a pegRNA encoding the third loxPsym integration. A clone was derived from this population with all three loxPsym sites. This clone was transfected with Cas9 and an HDR-donor DNA encoding *T2A-mScarlet*. After two weeks, mScarlet-positive cells were single-cell sorted, and successful tagging of *OTX2* was confirmed. Finally, this clone was induced with doxycycline and transfected with three epegRNAs and PE2-puromycin to increase the number of loxPsym integrations to six. The mixed population was single-cell sorted to derive a final clone derived with all six integrations.

Cre protein treatment and sorting-based screening

OTX2-mScarlet cells with 6 loxPsym integrations were plated at 100,000 cells per well in a 24-well plate. 1 hour before Cre protein treatment, full media was replaced with media that did not contain FBS or PSG. Purified TAT-Cre protein (Cambridge University Biochemistry Department) was added to the wells at concentrations of 500 nM, 0.334 nM, 0.223 nM, 0.149 nM, 0.099 nM or 0 nM and incubated on the cells for 2 hours. Extracellular Cre protein was removed by incubation in TrypLE Express (ThermoFisher) for 10 minutes followed by reseeding in full media in 6 well plates. After two weeks, cells were harvested, and stained with 10 µg/ml Hoechst 33342, and fluorescence intensities of mScarlet in haploid G1 cells were assessed by flow cytometry. Cells from the 500 nM sample were sorted into four gates based on mScarlet expression (on Sony MA900). The sorted cells were grown out for another 14 days and then harvested for high molecular weight DNA extraction. Cells for sorting-based screening with three loxPsym sites were treated the same way but FBS/PSG were not removed before induction. And cells treated with 2 µM Cre were sorted instead of cells treated with 500 nM of Cre.

High molecular weight DNA extraction and adenine methylation

Between 2 and 4 million cells were harvested by centrifugation and high molecular weight (HMW) genomic DNA was extracted using the Monarch High Molecular Weight DNA Extraction kit (New England BioLabs) and agitation speeds of 2,000 rpm. For the scrambling with six loxPsym sites, a modified protocol of the Monarch High Molecular Weight DNA kit was

used that enabled simultaneous assessment of DNA accessibility. Here, 129 µl of nuclei prep buffer was mixed with 10 µl of recombinant Cutsmart, 1 µl of 32 mM S-adenosylmethionine (160 µM), 10 µl of the 5000 Units/ml EcoGII methyltransferase (50 Units, New England BioLabs), and 5 µl of 20 mg/ml RNase A (200 µg, New England BioLabs). 2-3 million cells were resuspended in this reaction buffer and incubated at 37ºC for 30 minutes. The reaction was incubated for 5 mins at 65ºC to inactivate the enzymes. 150 µl of nuclei prep buffer containing 10 µl of proteinase K were added to the prepped nuclei and HMW DNA was extracted according to the Monarch High Molecular Weight DNA kit following the manufacturer's instructions. For samples from the screen with six loxPsym sites an additional removal step of low-molecular-weight DNA was done using the Short Fragment Eliminator kit (Oxford Nanopore Technologies) and following the manufacturer's protocol.

Cas9-enrichment

Two crRNAs for each end of the enrichment were designed using CHOPCHOP (Labun et al. 2019; Montague et al. 2014) (Settings: hg38, CRISPR/Cas9, nanopore enrichment). Only protospacers with efficiency scores > 50, GC content > 30 and < 70, self-complementarity = 0, and no other target sites with one or two mismatches were considered. ALT-R tracrRNA and crRNAs were ordered from Integrated DNA Technologies. crRNAs used are shown in Table 4.4. Cas9 enrichment was performed following the Cas9 sequencing kit protocol with slight modifications (CAS9106 Protocolv109, Oxford Nanopore Technologies). Briefly, 5 µg of HMW DNA was dephosphorylated. tracrRNA and crRNA were annealed and complexed with HifiCas9 V3 (Integrated DNA Technologies) to form RNPs. Dephosphorylated DNA was treated with Cas9 RNPs for 30 minutes at 37ºC. Sequencing adapters were ligated to the cut DNA. The ligation step was extended to 1 hour at room temperature. The libraries were then purified using Ampure XP beads (Agilent), washed with Long fragment buffer, and eluted in elution buffer for 16 hours. Because the DNA was very long, it precipitated at the ligation step and only went back into solution following a long elution step. The libraries were sequenced with the MinION Mk1B using R9.4.1 flow cells (FLO-MIN106).

Base and variant calling

Dorado basecaller (v0.5.1) with the dna_r9.4.1_e8_sup@v3.6 model was used for base calling without modifications and for mCpG modified base calling. Guppy (v6.4.6) basecaller with the res_dna_r941_min_modbases-all-context_v001.cfg model was used for m6A modified base calling. The reads were aligned to the hg38 reference genome with minimap2 (H. Li 2016, 2018). Structural variants were called with sniffles2 (Smolka et al. 2023) and relaxed qc settings (--mosaic, --minsvlen 25, --qc-output-all, --mapq 1, --output-rnames) as well as nanomonsv (Using whole-genome sequencing data from HAP1 cells (Chapter 3) as control and requiring one supporting read for the tumor and 0 for the control). The raw variants were filtered for ones that start and end within 50 bp of a loxPsym insertion site and the start and end positions of the variant

were adjusted to the precise loxPsym insertion position. Variants from sniffles and nanomonsv were combined, using the results from the caller that identified more reads for each variant.

For the screen scrambling three loxPsym sites, enough contiguous reads were obtained to distinguish simple variants, complex variants, and wild-type reads. Therefore, read names for filtered variants were extracted. Simple variants were classified as ones whose reads were exclusive to one rearrangement. For complex variants, the same read re-occurred in at least two different rearrangements. Finally, a wild-type sequence was assumed if one read did not occur in any rearrangement and had evidence of three non-rearranged loxPsym sequences. For the screen scrambling six loxPsym sites, no attempt was made to classify variants into simple and complex. To estimate the number of non-rearranged variants, all variant supporting reads across each sorting gate were summed up and subtracted from the maximum coverage for each sorting gate.

Analysis of mCpG and m6A DNA modifications

Analysis across the whole cluster: For mCpG, all aligned sequencing results (bam files with modification information) from various gates and unsorted samples of the screens with cell lines that had integrated three and six loxPsym sites were combined (samtools merge). For m6A, all sequencing results from DNA that was treated with EcoGII methyltransferase (samples from the screen with six loxPsym sites) were combined in the same way. The modification information was aggregated into per-position modification tables (bed files) using modkit pileup (--preset traditional) for CpG methylation and modbam2bed (-e -m m6A) for m6A methylation. Modification tables were further analyzed in R. Only positions with coverage > 5 were kept. A smoothing function was applied to the raw calls. For mCpG, for each position, an average of the preceding and following positions was calculated. For m6A, the average was calculated for the preceding and following 100 positions.

Individual variants: The merged bam files described above were first subsetted for only the reads that were associated with each variant (samtools view, read names to variant association from the sniffles variant caller). For the counter set without the variant to compare against, the files were subsetted to have mapped reads in the deleted area (samtools view). Positional methylation information was aggregated from the subsetted bam files and processed in R as described above, with the following difference: positions with coverage >1 were kept for m6A and the del57973 variant due to sparsity of variant-specific reads.

RNA sequencing

RNA was extracted from 2-5 million flash-frozen cells using the RNeasy plus mini kit (Qiagen). Libraries were prepared using the New England BioLabsNext® Ultra™ II Directional RNA

Library Prep Kit for Illumina (New Englands Biolabs), multiplexed, and sequenced on two Illumina-HTP Novaseq 6000 lanes using 150 bp paired end reads. The median insert size was 280 bp (quartiles 231, 355). Between 58 and 191 million reads were generated per sample. Salmon (2.0.0 (Patro et al. 2017)) was used to quantify transcripts against a salmon index built for the hg38 human reference genome. Further analysis was done using custom R scripts. Transcripts were collapsed to gene level using tximport (1.22.0) (Data S5). Normalization (rlog) and differential expression analysis were performed using DESeq2 (1.34.0). The ensembl v110 release was used for gene structure annotation (imported into R using biomaRt (2.50.3). For all analyses involving genes, the gene lists were filtered to contain only "protein_coding", "lncRNA", "snRNA", "snoRNA", "miRNA", "rRNA", "ribozyme" genes with expression base means > 20. RNA sequencing data from HAP1 Δ*MLH1* cells expressing prime editor and without loxPsym site integrations were plotted in Figure 4.3b. RNA sequencing data from an unscrambled clone with 3 loxPsym sites and 2 scrambled clones with enhancer cluster deletions were plotted in Figure 4.11b. I had help with RNA extraction from Valentin Rebernig.

qPCR

RNA was extracted from 2-5 million flash-frozen cells using the RNeasy plus mini kit (Qiagen) and reverse transcribed into cDNA using the SuperScript VILO cDNA Synthesis kit. 2 µl of cDNA was used in a 20 µl qPCR reaction using the SYBR Green real-time PCR master mix (ThermoFisher). For qPCR, two different fragments of *OTX2* and one fragment of the housekeeping gene *TBP* were amplified in replicates (P14-20, Table 4.3). The amplification was monitored in real-time through the detection of SYBR green fluorescence on the StepOne Plus Real-Time PCR system (ThermoFisher). The differences in PCR cycle number (Δct) between *OTX2* and *TBD* were calculated and averaged across replicates and amplicons. I had help performing qPCRs from Valentin Rebernig.

Software

Genomics: bwa (0.7.17), benchling (accessed between 2019-2023), bedtools (v2.29.0), ChromHMM (1.24), dorado (0.5.1), guppy (6.3.8 and 6.4.6), minimap2 (2.22), modbam2bed (0.9.5), modkit (0.1.12), mosdepth (0.3.3), sniffles2 (2.0.7), samtools (1.14), salmon, IGV (2.16.2), R (4.1.3), nanomonsv (0.6.0). Flow cytometry: FlowJo (v10), CytoExploreR (1.1.0). CytExpert.

R packages: biomaRt (2.50.3), DESeq2 (1.34.0), fgsea (1.20.0), ggrepel (0.9.3), ggpointdensity (0.1.0), plotgardnerer (1.4.2), StructalVariantAnnotation (1.10.1), tidyverse (1.3.2), viridis (0.6.2). tximport (1.22.0). ggpointdensity (0.1.0), RBioinf (1.48.0), reversetranslate (1.0.0), ShortRead (1.46.0), spgs (1.0-3), Tidyverse (1.3.1), Viridis (0.6.1).

5

Conclusions and Outlook

The human genome is coordinating the functions of trillions of cells in the around eight billion humans alive today. Diversity in its precise spelling together with environmental factors predict disease susceptibility (Jonsson et al. 2012; Cohen et al. 2006; Uda et al. 2008), phenotypic appearance (children look like their parents), and other aspects of our lives. Yet, the genomes' content and organization are highly non-streamlined and a result of natural selection acting on an endless series of coincidences and encounters with selfish genetic elements that reach back to the beginning of life. Despite the wonderful diversity across humans, our genomes differ from one another by only about 0.15% (Byrska-Bishop et al. 2022; 1000 Genomes Project Consortium et al. 2010). I believe that to understand the human genome at the end of the one path that was taken by evolution we need to engineer and compare the outputs and behaviors of genomes along the infinite paths that were never taken, and that are orders of magnitude more different than what we can currently find in the human population while probing sequence spaces more diverse in sequence and organization than those found in related species.

To achieve this, I developed strategies to write into genomes of human cell lines more efficiently and thereby prepare entire genomes or regulatory regions for subsequent large-scale randomization. Comparing these novel configurations to unchanged genomes revealed selective pressures acting on structural variants, but also highlighted the flexibility of even haploid genomes to accommodate large changes while still maintaining similar gene expression, morphology, and growth characteristics. More targeted randomization in gene regulatory regions allowed the dissection of an enhancer cluster and highlighted how various novel architectures can drive strong expression if they simultaneously delete some but move other enhancers closer to the transcription start site of the target gene. The projects described in this thesis are promising starting points but their full potential would only unfold if we can scale both the generation and phenotyping of alternative genomes.

5.1 Improvements to prime editing

At the core of the strategies to scramble genomes and regulatory regions lies the capability of prime editors to insert recombinase handles at desired locations. Given the remarkable advances in the prime editing field during the short four years since their initial publication, I am optimistic that the large-scale genome engineering I have undertaken in this thesis could be executed with more ease and on a larger scale today. The advances summarized below could be the starting point for 'mammalian scramble 2.0'.

Prime editing is a complex system involving many components and each one can be independently improved. The nCas9-RT protein itself was codon optimized and endowed with better nuclear localization sequences resulting in the PEmax and PE* architectures (P. J. Chen et al. 2021; P. Liu et al. 2021). Doman et al. further used phage-assisted evolution and protein engineering of the reverse transcriptase to create prime editor variants that are better at inserting structured sequences (PE6c, PE6d) (Doman et al. 2023). Adamson et al. communicated even further improved prime editors through fusion with a safeguarding protein at the 2023 Cold Spring Harbor CRISPR conference. On the side of the pegRNA, prediction tools trained on massive amounts of prime editing data in diverse systems are improving rapidly (G. Yu et al. 2023; Mathis, Allam, Kissling, et al. 2023; Mathis, Allam, Tálas, et al. 2023). Many pegRNAs I tested throughout this thesis were inactive and the fraction of these should decrease drastically with better prediction tools. The flurry of activity we have seen in recent years is likely only going to accelerate until prime editing becomes as efficient as base editors or making knockouts with Cas9.

5.2 Scaling genome scramble

The 17-year synthetic yeast genome project is nearing completion and 6.5 synthetic chromosomes have been consolidated in a single yeast strain (Y. Zhao et al. 2023). The completed synthetic yeast genome will have a total of just under 4,000 loxPsym integrations spaced 2.9 kb from each other on average (Richardson et al. 2017). I achieved around one-tenth the number of loxPsym integrations in a single experiment but at three orders of magnitude lower density across the genome. Increasing the number of sites will increase their frequency of rearrangement and create more potentially viable variants that do not affect essential DNA.

The three main drivers for more efficient recombinase site integration should be (1) the fraction of DNA single-strand breaks (the main driver of toxicity (Smith et al. 2020)) that will convert to faithful integrations, (2) the copy number and diversity of targeted high copy sequences, and (3) the experimental protocol that ensures highest cell viability. Better prime editors (discussed

above), and more active pegRNAs should help increase the editing efficiency. I only tested a single pegRNA design in this thesis and settled on it after I saw signals of editing. Adjusting parameters such as the length of the PBS, RTT, and precise insert sequence should yield more active pegRNAs. In addition, the human genome is replete with repetitive elements that could be targeted, some with much higher copy numbers than LINE-1. For example, Smith et al. designed guides against the Alu element that have copy numbers in the hundreds of thousands (Smith et al. 2020). However, it will be important to balance between copy number and toxicity.

Arguably the largest bottleneck is viability after scrambling. Two main forces cause cell death or stalled growth after genome randomization. The first one is an abundance of a- and dicentric chromosomes that arise in fold-backs as well as translocations and the second one is loss of essential DNA. Increasing the number of loxPsym sequences could ameliorate both. Fold-backs form when two loxPsym sites on sister chromatids or non-homologous chromosomes rearrange. I speculate that fold-backs might preferentially form if distances to the nearest loxPsym site *in cis* are far. Even if two sites rearrange *in cis*, the distance between them might be large and include many essential genes. Besides distance to the nearest sites, I demonstrated that fewer fold-backs are formed, more cells survive, and more variants remain if cells were arrested in the G1 phase of the cell cycle before scrambling. Building on this finding, the activity of Cre protein could be restricted to G1, for example by fusion with a cyclin-dependent kinase degron. Similarly, the original SCRaMbLE toolkit in yeast cells made use of a daughter cell-specific, oestradiol-inducible Cre recombinase that ensures only a single pulse of recombination per cell generation, and only when the genome is present in a single copy (J. S. Dymond et al. 2011).

Restricting Cre-activity will be another important knob to tune cell survival after scrambling. The membrane permeable protein approach I used in this thesis is difficult to titrate and internalization depends on protein precipitation on the membrane (Wadia, Stan, and Dowdy 2004). Consequently, cells will either take up none or large quantities of Cre. An ideal system should ensure that each cell in the population is exposed to a precise amount of Cre. I hypothesize that one way to achieve this is by stably integrating oestradiol or doxycycline-inducible Cre recombinase in a safe harbor locus. Cre expression should be coupled to an antibiotic resistance gene and a fluorescent reporter to monitor and select against silencing. With a system like this, rearrangements could accumulate over a prolonged period, similar to the strategy that enabled high copy number loxPsym integrations for me (section 3.2.1).

What could we do with a cell line with thousands of recombinase sites and a recombinase cassette that enables a slow but steady accumulation of structural variants? We could perform continuous directed evolution on the level of cells and entire genomes. For example, cells could be grown

under suboptimal conditions or in drug-containing media. Under these conditions, genome changes that support growth should start to take over the population. Cells could also be made to produce biologics, such as membrane-bound antibodies, and selected for engagement with antigens. Finally, if we can select cells for genome content (e.g. by staining of DNA and sorting with flow cytometry) it gives a path to minimize genomes. While there might not be the evolutionary pressure to keep repetitive DNA sequences around (especially in the context of a cell line), there also exists no mechanism to easily evict them from the genome. Scramble could provide that mechanism. The amount of DNA a cell line could lose might be substantial. For example, the genome size of the puffer fish (*Fugu rubripes*) is only 400 Mb (12% of ours), and similarly, comparison to other placental mammals shows that only ~11% of the human genome is under constraint (Sullivan et al. 2023).

5.3 Scaling enhancer scramble

The strategies developed in this thesis make it possible to engineer and phenotype many alternative non-coding architectures. I started by randomizing the regulatory region of a single gene (*OTX2*). However, to more comprehensively understand the grammar of gene expression and enable more accurate predictive models, we will likely need to create and measure the consequences of thousands of structural changes in hundreds of genes. To achieve this, there are four natural opportunities to scale enhancer scrambling: (1) Find more suitable target genes. (2) Improve the efficiency of recombinase site integration. (3) Increase the diversity of maintained rearrangement events. (4) Increase the throughput for reading out variants and mapping them to gene expression changes.

To expand to more genes, we first need a strategy to prioritize. There will be many ways to go about this. I teamed up with Ronnie Crawford from the Human Genetics Informatics team who developed an algorithm that ranked genes by their expression in HAP1, diversity of expression across other cell lines, fraction of nearby sequence that is in the enhancer chromatin state, the absolute number of nearby enhancers, distance to nearby genes, and gene size. The type of genes that floated to the top with such an approach was promising and retrieved known models of enhancer biology such as *SOX2* (rank 3) and *MYC* (rank 11).

Improvements in prime editing and pegRNA design (discussed above) will be the main driver for more efficient recombinase site integration. In addition, not every experiment will require clonal lines with many integrations. Having a more diverse pool of integrations could even increase the diversity of events (discussed next).

The diversity of rearrangement events was presumably limited in our experiments because partial deletions and inversions were unstable and continued rearranging until the entire locus was lost. The first way to address this issue is by tightly controlling Cre activity. For example, Cre recombinase expression could be driven by an inducible promoter and additionally be gated by small molecules (e.g. via ligand-induced translocation to the nucleus (Feil et al. 1996)). Moreover, Cre can be fused to a small molecule-induced degron (e.g. (Nishimura et al. 2009; Nabet et al. 2018)) to temporarily restrict its activity and freeze a diverse set of variants before rearrangement towards the terminal full deletion. The second option is to create more stable endpoints by swapping the loxPsym sites used for recombinase sites whose rearrangement product can no longer rearrange (e.g. (Araki, Araki, and Yamamura 1997)). Finally, not every cell requires the full set of all recombinase sites. Instead, we could scramble a cell pool where every cell has integrated a different set of two or more asymmetrical recombinase sites. Even if each set rearranges to its final product (or oscillates for inversions), the resulting cell pool will be diverse. This strategy also speeds up the process of cell line generation as it will no longer be necessary to create clonal lines harboring all integrations. The idea of integrating two recombinase sites also has advantages to improve readout strategies (discussed next).

Arguably the largest bottleneck is reading out enhancer architectures. While Cas9-enriched nanopore sequencing provided locus architecture, methylation, and accessibility on the same read, the enrichment over the genome background was poor (~100-fold resulting in ~1% of all reads mapped to the region of interest) and larger regions were difficult to capture end-to-end. Battaglia et al. modified the Cas9 sequencing protocol and successfully captured contiguous reads for a 116 kb locus with > 300-fold enrichment (Battaglia et al. 2022). These techniques could be combined with enrichment methods directly on the flow cell level such as adaptive sampling which reverses the current and ejects unwanted reads (Payne et al. 2021). Nanopore sequencing itself is also continuously improving and yield as well as base and modification-calling accuracy have increased throughout the years (Kolmogorov et al. 2023).

Classic CRISPR knockout screens do not sequence each target site to confirm knockout but instead sequence sgRNAs and infer a knockout from the presence of guides. If the insertion of recombinase sites with prime editing becomes sufficiently efficient, resulting architectures could also be inferred by sequencing pegRNAs. Instead of relying on stochastic recombination between many sites, each variant would then be encoded by a pair of pegRNAs and hundreds of different pairs could be delivered to a cell population akin to a CRISPR knockout screen. Reading out pegRNAs directly would drastically improve the scale of variants that can be assayed at the same time, and circumvent the need to engineer a cell line altogether. sgRNAs can also be captured in single-cell RNA sequencing experiments alongside the transcriptome. Similarly, it should be

possible to capture pegRNA pairs and infer changes to the target and nearby gene directly instead of relying on indirect sorting for reporter expression. However, more complex variants that combine several deletions and inversions cannot be generated with this strategy (unless using multiple pegRNA pairs encoding orthogonal recombinase sites) and the underlying changes to DNA methylation and accessibility information could not be inferred.

In summary, there will be many promising ways to scale the strategies laid out in this thesis and engineer genomes vastly different from ones that evolved naturally. Hopefully, we will learn lots of novel biology in the process and better understand the peculiarities of our genome. Off to new shores!

References

1000 Genomes Project Consortium, Gonçalo R. Abecasis, David Altshuler, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Richard A. Gibbs, Matt E. Hurles, and Gil A. McVean. 2010. "A Map of Human Genome Variation from Population-Scale Sequencing." *Nature* 467 (7319): 1061–73.

Abel, Haley J., David E. Larson, Allison A. Regier, Colby Chiang, Indraniel Das, Krishna L. Kanchi, Ryan M. Layer, et al. 2020. "Mapping and Characterization of Structural Variation in 17,795 Human Genomes." *Nature* 583 (7814): 83–89.

Abremski, K., and S. Gottesman. 1982. "Purification of the Bacteriophage Lambda Xis Gene Product Required for Lambda Excisive Recombination." *The Journal of Biological Chemistry* 257 (16): 9658–62.

Acampora, D., S. Mazan, Y. Lallemand, V. Avantaggiato, M. Maury, A. Simeone, and P. Brûlet. 1995. "Forebrain and Midbrain Regions Are Deleted in Otx2-/- Mutants due to a Defective Anterior Neuroectoderm Specification during Gastrulation." *Development* 121 (10): 3279–90.

Adikusuma, Fatwa, Sandra Piltz, Mark A. Corbett, Michelle Turvey, Shaun R. McColl, Karla J. Helbig, Michael R. Beard, James Hughes, Richard T. Pomerantz, and Paul Q. Thomas. 2018. "Large Deletions Induced by Cas9 Cleavage." *Nature*.

Allen, Felicity, Luca Crepaldi, Clara Alsinet, Alexander J. Strong, Vitalii Kleshchevnikov, Pietro De Angeli, Petra Páleníková, et al. 2018. "Predicting the Mutations Generated by Repair of Cas9-Induced Double-Strand Breaks." *Nature Biotechnology*, November. https://doi.org/10.1038/nbt.4317.

Altemose, Nicolas, Glennis A. Logsdon, Andrey V. Bzikadze, Pragya Sidhwani, Sasha A. Langley, Gina V. Caldas, Savannah J. Hoyt, et al. 2022. "Complete Genomic and Epigenetic Maps of Human Centromeres." *Science* 376 (6588): eabl4178.

Andersson, Robin, and Albin Sandelin. 2020. "Determinants of Enhancer and Promoter Activities of Regulatory Elements." *Nature Reviews. Genetics* 21 (2): 71–87.

Anzalone, Andrew V., Xin D. Gao, Christopher J. Podracky, Andrew T. Nelson, Luke W. Koblan, Aditya Raguram, Jonathan M. Levy, Jaron A. M. Mercer, and David R. Liu. 2022. "Programmable Deletion, Replacement, Integration and Inversion of Large DNA Sequences with Twin Prime Editing." *Nature Biotechnology* 40 (5): 731–40.

Anzalone, Andrew V., Luke W. Koblan, and David R. Liu. 2020. "Genome Editing with CRISPR–Cas Nucleases, Base Editors, Transposases and Prime Editors." *Nature Biotechnology* 38 (7): 824–44.

Anzalone, Andrew V., Peyton B. Randolph, Jessie R. Davis, Alexander A. Sousa, Luke W. Koblan, Jonathan M. Levy, Peter J. Chen, et al. 2019. "Search-and-Replace Genome Editing without Double-

Strand Breaks or Donor DNA." *Nature* 576 (7785): 149–57.

Araki, K., M. Araki, and K. Yamamura. 1997. "Targeted Integration of DNA Using Mutant Lox Sites in Embryonic Stem Cells." *Nucleic Acids Research* 25 (4): 868–72.

Arbab, Mandana, Max W. Shen, Beverly Mok, Christopher Wilson, Żaneta Matuszek, Christopher A. Cassa, and David R. Liu. 2020. "Determinants of Base Editing Outcomes from Target Library Analysis and Machine Learning." *Cell* 182 (2): 463–80.e30.

Arnold, Cosmas D., Daniel Gerlach, Christoph Stelzer, Łukasz M. Boryń, Martina Rath, and Alexander Stark. 2013. "Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-Seq." *Science* 339 (6123): 1074–77.

Bahr, Carsten, Lisa von Paleske, Veli V. Uslu, Silvia Remeseiro, Naoya Takayama, Stanley W. Ng, Alex Murison, et al. 2018. "Author Correction: A Myc Enhancer Cluster Regulates Normal and Leukaemic Haematopoietic Stem Cell Hierarchies." *Nature* 558 (7711): E4.

Banerji, J., L. Olson, and W. Schaffner. 1983. "A Lymphocyte-Specific Cellular Enhancer Is Located Downstream of the Joining Region in Immunoglobulin Heavy Chain Genes." *Cell* 33 (3): 729–40.

Banerji, J., S. Rusconi, and W. Schaffner. 1981. "Expression of a Beta-Globin Gene Is Enhanced by Remote SV40 DNA Sequences." *Cell* 27 (2 Pt 1): 299–308.

Barretina, Jordi, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A. Margolin, Sungjoon Kim, Christopher J. Wilson, et al. 2012. "The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity." *Nature* 483 (7391): 603–7.

Barski, Artem, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E. Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. 2007. "High-Resolution Profiling of Histone Methylations in the Human Genome." *Cell* 129 (4): 823–37.

Battaglia, Sofia, Kevin Dong, Jingyi Wu, Zeyu Chen, Fadi J. Najm, Yuanyuan Zhang, Molly M. Moore, Vivian Hecht, Noam Shoresh, and Bradley E. Bernstein. 2022. "Long-Range Phasing of Dynamic, Tissue-Specific and Allele-Specific Regulatory Elements." *Nature Genetics* 54 (10): 1504–13.

Beby, Francis, and Thomas Lamonerie. 2013. "The Homeobox Gene Otx2 in Development and Disease." *Experimental Eye Research* 111 (June): 9–16.

Beck, Christine R., José Luis Garcia-Perez, Richard M. Badge, and John V. Moran. 2011. "LINE-1 Elements in Structural Variation and Disease." *Annual Review of Genomics and Human Genetics* 12: 187–215.

Belancio, Victoria P., Dale J. Hedges, and Prescott Deininger. 2008. "Mammalian Non-LTR Retrotransposons: For Better or Worse, in Sickness and in Health." *Genome Research* 18 (3): 343–58.

Bergman, Drew T., Thouis R. Jones, Vincent Liu, Judhajeet Ray, Evelyn Jagoda, Layla Siraj, Helen Y. Kang, et al. 2022. "Compatibility Rules of Human Enhancer and Promoter Sequences." *Nature* 607 (7917): 176–84.

Bernstein, Bradley E., Chih Long Liu, Emily L. Humphrey, Ethan O. Perlstein, and Stuart L. Schreiber. 2004. "Global Nucleosome Occupancy in Yeast." *Genome Biology* 5 (9): R62.

Bhansali, Punita, Ales Cvekl, and Wei Liu. 2020. "A Distal Enhancer That Directs Otx2 Expression in the Retinal Pigment Epithelium and Neuroretina." *Developmental Dynamics: An Official Publication of the American Association of Anatomists* 249 (2): 209–21.

Bhaya, Devaki, Michelle Davison, and Rodolphe Barrangou. 2011. "CRISPR-Cas Systems in Bacteria and Archaea: Versatile Small RNAs for Adaptive Defense and Regulation." *Annual Review of Genetics* 45: 273–97.

Birney, Ewan, T. Daniel Andrews, Paul Bevan, Mario Caccamo, Yuan Chen, Laura Clarke, Guy Coates, et al. 2004. "An Overview of Ensembl." *Genome Research* 14 (5): 925–28.

Blayney, Joseph W., Helena Francis, Alexandra Rampasekova, Brendan Camellato, Leslie Mitchell, Rosa Stolper, Lucy Cornell, et al. 2023. "Super-Enhancers Include Classical Enhancers and Facilitators to Fully Activate Gene Expression." *Cell*, December. https://doi.org/10.1016/j.cell.2023.11.030.

Blomen, Vincent A., Peter Májek, Lucas T. Jae, Johannes W. Bigenzahn, Joppe Nieuwenhuis, Jacqueline Staring, Roberto Sacco, et al. 2015. "Gene Essentiality and Synthetic Lethality in Haploid Human Cells." *Science* 350 (6264): 1092–96.

Blount, B. A., G-O F. Gowers, J. C. H. Ho, R. Ledesma-Amaro, D. Jovicevic, R. M. McKiernan, Z. X. Xie, B. Z. Li, Y. J. Yuan, and T. Ellis. 2018. "Rapid Host Strain Improvement by in Vivo Rearrangement of a Synthetic Yeast Chromosome." *Nature Communications* 9 (1): 1932.

Boer, Carl G. de, Eeshit Dhaval Vaishnav, Ronen Sadeh, Esteban Luis Abeyta, Nir Friedman, and Aviv Regev. 2020. "Deciphering Eukaryotic Gene-Regulatory Logic with 100 Million Random Promoters." *Nature Biotechnology* 38 (1): 56–65.

Bogdanove, Adam J., and Daniel F. Voytas. 2011. "TAL Effectors: Customizable Proteins for DNA Targeting." *Science* 333 (6051): 1843–46.

Boix, Carles A., Benjamin T. James, Yongjin P. Park, Wouter Meuleman, and Manolis Kellis. 2021. "Regulatory Genomic Circuitry of Human Disease Loci by Integrative Epigenomics." *Nature* 590 (7845): 300–307.

Bonev, Boyan, and Giacomo Cavalli. 2016. "Organization and Function of the 3D Genome." *Nature Reviews. Genetics* 17 (11): 661–78.

Boyle, Alan P., Sean Davis, Hennady P. Shulha, Paul Meltzer, Elliott H. Margulies, Zhiping Weng, Terrence S. Furey, and Gregory E. Crawford. 2008. "High-Resolution Mapping and Characterization of Open Chromatin across the Genome." *Cell* 132 (2): 311–22.

Britten, R. J., and D. E. Kohne. 1968. "Repeated Sequences in DNA. Hundreds of Thousands of Copies of DNA Sequences Have Been Incorporated into the Genomes of Higher Organisms." *Science* 161 (3841): 529–40.

Brooks, Aaron N., Amanda L. Hughes, Sandra Clauder-Münster, Leslie A. Mitchell, Jef D. Boeke, and Lars

M. Steinmetz. 2022. “Transcriptional Neighborhoods Regulate Transcript Isoform Lengths and Expression Levels.” *Science* 375 (6584): 1000–1005.

Brosh, Ran, Camila Coelho, André M. Ribeiro-Dos-Santos, Gwen Ellis, Megan S. Hogan, Hannah J. Ashe, Nicolette Somogyi, et al. 2023. “Synthetic Regulatory Genomics Uncovers Enhancer Context Dependence at the Sox2 Locus.” *Molecular Cell* 83 (7): 1140–52.e7.

Brosh, Ran, Jon M. Laurent, Raquel Ordoñez, Emily Huang, Megan S. Hogan, Angela M. Hitchcock, Leslie A. Mitchell, et al. 2021. “A Versatile Platform for Locus-Scale Genome Rewriting and Verification.” *Proceedings of the National Academy of Sciences of the United States of America* 118 (10). https://doi.org/10.1073/pnas.2023952118.

Brown, William R. A., Nicholas C. O. Lee, Zhengyao Xu, and Margaret C. M. Smith. 2011. “Serine Recombinases as Tools for Genome Engineering.” *Methods* 53 (4): 372–79.

Buchholz, F., Y. Refaeli, A. Trumpp, and J. M. Bishop. 2000. “Inducible Chromosomal Translocation of AML1 and ETO Genes through Cre/loxP-Mediated Recombination in the Mouse.” *EMBO Reports* 1 (2): 133–39.

Buckley, Megan, Christina M. Kajba, Nicole Forrester, Chloé Terwagne, Chelsea Sawyer, Scott T. C. Shepherd, Joachim De Jonghe, Phoebe Dace, Samra Turajlic, and Gregory M. Findlay. 2023. “Saturation Genome Editing Resolves the Functional Spectrum of Pathogenic VHL Alleles.” *bioRxiv*. https://doi.org/10.1101/2023.06.10.542698.

Buenrostro, Jason D., Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. 2013. “Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position.” *Nature Methods* 10 (12): 1213–18.

Byrska-Bishop, Marta, Uday S. Evani, Xuefang Zhao, Anna O. Basile, Haley J. Abel, Allison A. Regier, André Corvelo, et al. 2022. “High-Coverage Whole-Genome Sequencing of the Expanded 1000 Genomes Project Cohort Including 602 Trios.” *Cell* 185 (18): 3426–40.e19.

Cameron, Daniel L., Jonathan Baber, Charles Shale, Anthony T. Papenfuss, Jose Espejo Valle-Inclan, Nicolle Besselink, Edwin Cuppen, and Peter Priestley. 2019. “GRIDSS, PURPLE, LINX: Unscrambling the Tumor Genome via Integrated Analysis of Structural Variation and Copy Number.” *bioRxiv*. https://doi.org/10.1101/781013.

Carette, Jan E., Matthijs Raaben, Anthony C. Wong, Andrew S. Herbert, Gregor Obernosterer, Nirupama Mulherkar, Ana I. Kuehne, et al. 2011. “Ebola Virus Entry Requires the Cholesterol Transporter Niemann-Pick C1.” *Nature* 477 (7364): 340–43.

Cerbini, Trevor, Ray Funahashi, Yongquan Luo, Chengyu Liu, Kyeyoon Park, Mahendra Rao, Nasir Malik, and Jizhong Zou. 2015. “Transcription Activator-like Effector Nuclease (TALEN)-Mediated CLYBL Targeting Enables Enhanced Transgene Expression and One-Step Generation of Dual Reporter Human Induced Pluripotent Stem Cell (iPSC) and Neural Stem Cell (NSC) Lines.” *PloS One* 10 (1): e0116032.

Chen, Baohui, Luke A. Gilbert, Beth A. Cimini, Joerg Schnitzbauer, Wei Zhang, Gene-Wei Li, Jason Park, et al. 2013. "Dynamic Imaging of Genomic Loci in Living Human Cells by an Optimized CRISPR/Cas System." *Cell* 155 (7): 1479–91.

Chen, Peter J., Jeffrey A. Hussmann, Jun Yan, Friederike Knipping, Purnima Ravisankar, Pin-Fang Chen, Cidi Chen, et al. 2021. "Enhanced Prime Editing Systems by Manipulating Cellular Determinants of Editing Outcomes." *Cell* 184 (22): 5635–52.e29.

Chen, Peter J., and David R. Liu. 2022. "Prime Editing for Precise and Highly Versatile Genome Manipulation." *Nature Reviews. Genetics*, November. https://doi.org/10.1038/s41576-022-00541-1.

Chen, R., J. B. Bouck, G. M. Weinstock, and R. A. Gibbs. 2001. "Comparing Vertebrate Whole-Genome Shotgun Reads to the Human Genome." *Genome Research* 11 (11): 1807–16.

Chen, Siwei, Laurent C. Francioli, Julia K. Goodrich, Ryan L. Collins, Masahiro Kanai, Qingbo Wang, Jessica Alföldi, et al. 2023. "A Genomic Mutational Constraint Map Using Variation in 76,156 Human Genomes." *Nature*, December. https://doi.org/10.1038/s41586-023-06045-0.

Chen, Wei, Junhong Choi, Jenny F. Nathans, Vikram Agarwal, Beth Martin, Eva Nichols, Anh Leith, Choli Lee, and Jay Shendure. 2021 "Multiplex Genomic Recording of Enhancer and Signal Transduction Activity in Mammalian Cells." https://doi.org/10.1101/2021.11.05.467434.

Chen, Yuting, Eriona Hysolli, Anlu Chen, Stephen Casper, Songlei Liu, Kevin Yang, Chenli Liu, and George Church. 2022. "Multiplex Base Editing to Convert TAG into TAA Codons in the Human Genome." *Nature Communications* 13 (1): 4482.

Choi, Junhong, Wei Chen, Anna Minkina, Florence M. Chardon, Chase C. Suiter, Samuel G. Regalado, Silvia Domcke, et al. 2022. "A Time-Resolved, Multi-Symbol Molecular Recorder via Sequential Genome Editing." *Nature* 608 (7921): 98–107.

Choi, Junhong, Wei Chen, Chase C. Suiter, Choli Lee, Florence M. Chardon, Wei Yang, Anh Leith, Riza M. Daza, Beth Martin, and Jay Shendure. 2022. "Precise Genomic Deletions Using Paired Prime Editing." *Nature Biotechnology* 40 (2): 218–26.

Christmas, Matthew J., Irene M. Kaplow, Diane P. Genereux, Michael X. Dong, Graham M. Hughes, Xue Li, Patrick F. Sullivan, et al. 2023. "Evolutionary Constraint and Innovation across Hundreds of Placental Mammals." *Science* 380 (6643): eabn3943.

Cohen, Jonathan C., Eric Boerwinkle, Thomas H. Mosley Jr, and Helen H. Hobbs. 2006. "Sequence Variations in PCSK9, Low LDL, and Protection against Coronary Heart Disease." *The New England Journal of Medicine* 354 (12): 1264–72.

Collins, E. C., R. Pannell, E. M. Simpson, A. Forster, and T. H. Rabbitts. 2000. "Inter-Chromosomal Recombination of Mll and Af9 Genes Mediated by Cre-loxP in Mouse Development." *EMBO Reports* 1 (2): 127–32.

Collins, Ryan L., Harrison Brand, Konrad J. Karczewski, Xuefang Zhao, Jessica Alföldi, Laurent C. Francioli, Amit V. Khera, et al. 2020. "A Structural Variation Reference for Medical and Population

Genetics." *Nature* 581 (7809): 444–51.

Cong, Le, F. Ann Ran, David Cox, Shuailiang Lin, Robert Barretto, Naomi Habib, Patrick D. Hsu, et al. 2013. "Multiplex Genome Engineering Using CRISPR/Cas Systems." *Science* 339 (6121): 819–23.

Cortés-Ciriano, Isidro, Jake June-Koo Lee, Ruibin Xi, Dhawal Jain, Youngsook L. Jung, Lixing Yang, Dmitry Gordenin, et al. 2020. "Comprehensive Analysis of Chromothripsis in 2,658 Human Cancers Using Whole-Genome Sequencing." *Nature Genetics* 52 (3): 331–41.

Cosenza, Marco Raffaele, Bernardo Rodriguez-Martin, and Jan O. Korbel. 2022. "Structural Variation in Cancer: Role, Prevalence, and Mechanisms." *Annual Review of Genomics and Human Genetics* 23 (August): 123–52.

Costanzo, Michael, Anastasia Baryshnikova, Jeremy Bellay, Yungil Kim, Eric D. Spear, Carolyn S. Sevier, Huiming Ding, et al. 2010. "The Genetic Landscape of a Cell." *Science* 327 (5964): 425–31.

Craig, I. W. 1994. "Organization of the Human Genome." *Journal of Inherited Metabolic Disease* 17 (4): 391–402.

Crasta, Karen, Neil J. Ganem, Regina Dagher, Alexandra B. Lantermann, Elena V. Ivanova, Yunfeng Pan, Luigi Nezi, Alexei Protopopov, Dipanjan Chowdhury, and David Pellman. 2012. "DNA Breaks and Chromosome Pulverization from Errors in Mitosis." *Nature* 482 (7383): 53–58.

Cui, Tongtong, Zhikun Li, Qi Zhou, and Wei Li. 2020. "Current Advances in Haploid Stem Cells." *Protein & Cell* 11 (1): 23–33.

Davydov, Eugene V., David L. Goode, Marina Sirota, Gregory M. Cooper, Arend Sidow, and Serafim Batzoglou. 2010. "Identifying a High Fraction of the Human Genome to Be under Selective Constraint Using GERP++." *PLoS Computational Biology* 6 (12): e1001025.

DeBose-Scarlett, Evon M., and Beth A. Sullivan. 2021. "Genomic and Epigenetic Foundations of Neocentromere Formation." *Annual Review of Genetics* 55 (November): 331–48.

Despang, Alexandra, Robert Schöpflin, Martin Franke, Salaheddine Ali, Ivana Jerković, Christina Paliou, Wing-Lee Chan, et al. 2019. "Functional Dissection of the Sox9-Kcnj2 Locus Identifies Nonessential and Instructive Roles of TAD Architecture." *Nature Genetics* 51 (8): 1263–71.

Dinkel, Holger, Kim Van Roey, Sushama Michael, Norman E. Davey, Robert J. Weatheritt, Diana Born, Tobias Speck, et al. 2014. "The Eukaryotic Linear Motif Resource ELM: 10 Years and Counting." *Nucleic Acids Research* 42 (Database issue): D259–66.

Dinkel, Holger, Kim Van Roey, Sushama Michael, Manjeet Kumar, Bora Uyar, Brigitte Altenberg, Vladislava Milchevskaya, et al. 2016. "ELM 2016—data Update and New Functionality of the Eukaryotic Linear Motif Resource." *Nucleic Acids Research* 44 (D1): D294–300.

Doench, John G., Nicolo Fusi, Meagan Sullender, Mudra Hegde, Emma W. Vaimberg, Katherine F. Donovan, Ian Smith, et al. 2016. "Optimized sgRNA Design to Maximize Activity and Minimize off-Target Effects of CRISPR-Cas9." *Nature Biotechnology* 34 (2): 184–91.

Doman, Jordan L., Smriti Pandey, Monica E. Neugebauer, Meirui An, Jessie R. Davis, Peyton B. Randolph, Amber McElroy, et al. 2023. "Phage-Assisted Evolution and Protein Engineering Yield Compact, Efficient Prime Editors." *Cell* 186 (18): 3983–4002.e26.

Doman, Jordan L., Alexander A. Sousa, Peyton B. Randolph, Peter J. Chen, and David R. Liu. 2022. "Designing and Executing Prime Editing Experiments in Mammalian Cells." *Nature Protocols* 17 (11): 2431–68.

Doni Jayavelu, Naresh, Ajay Jajodia, Arpit Mishra, and R. David Hawkins. 2020. "Candidate Silencer Elements for the Human and Mouse Genomes." *Nature Communications* 11 (1): 1061.

Drumm, Mitchell L., Assem G. Ziady, and Pamela B. Davis. 2012. "Genetic Variation and Clinical Heterogeneity in Cystic Fibrosis." *Annual Review of Pathology* 7: 267–82.

DuBridge, R. B., P. Tang, H. C. Hsia, P. M. Leong, J. H. Miller, and M. P. Calos. 1987. "Analysis of Mutation in Human Cells by Using an Epstein-Barr Virus Shuttle System." *Molecular and Cellular Biology* 7 (1): 379–87.

Durrant, Matthew G., Alison Fanton, Josh Tycko, Michaela Hinks, Sita S. Chandrasekaran, Nicholas T. Perry, Julia Schaepe, et al. 2023. "Systematic Discovery of Recombinases for Efficient Integration of Large DNA Sequences into the Human Genome." *Nature Biotechnology* 41 (4): 488–99.

Dymond, Jessica, and Jef Boeke. 2012. "The Saccharomyces Cerevisiae SCRaMbLE System and Genome Minimization." *Bioengineered Bugs* 3 (3): 168–71.

Dymond, Jessica S., Sarah M. Richardson, Candice E. Coombes, Timothy Babatz, Héloïse Muller, Narayana Annaluru, William J. Blake, et al. 2011. "Synthetic Chromosome Arms Function in Yeast and Generate Phenotypic Diversity by Design." *Nature* 477 (7365): 471–76.

Emerson, Mark M., and Constance L. Cepko. 2011. "Identification of a Retina-Specific Otx2 Enhancer Element Active in Immature Developing Photoreceptors." *Developmental Biology* 360 (1): 241–55.

ENCODE Project Consortium. 2004. "The ENCODE (ENCyclopedia Of DNA Elements) Project." *Science* 306 (5696): 636–40.

ENCODE Project Consortium. 2012. "An Integrated Encyclopedia of DNA Elements in the Human Genome." *Nature* 489 (7414): 57–74.

ENCODE Project Consortium, Jill E. Moore, Michael J. Purcaro, Henry E. Pratt, Charles B. Epstein, Noam Shoresh, Jessika Adrian, et al. 2020. "Expanded Encyclopaedias of DNA Elements in the Human and Mouse Genomes." *Nature* 583 (7818): 699–710.

Ernst, Jason, and Manolis Kellis. 2017. "Chromatin-State Discovery and Genome Annotation with ChromHMM." *Nature Protocols* 12 (12): 2478–92.

Erwood, Steven, Teija M. I. Bily, Jason Lequyer, Joyce Yan, Nitya Gulati, Reid A. Brewer, Liangchi Zhou, Laurence Pelletier, Evgueni A. Ivakine, and Ronald D. Cohn. 2022. "Saturation Variant Interpretation Using CRISPR Prime Editing." *Nature Biotechnology* 40 (6): 885–95.

Feil, R., J. Brocard, B. Mascrez, M. LeMeur, D. Metzger, and P. Chambon. 1996. "Ligand-Activated Site-Specific Recombination in Mice." *Proceedings of the National Academy of Sciences of the United States of America* 93 (20): 10887–90.

Feng, Siyu, Sayaka Sekine, Veronica Pessino, Han Li, Manuel D. Leonetti, and Bo Huang. 2017. "Improved Split Fluorescent Proteins for Endogenous Protein Labeling." *Nature Communications* 8 (1): 370.

Ferreira da Silva, J., G. P. Oliveira, E. A. Arasa-Verge, C. Kagiou, A. Moretton, G. Timelthaler, J. Jiricny, and J. I. Loizou. 2022. "Prime Editing Efficiency and Fidelity Are Enhanced in the Absence of Mismatch Repair." *Nature Communications* 13 (1): 760.

Findlay, Gregory M., Evan A. Boyle, Ronald J. Hause, Jason C. Klein, and Jay Shendure. 2014. "Saturation Editing of Genomic Regions by Multiplex Homology-Directed Repair." *Nature* 513 (7516): 120–23.

Findlay, Gregory M., Riza M. Daza, Beth Martin, Melissa D. Zhang, Anh P. Leith, Molly Gasperini, Joseph D. Janizek, Xingfan Huang, Lea M. Starita, and Jay Shendure. 2018. "Accurate Classification of BRCA1 Variants with Saturation Genome Editing." *Nature* 562 (7726): 217–22.

Forment, Josep V., Abderrahmane Kaidi, and Stephen P. Jackson. 2012. "Chromothripsis and Cancer: Causes and Consequences of Chromosome Shattering." *Nature Reviews. Cancer* 12 (10): 663–70.

Fowler, Douglas M., David J. Adams, Anna L. Gloyn, William C. Hahn, Debora S. Marks, Lara A. Muffley, James T. Neal, et al. 2023. "An Atlas of Variant Effects to Understand the Genome at Nucleotide Resolution." *Genome Biology* 24 (1): 147.

Gardiner-Garden, M., and M. Frommer. 1987. "CpG Islands in Vertebrate Genomes." *Journal of Molecular Biology* 196 (2): 261–82.

Gaudelli, Nicole M., Alexis C. Komor, Holly A. Rees, Michael S. Packer, Ahmed H. Badran, David I. Bryson, and David R. Liu. 2017. "Programmable Base Editing of A•T to G•C in Genomic DNA without DNA Cleavage." *Nature* 551 (7681): 464–71.

Gershenzon, Naum I., and Ilya P. Ioshikhes. 2005. "Synergy of Human Pol II Core Promoter Elements Revealed by Statistical Sequence Analysis." *Bioinformatics* 21 (8): 1295–1300.

Geurts, Maarten H., Eyleen de Poel, Cayetano Pleguezuelos-Manzano, Rurika Oka, Léo Carrillo, Amanda Andersson-Rolf, Matteo Boretto, et al. 2021. "Evaluating CRISPR-Based Prime Editing for Cancer Modeling and CFTR Repair in Organoids." *Life Science Alliance* 4 (10). https://doi.org/10.26508/lsa.202000940.

Ghavi-Helm, Yad, Aleksander Jankowski, Sascha Meiers, Rebecca R. Viales, Jan O. Korbel, and Eileen E. M. Furlong. 2019. "Highly Rearranged Chromosomes Reveal Uncoupling between Genome Topology and Gene Expression." *Nature Genetics* 51 (8): 1272–82.

Gibson, Daniel G., John I. Glass, Carole Lartigue, Vladimir N. Noskov, Ray-Yuan Chuang, Mikkel A. Algire, Gwynedd A. Benders, et al. 2010. "Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome." *Science* 329 (5987): 52–56.

Gilbert, Luke A., Matthew H. Larson, Leonardo Morsut, Zairan Liu, Gloria A. Brar, Sandra E. Torres,

Noam Stern-Ginossar, et al. 2013. “CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes.” *Cell* 154 (2): 442–51.

Gillies, S. D., S. L. Morrison, V. T. Oi, and S. Tonegawa. 1983. “A Tissue-Specific Transcription Enhancer Element Is Located in the Major Intron of a Rearranged Immunoglobulin Heavy Chain Gene.” *Cell* 33 (3): 717–28.

Gilpatrick, Timothy, Isac Lee, James E. Graham, Etienne Raimondeau, Rebecca Bowen, Andrew Heron, Bradley Downs, Saraswati Sukumar, Fritz J. Sedlazeck, and Winston Timp. 2020. “Targeted Nanopore Sequencing with Cas9-Guided Adapter Ligation.” *Nature Biotechnology* 38 (4): 433–38.

Girskis, Kelly M., Andrew B. Stergachis, Ellen M. DeGennaro, Ryan N. Doan, Xuyu Qian, Matthew B. Johnson, Peter P. Wang, et al. 2021. “Rewiring of Human Neurodevelopmental Gene Regulatory Programs by Human Accelerated Regions.” *Neuron* 109 (20): 3239–51.e7.

Goldberg, Michael L. “SEQUENCE ANALYSIS OF DROSOPHILA HISTONE GENES.” 1979. Thesis, Stanford Univ. (1979).

Golic, K. G., and S. Lindquist. 1989. “The FLP Recombinase of Yeast Catalyzes Site-Specific Recombination in the Drosophila Genome.” *Cell* 59 (3): 499–509.

Gordon, M. Grace, Fumitaka Inoue, Beth Martin, Max Schubach, Vikram Agarwal, Sean Whalen, Shiyun Feng, et al. 2020. “lentiMPRA and MPRAflow for High-Throughput Functional Characterization of Gene Regulatory Elements.” *Nature Protocols* 15 (8): 2387–2412.

Gowers, G-O F., S. M. Chee, D. Bell, L. Suckling, M. Kern, D. Tew, D. W. McClymont, and T. Ellis. 2020. “Improved Betulinic Acid Biosynthesis Using Synthetic Yeast Chromosome Recombination and Semi-Automated Rapid LC-MS Screening.” *Nature Communications* 11 (1): 868.

Gruber, Andreas R., Ronny Lorenz, Stephan H. Bernhart, Richard Neuböck, and Ivo L. Hofacker. 2008. “The Vienna RNA Websuite.” *Nucleic Acids Research* 36 (Web Server issue): W70–74.

Gupta, Shikha, Martin Gellert, and Wei Yang. 2011. “Mechanism of Mismatch Recognition Revealed by Human MutSβ Bound to Unpaired DNA Loops.” *Nature Structural & Molecular Biology* 19 (1): 72–78.

Harrow, Jennifer, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken, et al. 2012. “GENCODE: The Reference Human Genome Annotation for The ENCODE Project.” *Genome Research* 22 (9): 1760–74.

Hart, Traver, Megha Chandrashekhar, Michael Aregger, Zachary Steinhart, Kevin R. Brown, Graham MacLeod, Monika Mis, et al. 2015. “High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities.” *Cell* 163 (6): 1515–26.

Hatch, Emily M., Andrew H. Fischer, Thomas J. Deerinck, and Martin W. Hetzer. 2013. “Catastrophic Nuclear Envelope Collapse in Cancer Cell Micronuclei.” *Cell* 154 (1): 47–60.

Hoess, R. H., A. Wierzbicki, and K. Abremski. 1986. “The Role of the loxP Spacer Region in P1 Site-Specific Recombination.” *Nucleic Acids Research* 14 (5): 2287–2300.

Hofacker, Ivo L. 2003. "Vienna RNA Secondary Structure Server." *Nucleic Acids Research* 31 (13): 3429–31.

Howe, Kevin L., Premanand Achuthan, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M. Ridwan Amode, Irina M. Armean, et al. 2021. "Ensembl 2021." *Nucleic Acids Research* 49 (D1): D884–91.

Hoyt, Savannah J., Jessica M. Storer, Gabrielle A. Hartley, Patrick G. S. Grady, Ariel Gershman, Leonardo G. de Lima, Charles Limouse, et al. 2022. "From Telomere to Telomere: The Transcriptional and Epigenetic State of Human Repeat Elements." *Science* 376 (6588): eabk3112.

Huang, Di, and Ivan Ovcharenko. 2022. "Enhancer-Silencer Transitions in the Human Genome." *Genome Research* 32 (3): 437–48.

Hussmann, Jeffrey A., Jia Ling, Purnima Ravisankar, Jun Yan, Ann Cirincione, Albert Xu, Danny Simpson, et al. 2021. "Mapping the Genetic Landscape of DNA Double-Strand Break Repair." *Cell* 184 (22): 5653–69.e25.

Ijaz, Jannat. 2021. *Reconstructing Chromothriptic Chromosomes in Oesophageal Adenocarcinomas*. University of Cambridge.

Jia, Bin, Yi Wu, Bing-Zhi Li, Leslie A. Mitchell, Hong Liu, Shuo Pan, Juan Wang, et al. 2018. "Precise Control of SCRaMbLE in Synthetic Haploid and Diploid Yeast." *Nature Communications* 9 (1): 1933.

Jiang, Kaiyi, Justin Lim, Samantha Sgrizzi, Michael Trinh, Alisan Kayabolen, Natalya Yutin, Eugene V. Koonin, Omar O. Abudayyeh, and Jonathan S. Gootenberg. 2023. "Programmable RNA-Guided Endonucleases Are Widespread in Eukaryotes and Their Viruses." *bioRxiv : The Preprint Server for Biology*, June. https://doi.org/10.1101/2023.06.13.544871.

Jiang, Tingting, Xiao-Ou Zhang, Zhiping Weng, and Wen Xue. 2022. "Deletion and Replacement of Long Genomic Sequences Using Prime Editing." *Nature Biotechnology* 40 (2): 227–34.

Jinek, Martin, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A. Doudna, and Emmanuelle Charpentier. 2012. "A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity." *Science* 337 (6096): 816–21.

Jinek, Martin, Alexandra East, Aaron Cheng, Steven Lin, Enbo Ma, and Jennifer Doudna. 2013. "RNA-Programmed Genome Editing in Human Cells." *eLife* 2 (January): e00471.

Jonsson, Thorlakur, Jasvinder K. Atwal, Stacy Steinberg, Jon Snaedal, Palmi V. Jonsson, Sigurbjorn Bjornsson, Hreinn Stefansson, et al. 2012. "A Mutation in APP Protects against Alzheimer's Disease and Age-Related Cognitive Decline." *Nature* 488 (7409): 96–99.

Jost, Marco, Daniel A. Santos, Reuben A. Saunders, Max A. Horlbeck, John S. Hawkins, Sonia M. Scaria, Thomas M. Norman, et al. 2020. "Titrating Gene Expression Using Libraries of Systematically Attenuated CRISPR Guide RNAs." *Nature Biotechnology* 38 (3): 355–64.

Kang, Jianping, Jieyi Li, Zhou Guo, Sijie Zhou, Shuxin Su, Wenhai Xiao, Yi Wu, and Yingjin Yuan. 2022. "Enhancement and Mapping of Tolerance to Salt Stress and 5-Fluorocytosine in Synthetic Yeast Strains via SCRaMbLE." *Synthetic and Systems Biotechnology* 7 (3): 869–77.

Karimova, Madina, Josephine Abi-Ghanem, Nicolas Berger, Vineeth Surendranath, Maria Teresa Pisabarro, and Frank Buchholz. 2013. "Vika/vox, a Novel Efficient and Specific Cre/loxP-like Site-Specific Recombination System." *Nucleic Acids Research* 41 (2): e37.

Kaufman, Michael L., Noah B. Goodson, Ko Uoon Park, Michael Schwanke, Emma Office, Sophia R. Schneider, Joy Abraham, Austin Hensley, Kenneth L. Jones, and Joseph A. Brzezinski. 2021. "Initiation of Otx2 Expression in the Developing Mouse Retina Requires a Unique Enhancer and Either Ascl1 or Neurog2 Activity." *Development* 148 (12). https://doi.org/10.1242/dev.199399.

Kilby, N. J., M. R. Snaith, and J. A. Murray. 1993. "Site-Specific Recombinases: Tools for Genome Engineering." *Trends in Genetics: TIG* 9 (12): 413–21.

Kim, Hoon, Nam-Phuong Nguyen, Kristen Turner, Sihan Wu, Amit D. Gujar, Jens Luebeck, Jihe Liu, et al. 2020. "Extrachromosomal DNA Is Associated with Oncogene Amplification and Poor Outcome across Multiple Cancers." *Nature Genetics* 52 (9): 891–97.

Kim, Hui Kwon, Younggwang Kim, Sungtae Lee, Seonwoo Min, Jung Yoon Bae, Jae Woo Choi, Jinman Park, Dongmin Jung, Sungroh Yoon, and Hyongbum Henry Kim. 2019. "SpCas9 Activity Prediction by DeepSpCas9, a Deep Learning–based Model with High Generalization Performance." *Science Advances* 5 (11): eaax9249.

Kim, Hui Kwon, Goosang Yu, Jinman Park, Seonwoo Min, Sungtae Lee, Sungroh Yoon, and Hyongbum Henry Kim. 2021. "Predicting the Efficiency of Prime Editing Guide RNAs in Human Cells." *Nature Biotechnology* 39 (2): 198–206.

Kioussis, D., E. Vanin, T. deLange, R. A. Flavell, and F. G. Grosveld. 1983. "Beta-Globin Gene Inactivation by DNA Translocation in Gamma Beta-Thalassaemia." *Nature* 306 (5944): 662–66.

Kircher, Martin, Daniela M. Witten, Preti Jain, Brian J. O'Roak, Gregory M. Cooper, and Jay Shendure. 2014. "A General Framework for Estimating the Relative Pathogenicity of Human Genetic Variants." *Nature Genetics* 46 (3): 310–15.

Kit, S. 1961. "Equilibrium Sedimentation in Density Gradients of DNA Preparations from Animal Tissues." *Journal of Molecular Biology* 3 (December): 711–16.

Klug, Aaron. 2010. "The Discovery of Zinc Fingers and Their Applications in Gene Regulation and Genome Manipulation." *Annual Review of Biochemistry* 79: 213–31.

Koeppel, Jonas, Juliane Weller, Elin Madli Peets, Ananth Pallaseni, Ivan Kuzmin, Uku Raudvere, Hedi Peterson, Fabio Giuseppe Liberante, and Leopold Parts. 2023. "Prediction of Prime Editing Insertion Efficiencies Using Sequence Features and DNA Repair Determinants." *Nature Biotechnology*, February. https://doi.org/10.1038/s41587-023-01678-y.

Kolmogorov, Mikhail, Kimberley J. Billingsley, Mira Mastoras, Melissa Meredith, Jean Monlong, Ryan Lorig-Roach, Mobin Asri, et al. 2023. "Scalable Nanopore Sequencing of Human Genomes Provides a Comprehensive View of Haplotype-Resolved Variation and Methylation." *Nature Methods* 20 (10): 1483–92.

Komor, Alexis C., Yongjoo B. Kim, Michael S. Packer, John A. Zuris, and David R. Liu. 2016. "Programmable Editing of a Target Base in Genomic DNA without Double-Stranded DNA Cleavage." *Nature* 533 (7603): 420–24.

Korbel, Jan O., and Peter J. Campbell. 2013. "Criteria for Inference of Chromothripsis in Cancer Genomes." *Cell* 152 (6): 1226–36.

Kosicki, Michael, Kärt Tomberg, and Allan Bradley. 2018. "Repair of Double-Strand Breaks Induced by CRISPR-Cas9 Leads to Large Deletions and Complex Rearrangements." *Nature Biotechnology* 36 (8): 765–71.

Kouzarides, Tony. 2007. "Chromatin Modifications and Their Function." *Cell* 128 (4): 693–705.

Kramer, Nicole E., Eric S. Davis, Craig D. Wenger, Erika M. Deoudes, Sarah M. Parker, Michael I. Love, and Douglas H. Phanstiel. 2022. "Plotgardener: Cultivating Precise Multi-Panel Figures in R." *Bioinformatics* 38 (7): 2042–45.

Krupina, Ksenia, Alexander Goginashvili, and Don W. Cleveland. 2021. "Causes and Consequences of Micronuclei." *Current Opinion in Cell Biology* 70 (June): 91–99.

Krupina, Ksenia, Alexander Goginashvili, and Don W. Cleveland. 2023. "Scrambling the Genome in Cancer: Causes and Consequences of Complex Chromosome Rearrangements." *Nature Reviews. Genetics*, November. https://doi.org/10.1038/s41576-023-00663-0.

Kuderna, Lukas F. K., Jacob C. Ulirsch, Sabrina Rashid, Mohamed Ameen, Laksshman Sundaram, Glenn Hickey, Anthony J. Cox, et al. 2023. "Identification of Constrained Sequence Elements across 239 Primate Genomes." *Nature*, November. https://doi.org/10.1038/s41586-023-06798-8.

Kurokawa, Daisuke, Tomomi Ohmura, Yusuke Sakurai, Kenichi Inoue, Yoko Suda, and Shinichi Aizawa. 2014. "Otx2 Expression in Anterior Neuroectoderm and Forebrain/midbrain Is Directed by More than Six Enhancers." *Developmental Biology* 387 (2): 203–13.

Kweon, Jiyeon, Hye-Yeon Hwang, Haesun Ryu, An-Hee Jang, Daesik Kim, and Yongsub Kim. 2022. "Targeted Genomic Translocations and Inversions Generated Using a Paired Prime Editing Strategy." *Molecular Therapy: The Journal of the American Society of Gene Therapy*, September. https://doi.org/10.1016/j.ymthe.2022.09.008.

Kweon, Jiyeon, Jung-Ki Yoon, An-Hee Jang, Ha Rim Shin, Ji-Eun See, Gayoung Jang, Jong-Il Kim, and Yongsub Kim. 2021. "Engineered Prime Editors with PAM Flexibility." *Molecular Therapy: The Journal of the American Society of Gene Therapy* 29 (6): 2001–7.

Labun, Kornel, Tessa G. Montague, Maximilian Krause, Yamila N. Torres Cleuren, Håkon Tjeldnes, and Eivind Valen. 2019. "CHOPCHOP v3: Expanding the CRISPR Web Toolbox beyond Genome Editing." *Nucleic Acids Research* 47 (W1): W171–74.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, et al. 2001. "Initial Sequencing and Analysis of the Human Genome." *Nature* 409 (6822): 860–921.

Landrum, Melissa J., Jennifer M. Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla,

Baoshan Gu, et al. 2016. “ClinVar: Public Archive of Interpretations of Clinically Relevant Variants.” *Nucleic Acids Research* 44 (D1): D862–68.

Landrum, Melissa J., Jennifer M. Lee, Mark Benson, Garth R. Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, et al. 2018. “ClinVar: Improving Access to Variant Interpretations and Supporting Evidence.” *Nucleic Acids Research* 46 (D1): D1062–67.

Lee, E-Chiang, Qi Liang, Hanif Ali, Luke Bayliss, Alastair Beasley, Tara Bloomfield-Gerdes, Laura Bonoli, et al. 2014. “Complete Humanization of the Mouse Immunoglobulin Loci Enables Efficient Therapeutic Antibody Discovery.” *Nature Biotechnology* 32 (4): 356–63.

Lee, G., and I. Saito. 1998. “Role of Nucleotide Sequences of loxP Spacer Region in Cre-Mediated Recombination.” *Gene* 216 (1): 55–65.

Lee, Isac, Roham Razaghi, Timothy Gilpatrick, Michael Molnar, Ariel Gershman, Norah Sadowski, Fritz J. Sedlazeck, Kasper D. Hansen, Jared T. Simpson, and Winston Timp. 2020. “Simultaneous Profiling of Chromatin Accessibility and Methylation on Human Cell Lines with Nanopore Sequencing.” *Nature Methods* 17 (12): 1191–99.

Leibowitz, Mitchell L., Stamatis Papathanasiou, Phillip A. Doerfler, Logan J. Blaine, Lili Sun, Yu Yao, Cheng-Zhong Zhang, Mitchell J. Weiss, and David Pellman. 2021. “Chromothripsis as an on-Target Consequence of CRISPR-Cas9 Genome Editing.” *Nature Genetics* 53 (6): 895–905.

Lettice, Laura A., Simon J. H. Heaney, Lorna A. Purdie, Li Li, Philippe de Beer, Ben A. Oostra, Debbie Goode, Greg Elgar, Robert E. Hill, and Esther de Graaff. 2003. “A Long-Range Shh Enhancer Regulates Expression in the Developing Limb and Fin and Is Associated with Preaxial Polydactyly.” *Human Molecular Genetics* 12 (14): 1725–35.

Liberante, Fabio Giuseppe, and Tom Ellis. 2021. “From Kilobases to Megabases: Design and Delivery of Large DNA Constructs into Mammalian Genomes.” *Current Opinion in Systems Biology* 25 (March): 1–10.

Li, Heng. 2016. “Minimap and Miniasm: Fast Mapping and de Novo Assembly for Noisy Long Sequences.” *Bioinformatics* 32 (14): 2103–10.

Li, Heng. 2018. “Minimap2: Pairwise Alignment for Nucleotide Sequences.” *Bioinformatics* 34 (18): 3094–3100.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics* 25 (16): 2078–79.

Lin, Xueqiu, Yanxia Liu, Shuai Liu, Xiang Zhu, Lingling Wu, Yanyu Zhu, Dehua Zhao, et al. 2022. “Nested Epistasis Enhancer Networks for Robust Genome Regulation.” *Science* 377 (6610): 1077–85.

Liu, Chih Long, Tommy Kaplan, Minkyu Kim, Stephen Buratowski, Stuart L. Schreiber, Nir Friedman, and Oliver J. Rando. 2005. “Single-Nucleosome Mapping of Histone Modifications in S. Cerevisiae.” *PLoS Biology* 3 (10): e328.

Liu, Guanwen, Qiupeng Lin, Shuai Jin, and Caixia Gao. 2022. "The CRISPR-Cas Toolbox and Gene Editing Technologies." *Molecular Cell* 82 (2): 333–47.

Liu, Pengpeng, Shun-Qing Liang, Chunwei Zheng, Esther Mintzer, Yan G. Zhao, Karthikeyan Ponnienselvan, Aamir Mir, et al. 2021. "Improved Prime Editors Enable Pathogenic Allele Correction and Cancer Modelling in Adult Mice." *Nature Communications* 12 (1): 2121.

Liu, Wei, Zhouqing Luo, Yun Wang, Nhan T. Pham, Laura Tuck, Irene Pérez-Pi, Longying Liu, et al. 2018. "Rapid Pathway Prototyping and Engineering Using in Vitro and in Vivo Synthetic Genome SCRaMbLE-in Methods." *Nature Communications* 9 (1): 1936.

Liu, Yao, Xiangyang Li, Siting He, Shuhong Huang, Chao Li, Yulin Chen, Zhen Liu, Xingxu Huang, and Xiaolong Wang. 2020. "Efficient Generation of Mouse Models with the Prime Editing System." *Cell Discovery* 6 (April): 27.

Li, Xiangyang, Xin Wang, Wenjun Sun, Shisheng Huang, Mingtian Zhong, Yuan Yao, Quanjiang Ji, and Xingxu Huang. 2022. "Enhancing Prime Editing Efficiency by Modified pegRNA with RNA G-Quadruplexes." *Journal of Molecular Cell Biology* 14 (4). https://doi.org/10.1093/jmcb/mjac022.

Li, Xiaosa, Lina Zhou, Bao-Qing Gao, Guangye Li, Xiao Wang, Ying Wang, Jia Wei, et al. 2022. "Highly Efficient Prime Editing by Introducing Same-Sense Mutations in pegRNA or Stabilizing Its Structure." *Nature Communications* 13 (1): 1669.

Li, Xiaoyi, Wei Chen, Beth K. Martin, Diego Calderon, Choli Lee, Junhong Choi, Florence M. Chardon, et al. 2023. "Chromatin Context-Dependent Regulation and Epigenetic Manipulation of Prime Editing." *bioRxiv : The Preprint Server for Biology*, April. https://doi.org/10.1101/2023.04.12.536587.

Li, Yilong, Nicola D. Roberts, Jeremiah A. Wala, Ofer Shapira, Steven E. Schumacher, Kiran Kumar, Ekta Khurana, et al. 2020. "Patterns of Somatic Structural Variation in Human Cancer Genomes." *Nature* 578 (7793): 112–21.

Loeb, Lawrence A., and Raymond J. Monnat Jr. 2008. "DNA Polymerases and Human Disease." *Nature Reviews. Genetics* 9 (8): 594–604.

Logsdon, Glennis A., Allison N. Rozanski, Fedor Ryabov, Tamara Potapova, Valery A. Shepelev, Yafei Mao, Mikko Rautiainen, et al. 2023. "The Variation and Evolution of Complete Human Centromeres." *bioRxiv : The Preprint Server for Biology*, May. https://doi.org/10.1101/2023.05.30.542849.

Lohia, Ruchi, Nathan Fox, and Jesse Gillis. 2022. "A Global High-Density Chromatin Interaction Network Reveals Functional Long-Range and Trans-Chromosomal Relationships." *Genome Biology* 23 (1): 238.

Long, Hannah K., Marco Osterwalder, Ian C. Welsh, Karissa Hansen, James O. J. Davies, Yiran E. Liu, Mervenaz Koska, et al. 2020. "Loss of Extreme Long-Range Enhancers in Human Neural Crest Drives a Craniofacial Disorder." *Cell Stem Cell* 27 (5): 765–83.e14.

Lonsdale, John, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, et al. 2013. "The Genotype-Tissue Expression (GTEx) Project." *Nature Genetics* 45 (6): 580–85.

Lopes, Rui, Gozde Korkmaz, and Reuven Agami. 2016. "Applying CRISPR-Cas9 Tools to Identify and Characterize Transcriptional Enhancers." *Nature Reviews. Molecular Cell Biology* 17 (9): 597–604.

Loveless, Theresa B., Courtney K. Carlson, Vincent J. Hu, Catalina A. Dentzel Helmy, Guohao Liang, Michelle Ficht, Arushi Singhai, and Chang C. Liu. 2021. "Molecular Recording of Sequential Cellular Events into DNA." *bioRxiv*. https://doi.org/10.1101/2021.11.05.467507.

Luo, Zhouqing, Lihui Wang, Yun Wang, Weimin Zhang, Yakun Guo, Yue Shen, Linghuo Jiang, et al. 2018. "Identifying and Characterizing SCRaMbLEd Synthetic Yeast Using ReSCuES." *Nature Communications* 9 (1): 1930.

Lynch, Michael, and John S. Conery. 2003. "The Origins of Genome Complexity." *Science* 302 (5649): 1401–4.

Ly, Peter, Simon F. Brunner, Ofer Shoshani, Dong Hyun Kim, Weijie Lan, Tatyana Pyntikova, Adrienne M. Flanagan, et al. 2019. "Chromosome Segregation Errors Generate a Diverse Spectrum of Simple and Complex Genomic Rearrangements." *Nature Genetics* 51 (4): 705–15.

Mali, Prashant, Luhan Yang, Kevin M. Esvelt, John Aach, Marc Guell, James E. DiCarlo, Julie E. Norville, and George M. Church. 2013. "RNA-Guided Human Genome Engineering via Cas9." *Science* 339 (6121): 823–26.

Ma, Lu, Yunxiang Li, Xinyu Chen, Mingzhu Ding, Yi Wu, and Ying-Jin Yuan. 2019. "SCRaMbLE Generates Evolved Yeasts with Increased Alkali Tolerance." *Microbial Cell Factories* 18 (1): 52.

Martinez-Ara, Miguel, Federico Comoglio, Joris van Arensbergen, and Bas van Steensel. 2022. "Systematic Analysis of Intrinsic Enhancer-Promoter Compatibility in the Mouse Genome." *Molecular Cell* 82 (13): 2519–31.e6.

Martyn, Gabriella E., Michael T. Montgomery, Hank Jones, Katherine Guo, Benjamin R. Doughty, Johannes Linder, Ziwei Chen, et al. 2023. "Rewriting Regulatory DNA to Dissect and Reprogram Gene Expression." *bioRxiv*. https://doi.org/10.1101/2023.12.20.572268.

Maston, Glenn A., Sara K. Evans, and Michael R. Green. 2006. "Transcriptional Regulatory Elements in the Human Genome." *Annual Review of Genomics and Human Genetics* 7: 29–59.

Mathis, Nicolas, Ahmed Allam, Lucas Kissling, Kim Fabiano Marquart, Lukas Schmidheini, Cristina Solari, Zsolt Balázs, Michael Krauthammer, and Gerald Schwank. 2023. "Predicting Prime Editing Efficiency and Product Purity by Deep Learning." *Nature Biotechnology* 41 (8): 1151–59.

Mathis, Nicolas, Ahmed Allam, András Tálas, Elena Benvenuto, Ruben Schep, Tanav Damodharan, Zsolt Balázs, et al. 2023. "Predicting Prime Editing Efficiency across Diverse Edit Types and Chromatin Contexts with Machine Learning." *bioRxiv*. https://doi.org/10.1101/2023.10.09.561414.

McDonald, Torrin L., Weichen Zhou, Christopher P. Castro, Camille Mumm, Jessica A. Switzenberg, Ryan

E. Mills, and Alan P. Boyle. 2021. “Cas9 Targeted Enrichment of Mobile Elements Using Nanopore Sequencing.” *Nature Communications* 12 (1): 3586.

McKinley, Kara L., and Iain M. Cheeseman. 2016. “The Molecular Basis for Centromere Identity and Function.” *Nature Reviews. Molecular Cell Biology* 17 (1): 16–29.

Meers, Chance, Hoang C. Le, Sanjana R. Pesari, Florian T. Hoffmann, Matt W. G. Walker, Jeanine Gezelle, Stephen Tang, and Samuel H. Sternberg. 2023. “Transposon-Encoded Nucleases Use Guide RNAs to Promote Their Selfish Spread.” *Nature*, September. https://doi.org/10.1038/s41586-023-06597-1.

Meier, Joshua A., Feng Zhang, and Neville E. Sanjana. 2017. “GUIDES: sgRNA Design for Loss-of-Function Screens.” *Nature Methods* 14 (9): 831–32.

Melnikov, Alexandre, Anand Murugan, Xiaolan Zhang, Tiberiu Tesileanu, Li Wang, Peter Rogov, Soheil Feizi, et al. 2012. “Systematic Dissection and Optimization of Inducible Enhancers in Human Cells Using a Massively Parallel Reporter Assay.” *Nature Biotechnology* 30 (3): 271–77.

Mewes, H. W., K. Albermann, M. Bähr, D. Frishman, A. Gleissner, J. Hani, K. Heumann, et al. 1997. “Overview of the Yeast Genome.” *Nature* 387 (6632 Suppl): 7–65.

Meyers, Robin M., Jordan G. Bryan, James M. McFarland, Barbara A. Weir, Ann E. Sizemore, Han Xu, Neekesh V. Dharia, et al. 2017. “Computational Correction of Copy Number Effect Improves Specificity of CRISPR-Cas9 Essentiality Screens in Cancer Cells.” *Nature Genetics* 49 (12): 1779–84.

Michalik, Stephan, Florian Siegerist, Raghavendra Palankar, Kati Franzke, Maximilian Schindler, Alexander Reder, Ulrike Seifert, et al. 2022. “Comparative Analysis of ChAdOx1 nCoV-19 and Ad26.COV2.S SARS-CoV-2 Vector Vaccines.” *Haematologica* 107 (4): 947–57.

Montague, Tessa G., José M. Cruz, James A. Gagnon, George M. Church, and Eivind Valen. 2014. “CHOPCHOP: A CRISPR/Cas9 and TALEN Web Tool for Genome Editing.” *Nucleic Acids Research* 42 (Web Server issue): W401–7.

Nabet, Behnam, Justin M. Roberts, Dennis L. Buckley, Joshiawa Paulk, Shiva Dastjerdi, Annan Yang, Alan L. Leggett, et al. 2018. “The dTAG System for Immediate and Target-Specific Protein Degradation.” *Nature Chemical Biology* 14 (5): 431–41.

Nahmad, Alessio David, Eli Reuveni, Ella Goldschmidt, Tamar Tenne, Meytal Liberman, Miriam Horovitz-Fried, Rami Khosravi, et al. 2022. “Frequent Aneuploidy in Primary Human T Cells after CRISPR–Cas9 Cleavage.” *Nature Biotechnology* 40 (12): 1807–13.

Nam, Chang Hyun, Jeonghwan Youk, Jeong Yeon Kim, Joonoh Lim, Jung Woo Park, Soo A. Oh, Hyun Jung Lee, et al. 2023. “Widespread Somatic L1 Retrotransposition in Normal Colorectal Epithelium.” *Nature* 617 (7961): 540–47.

Nelson, James W., Peyton B. Randolph, Simon P. Shen, Kelcee A. Everette, Peter J. Chen, Andrew V. Anzalone, Meirui An, et al. 2022. “Engineered pegRNAs Improve Prime Editing Efficiency.” *Nature Biotechnology* 40 (3): 402–10.

Ngan, Chew Yee, Chee Hong Wong, Harianto Tjong, Wenbo Wang, Rachel L. Goldfeder, Cindy Choi, Hao He, et al. 2020. "Chromatin Interaction Analyses Elucidate the Roles of PRC2-Bound Silencers in Mouse Development." *Nature Genetics* 52 (3): 264–72.

Nishimura, Kohei, Tatsuo Fukagawa, Haruhiko Takisawa, Tatsuo Kakimoto, and Masato Kanemaki. 2009. "An Auxin-Based Degron System for the Rapid Depletion of Proteins in Nonplant Cells." *Nature Methods* 6 (12): 917–22.

Niu, Dong, Hong-Jiang Wei, Lin Lin, Haydy George, Tao Wang, I-Hsiu Lee, Hong-Ye Zhao, et al. 2017. "Inactivation of Porcine Endogenous Retrovirus in Pigs Using CRISPR-Cas9." *Science* 357 (6357): 1303–7.

Nora, Elphège P., Anton Goloborodko, Anne-Laure Valton, Johan H. Gibcus, Alec Uebersohn, Nezar Abdennur, Job Dekker, Leonid A. Mirny, and Benoit G. Bruneau. 2017. "Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization." *Cell* 169 (5): 930–44.e22.

Nurk, Sergey, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, Mitchell R. Vollger, et al. 2022. "The Complete Sequence of a Human Genome." *Science* 376 (6588): 44–53.

Olorunniji, Femi J., Susan J. Rosser, and W. Marshall Stark. 2016. "Site-Specific Recombinases: Molecular Machines for the Genetic Revolution." *Biochemical Journal* 473 (6): 673–84.

Ong, Chin-Tong, and Victor G. Corces. 2014. "CTCF: An Architectural Protein Bridging Genome Topology and Function." *Nature Reviews. Genetics* 15 (4): 234–46.

Ordoñez, Raquel, Weimin Zhang, Gwen Ellis, Florrie Zhu, Hannah J. Ashe, André M. Ribeiro-dos-Santos, Ran Brosh, et al. 2023. "Genomic Context Sensitizes Regulatory Elements to Genetic Disruption." *bioRxiv*. https://doi.org/10.1101/2023.07.02.547201.

Orphanides, G., T. Lagrange, and D. Reinberg. 1996. "The General Transcription Factors of RNA Polymerase II." *Genes & Development* 10 (21): 2657–83.

Oster, I. I. 1956. "A New Crossing-over Suppressor in Chromosome 2 Effective in the Presence of Heterologous Inversions." *Drosophila Information Service*.

Pallaseni, Ananth, Elin Madli Peets, Gareth Girling, Luca Crepaldi, Ivan Kuzmin, Uku Raudvere, Hedi Peterson, et al. 2023. "The Interplay of DNA Repair Context with Target Sequence Predictably Biasses Cas9-Generated Mutations." *bioRxiv : The Preprint Server for Biology*, June. https://doi.org/10.1101/2023.06.28.546891.

Pallaseni, Ananth, Elin Madli Peets, Jonas Koeppel, Juliane Weller, Thomas Vanderstichele, Uyen Linh Ho, Luca Crepaldi, Jolanda van Leeuwen, Felicity Allen, and Leopold Parts. 2022. "Predicting Base Editing Outcomes Using Position-Specific Sequence Determinants." *Nucleic Acids Research* 50 (6): 3551–64.

Pang, Baoxu, and Michael P. Snyder. 2020. "Systematic Identification of Silencers in Human Cells."

Nature Genetics 52 (3): 254–63.

Papathanasiou, Stamatis, Styliani Markoulaki, Logan J. Blaine, Mitchell L. Leibowitz, Cheng-Zhong Zhang, Rudolf Jaenisch, and David Pellman. 2021. "Whole Chromosome Loss and Genomic Instability in Mouse Embryos after CRISPR-Cas9 Genome Editing." *Nature Communications* 12 (1): 5855.

Patro, Rob, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. 2017. "Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression." *Nature Methods* 14 (4): 417–19.

Patwardhan, Rupali P., Joseph B. Hiatt, Daniela M. Witten, Mee J. Kim, Robin P. Smith, Dalit May, Choli Lee, et al. 2012. "Massively Parallel Functional Dissection of Mammalian Enhancers in Vivo." *Nature Biotechnology* 30 (3): 265–70.

Payne, Alexander, Nadine Holmes, Thomas Clarke, Rory Munro, Bisrat J. Debebe, and Matthew Loose. 2021. "Readfish Enables Targeted Nanopore Sequencing of Gigabase-Sized Genomes." *Nature Biotechnology* 39 (4): 442–50.

Pedersen, Brent S., and Aaron R. Quinlan. 2018. "Mosdepth: Quick Coverage Calculation for Genomes and Exomes." *Bioinformatics* 34 (5): 867–68.

Peets, Elin Madli, Luca Crepaldi, Yan Zhou, Felicity Allen, Rasa Elmentaite, Guillaume Noell, Gemma Turner, Vivek Iyer, and Leopold Parts. 2019. "Minimized Double Guide RNA Libraries Enable Scale-Limited CRISPR/Cas9 Screens." *bioRxiv*. https://doi.org/10.1101/859652.

Pinglay, Sudarshan, Milica Bulajić, Dylan P. Rahe, Emily Huang, Ran Brosh, Nicholas E. Mamrak, Benjamin R. King, et al. 2022. "Synthetic Regulatory Reconstitution Reveals Principles of Mammalian *Hox* Cluster Regulation." *Science* 377 (6601): eabk2820.

Plück, A. 1996. "Conditional Mutagenesis in Mice: The Cre/loxP Recombination System." *International Journal of Experimental Pathology* 77 (6): 269–78.

Pollard, Katherine S., Melissa J. Hubisz, Kate R. Rosenbloom, and Adam Siepel. 2010. "Detection of Nonneutral Substitution Rates on Mammalian Phylogenies." *Genome Research* 20 (1): 110–21.

Porrua, Odil, Marc Boudvillain, and Domenico Libri. 2016. "Transcription Termination: Variations on Common Themes." *Trends in Genetics: TIG* 32 (8): 508–22.

Pradella, Davide, Minsi Zhang, Rui Gao, Melissa A. Yao, Katarzyna M. Gluchowska, Ylenia Cendon Florez, Tanmay Mishra, et al. 2023. "Immortalization and Transformation of Primary Cells Mediated by Engineered ecDNAs." *bioRxiv : The Preprint Server for Biology*, June. https://doi.org/10.1101/2023.06.25.546239.

Puntervoll, Pål, Rune Linding, Christine Gemünd, Sophie Chabanis-Davidson, Morten Mattingsdal, Scott Cameron, David M. A. Martin, et al. 2003. "ELM Server: A New Resource for Investigating Short Functional Sites in Modular Eukaryotic Proteins." *Nucleic Acids Research* 31 (13): 3625–30.

Qi, Lei S., Matthew H. Larson, Luke A. Gilbert, Jennifer A. Doudna, Jonathan S. Weissman, Adam P. Arkin, and Wendell A. Lim. 2013. "Repurposing CRISPR as an RNA-Guided Platform for Sequence-

Specific Control of Gene Expression." *Cell* 152 (5): 1173–83.

Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26 (6): 841–42.

Radford, E. J., H. K. Tan, M. H. L. Andersson, J. D. Stephenson, E. J. Gardner, H. Ironfield, A. J. Waters, et al. 2022. "Saturation Genome Editing of DDX3X Clarifies Pathogenicity of Germline and Somatic Variation." *bioRxiv*. https://doi.org/10.1101/2022.06.10.22276179.

Ramírez-Solis, R., P. Liu, and A. Bradley. 1995. "Chromosome Engineering in Mice." *Nature* 378 (6558): 720–24.

Ramsköld, Daniel, Eric T. Wang, Christopher B. Burge, and Rickard Sandberg. 2009. "An Abundance of Ubiquitously Expressed Genes Revealed by Tissue Transcriptome Sequence Data." *PLoS Computational Biology* 5 (12): e1000598.

Rand, Arthur C., Miten Jain, Jordan M. Eizenga, Audrey Musselman-Brown, Hugh E. Olsen, Mark Akeson, and Benedict Paten. 2017. "Mapping DNA Methylation with High-Throughput Nanopore Sequencing." *Nature Methods* 14 (4): 411–13.

Reiff, Sarah B., Andrew J. Schroeder, Koray Kırlı, Andrea Cosolo, Clara Bakker, Luisa Mercado, Soohyun Lee, et al. 2022. "The 4D Nucleome Data Portal as a Resource for Searching and Visualizing Curated Nucleomics Data." *Nature Communications* 13 (1): 2365.

Reményi, Attila, Hans R. Schöler, and Matthias Wilmanns. 2004. "Combinatorial Control of Gene Expression." *Nature Structural & Molecular Biology* 11 (9): 812–15.

Rentzsch, Philipp, Daniela Witten, Gregory M. Cooper, Jay Shendure, and Martin Kircher. 2019. "CADD: Predicting the Deleteriousness of Variants throughout the Human Genome." *Nucleic Acids Research* 47 (D1): D886–94.

Richardson, Sarah M., Leslie A. Mitchell, Giovanni Stracquadanio, Kun Yang, Jessica S. Dymond, James E. DiCarlo, Dongwon Lee, et al. 2017. "Design of a Synthetic Yeast Genome." *Science* 355 (6329): 1040–44.

Riesenberg, Stephan, Philipp Kanis, Dominik Macak, Damian Wollny, Dorothee Düsterhöft, Johannes Kowalewski, Nelly Helmbrecht, Tomislav Maricic, and Svante Pääbo. 2023. "Efficient High-Precision Homology-Directed Repair-Dependent Genome Editing by HDRobust." *Nature Methods* 20 (9): 1388–99.

Robert, Francis, Mathilde Barbeau, Sylvain Éthier, Josée Dostie, and Jerry Pelletier. 2015. "Pharmacological Inhibition of DNA-PK Stimulates Cas9-Mediated Genome Editing." *Genome Medicine* 7 (1): 93.

Roeder, R. G. 1996. "The Role of General Initiation Factors in Transcription by RNA Polymerase II." *Trends in Biochemical Sciences* 21 (9): 327–35.

Roth, D. B., and N. L. Craig. 1998. "VDJ Recombination: A Transposase Goes to Work." *Cell* 94 (4): 411–14.

Rothkamm, Kai, Ines Krüger, Larry H. Thompson, and Markus Löbrich. 2003. "Pathways of DNA Double-Strand Break Repair during the Mammalian Cell Cycle." *Molecular and Cellular Biology* 23 (16): 5706–15.

Roy, Scott William, and Walter Gilbert. 2006. "The Evolution of Spliceosomal Introns: Patterns, Puzzles and Progress." *Nature Reviews. Genetics* 7 (3): 211–21.

Sadowski, P. 1986. "Site-Specific Recombinases: Changing Partners and Doing the Twist." *Journal of Bacteriology* 165 (2): 341–47.

Saito, Makoto, Peiyu Xu, Guilhem Faure, Samantha Maguire, Soumya Kannan, Han Altae-Tran, Sam Vo, Anan Desimone, Rhiannon K. Macrae, and Feng Zhang. 2023. "Fanzor Is a Eukaryotic Programmable RNA-Guided Endonuclease." *Nature* 620 (7974): 660–68.

Sanborn, Adrian L., Suhas S. P. Rao, Su-Chen Huang, Neva C. Durand, Miriam H. Huntley, Andrew I. Jewett, Ivan D. Bochkov, et al. 2015. "Chromatin Extrusion Explains Key Features of Loop and Domain Formation in Wild-Type and Engineered Genomes." *Proceedings of the National Academy of Sciences of the United States of America* 112 (47): E6456–65.

Sauer, Brian, and Jeffrey McDermott. 2004. "DNA Recombination with a Heterospecific Cre Homolog Identified from Comparison of the Pac-c1 Regions of P1-Related Phages." *Nucleic Acids Research* 32 (20): 6086–95.

Schaik, Tom van, Mabel Vos, Daan Peric-Hupkes, Patrick Hn Celie, and Bas van Steensel. 2020. "Cell Cycle Dynamics of Lamina-Associated DNA." *EMBO Reports* 21 (11): e50636.

Schene, Imre F., Indi P. Joore, Rurika Oka, Michal Mokry, Anke H. M. van Vugt, Ruben van Boxtel, Hubert P. J. van der Doef, et al. 2020. "Prime Editing for Functional Repair in Patient-Derived Disease Models." *Nature Communications* 11 (1): 5352.

Schindler, Daniel, Roy S. K. Walker, Shuangying Jiang, Aaron N. Brooks, Yun Wang, Carolin A. Müller, Charlotte Cockram, et al. 2023. "Design, Construction, and Functional Characterization of a tRNA Neochromosome in Yeast." *Cell* 186 (24): 5237–53.e22.

Schoenfelder, Stefan, and Peter Fraser. 2019. "Long-Range Enhancer-Promoter Contacts in Gene Expression Control." *Nature Reviews. Genetics* 20 (8): 437–55.

Scholefield, Janine, and Patrick T. Harrison. 2021. "Prime Editing – an Update on the Field." *Gene Therapy*. https://doi.org/10.1038/s41434-021-00263-9.

Shen, Max W., 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. 2018. "Predictable and Precise Template-Free CRISPR Editing of Pathogenic Variants." *Nature* 563 (7733): 646–51.

Shen, Yin, Feng Yue, David F. McCleary, Zhen Ye, Lee Edsall, Samantha Kuan, Ulrich Wagner, et al. 2012. "A Map of the Cis-Regulatory Sequences in the Mouse Genome." *Nature* 488 (7409): 116–20.

Shiraishi, Yuichi, Junji Koya, Kenichi Chiba, Ai Okada, Yasuhito Arai, Yuki Saito, Tatsuhiro Shibata, and Keisuke Kataoka. 2023. "Precise Characterization of Somatic Complex Structural Variations from

Tumor Control Paired Long-Read Sequencing Data with Nanomonsv." *Nucleic Acids Research* 51 (14): e74.

Shoshani, Ofer, Simon F. Brunner, Rona Yaeger, Peter Ly, Yael Nechemia-Arbely, Dong Hyun Kim, Rongxin Fang, et al. 2021. "Chromothripsis Drives the Evolution of Gene Amplification in Cancer." *Nature* 591 (7848): 137–41.

Simpson, Jared T., Rachael E. Workman, P. C. Zuzarte, Matei David, L. J. Dursi, and Winston Timp. 2017. "Detecting DNA Cytosine Methylation Using Nanopore Sequencing." *Nature Methods* 14 (4): 407–10.

Sloan, Cricket A., Esther T. Chan, Jean M. Davidson, Venkat S. Malladi, J. Seth Strattan, Benjamin C. Hitz, Idan Gabdank, et al. 2016. "ENCODE Data at the ENCODE Portal." *Nucleic Acids Research* 44 (D1): D726–32.

Smale, S. T., and D. Baltimore. 1989. "The 'Initiator' as a Transcription Control Element." *Cell* 57 (1): 103–13.

Smith, Cory J., Oscar Castanon, Khaled Said, Verena Volf, Parastoo Khoshakhlagh, Amanda Hornick, Raphael Ferreira, et al. 2020. "Enabling Large-Scale Genome Editing at Repetitive Elements by Reducing DNA Nicking." *Nucleic Acids Research* 48 (9): 5183–95.

Smolka, Moritz, Luis F. Paulin, Christopher M. Grochowski, Dominic W. Horner, Medhat Mahmoud, Sairam Behera, Ester Kalef-Ezra, et al. 2023. "Comprehensive Structural Variant Detection: From Mosaic to Population-Level." *bioRxiv*. https://doi.org/10.1101/2022.04.04.487055.

Song, Myungjae, Hui Kwon Kim, Sungtae Lee, Younggwang Kim, Sang-Yeon Seo, Jinman Park, Jae Woo Choi, et al. 2020. "Sequence-Specific Prediction of the Efficiencies of Adenine and Cytosine Base Editors." *Nature Biotechnology* 38 (9): 1037–43.

Stephens, Philip J., Chris D. Greenman, Beiyuan Fu, Fengtang Yang, Graham R. Bignell, Laura J. Mudie, Erin D. Pleasance, et al. 2011. "Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development." *Cell* 144 (1): 27–40.

Stergachis, Andrew B., Brian M. Debo, Eric Haugen, L. Stirling Churchman, and John A. Stamatoyannopoulos. 2020. "Single-Molecule Regulatory Architectures Captured by Chromatin Fiber Sequencing." *Science* 368 (6498): 1449–54.

Sternberg, N., and D. Hamilton. 1981. "Bacteriophage P1 Site-Specific Recombination. I. Recombination between loxP Sites." *Journal of Molecular Biology* 150 (4): 467–86.

Sudmant, Peter H., Tobias Rausch, Eugene J. Gardner, Robert E. Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, et al. 2015. "An Integrated Map of Structural Variation in 2,504 Human Genomes." *Nature* 526 (7571): 75–81.

Sullivan, Patrick F., Jennifer R. S. Meadows, Steven Gazal, Badoi N. Phan, Xue Li, Diane P. Genereux, Michael X. Dong, et al. 2023. "Leveraging Base-Pair Mammalian Constraint to Understand Genetic Variation and Human Disease." *Science* 380 (6643): eabn2937.

Tabula Sapiens Consortium, Robert C. Jones, Jim Karkanias, Mark A. Krasnow, Angela Oliveira Pisco, Stephen R. Quake, Julia Salzman, et al. 2022. "The Tabula Sapiens: A Multiple-Organ, Single-Cell Transcriptomic Atlas of Humans." *Science* 376 (6594): eabl4896.

Takahashi, Kazutoshi, Koji Tanabe, Mari Ohnuki, Megumi Narita, Tomoko Ichisaka, Kiichiro Tomoda, and Shinya Yamanaka. 2007. "Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors." *Cell* 131 (5): 861–72.

Tan, Evan, Cara Sze Hui Chin, Zhi Feng Sherman Lim, and Say Kong Ng. 2021. "HEK293 Cell Line as a Platform to Produce Recombinant Proteins and Viral Vectors." *Frontiers in Bioengineering and Biotechnology* 9 (December): 796991.

Thomas, C. A., Jr. 1971. "The Genetic Organization of Chromosomes." *Annual Review of Genetics* 5: 237–56.

Thomas, Henry, Songjie Feng, Marie Huber, Vincent Loubiere, Daria Vanina, Mattia Pitasi, Alexander Stark, and Christa Buecker. 2023. "Enhancer Cooperativity Can Compensate for Loss of Activity over Large Genomic Distances." *bioRxiv*. https://doi.org/10.1101/2023.12.06.570399.

Thomas, J. W., J. W. Touchman, R. W. Blakesley, G. G. Bouffard, S. M. Beckstrom-Sternberg, E. H. Margulies, M. Blanchette, et al. 2003. "Comparative Analyses of Multi-Species Sequences from Targeted Genomic Regions." *Nature* 424 (6950): 788–93.

Tian, Xueying, and Bin Zhou. 2021. "Strategies for Site-Specific Recombination with High Efficiency and Precise Spatiotemporal Resolution." *The Journal of Biological Chemistry* 296 (March): 100509.

Trojan, Joerg, Stefan Zeuzem, Ann Randolph, Christine Hemmerle, Angela Brieger, Jochen Raedle, Guido Plotz, Josef Jiricny, and Giancarlo Marra. 2002. "Functional Analysis of hMLH1 Variants and HNPCC-Related Mutations Using a Human Expression System." *Gastroenterology* 122 (1): 211–19.

Tubio, Jose M. C., Yilong Li, Young Seok Ju, Inigo Martincorena, Susanna L. Cooke, Marta Tojo, Gunes Gundem, et al. 2014. "Mobile DNA in Cancer. Extensive Transduction of Nonrepetitive DNA Mediated by L1 Retrotransposition in Cancer Genomes." *Science* 345 (6196): 1251343.

Uda, Manuela, Renzo Galanello, Serena Sanna, Guillaume Lettre, Vijay G. Sankaran, Weimin Chen, Gianluca Usala, et al. 2008. "Genome-Wide Association Study Shows BCL11A Associated with Persistent Fetal Hemoglobin and Amelioration of the Phenotype of Beta-Thalassemia." *Proceedings of the National Academy of Sciences of the United States of America* 105 (5): 1620–25.

Vanderstichele, Thomas, Katie L. Burnham, Niek de Klein, Manuel Tardaguila, Brittany Howell, Klaudia Walter, Kousik Kundu, et al. 2023. "Misexpression of Inactive Genes in Whole Blood Is Associated with Nearby Rare Structural Variants." *bioRxiv*. https://doi.org/10.1101/2023.11.17.567537.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, et al. 2001. "The Sequence of the Human Genome." *Science* 291 (5507): 1304–51.

Vinogradov, Alexander E. 2003. "Selfish DNA Is Maladaptive: Evidence from the Plant Red List." *Trends in Genetics: TIG* 19 (11): 609–14.

Wadia, Jehangir S., Radu V. Stan, and Steven F. Dowdy. 2004. "Transducible TAT-HA Fusogenic Peptide Enhances Escape of TAT-Fusion Proteins after Lipid Raft Macropinocytosis." *Nature Medicine* 10 (3): 310–15.

Wang, Jinlin, Zhou He, Guoquan Wang, Ruiwen Zhang, Junyi Duan, Pan Gao, Xinlin Lei, et al. 2022. "Efficient Targeted Insertion of Large DNA Fragments without DNA Donors." *Nature Methods* 19 (3): 331–40.

Wang, Juan, Ze-Xiong Xie, Yuan Ma, Xiang-Rong Chen, Yao-Qing Huang, Bo He, Bin Jia, Bing-Zhi Li, and Ying-Jin Yuan. 2018. "Ring Synthetic Chromosome V SCRaMbLE." *Nature Communications* 9 (1): 3783.

Wang, Yuezhu, Chao Song, Jun Zhao, Yuexin Zhang, Xilong Zhao, Chenchen Feng, Guorui Zhang, et al. 2023. "SEdb 2.0: A Comprehensive Super-Enhancer Database of Human and Mouse." *Nucleic Acids Research* 51 (D1): D280–90.

Waring, M., and R. J. Britten. 1966. "Nucleotide Sequence Repetition: A Rapidly Reassociating Fraction of Mouse DNA." *Science* 154 (3750): 791–94.

Weisheit, Isabel, Joseph A. Kroeger, Rainer Malik, Julien Klimmt, Dennis Crusius, Angelika Dannert, Martin Dichgans, and Dominik Paquet. 2020. "Detection of Deleterious On-Target Effects after HDR-Mediated CRISPR Editing." *Cell Reports* 31 (8): 107689.

Weller, Juliane, Ananth Pallaseni, Jonas Koeppel, and Leopold Parts. 2023. "Predicting Mutations Generated by Cas9, Base Editing, and Prime Editing in Mammalian Cells." *The CRISPR Journal* 6 (4): 325–38.

Winzeler, E. A., D. D. Shoemaker, A. Astromoff, H. Liang, K. Anderson, B. Andre, R. Bangham, et al. 1999. "Functional Characterization of the S. Cerevisiae Genome by Gene Deletion and Parallel Analysis." *Science* 285 (5429): 901–6.

Wolff, Jonas Holst, Jakob Haldrup, Emil Aagaard Thomsen, Sofie Andersen, and Jacob Giehm Mikkelsen. 2021. "piggyPrime: High-Efficacy Prime Editing in Human Cells Using piggyBac-Based DNA Transposition." *Frontiers in Genome Editing* 3 (November): 786893.

Wu, Yi, Rui-Ying Zhu, Leslie A. Mitchell, Lu Ma, Rui Liu, Meng Zhao, Bin Jia, et al. 2018. "In Vitro DNA SCRaMbLE." *Nature Communications* 9 (1): 1935.

Xia, Bo, Weimin Zhang, Aleksandra Wudzinska, Emily Huang, Ran Brosh, Maayan Pour, Alexander Miller, et al. 2021. "The Genetic Basis of Tail-Loss Evolution in Humans and Apes." *bioRxiv*. https://doi.org/10.1101/2021.09.14.460388.

Xu, Zhengyao, Louise Thomas, Ben Davies, Ronald Chalmers, Maggie Smith, and William Brown. 2013. "Accuracy and Efficiency Define Bxb1 Integrase as the Best of Fifteen Candidate Serine Recombinases for the Integration of DNA into the Human Genome." *BMC Biotechnology* 13 (October): 87.

Yang, Luhan, Marc Güell, Dong Niu, Haydy George, Emal Lesha, Dennis Grishin, John Aach, et al. 2015.

"Genome-Wide Inactivation of Porcine Endogenous Retroviruses (PERVs)." *Science* 350 (6264): 1101–4.

Yarnall, Matthew T. N., Eleonora I. Ioannidi, Cian Schmitt-Ulms, Rohan N. Krajeski, Justin Lim, Lukas Villiger, Wenyuan Zhou, et al. 2022. "Drag-and-Drop Genome Insertion of Large Sequences without Double-Strand DNA Cleavage Using CRISPR-Directed Integrases." *Nature Biotechnology*, November. https://doi.org/10.1038/s41587-022-01527-4.

Yi, Eunhee, Rocío Chamorro González, Anton G. Henssen, and Roel G. W. Verhaak. 2022. "Extrachromosomal DNA Amplifications in Cancer." *Nature Reviews. Genetics* 23 (12): 760–71.

Yu, Goosang, Hui Kwon Kim, Jinman Park, Hyunjong Kwak, Yumin Cheong, Dongyoung Kim, Jiyun Kim, Jisung Kim, and Hyongbum Henry Kim. 2023. "Prediction of Efficiencies for Diverse Prime Editing Systems in Multiple Cell Types." *Cell* 186 (10): 2256–72.e23.

Yusa, Kosuke, Liqin Zhou, Meng Amy Li, Allan Bradley, and Nancy L. Craig. 2011. "A Hyperactive piggyBac Transposase for Mammalian Applications." *Proceedings of the National Academy of Sciences of the United States of America* 108 (4): 1531–36.

Yu, Y., and A. Bradley. 2001. "Engineering Chromosomal Rearrangements in Mice." *Nature Reviews. Genetics* 2 (10): 780–90.

Zhang, Cheng-Zhong, Alexander Spektor, Hauke Cornils, Joshua M. Francis, Emily K. Jackson, Shiwei Liu, Matthew Meyerson, and David Pellman. 2015. "Chromothripsis from DNA Damage in Micronuclei." *Nature* 522 (7555): 179–84.

Zhang, Guiquan, Yao Liu, Shisheng Huang, Shiyuan Qu, Daolin Cheng, Yuan Yao, Quanjiang Ji, Xiaolong Wang, Xingxu Huang, and Jianghuai Liu. 2022. "Enhancement of Prime Editing via xrRNA Motif-Joined pegRNA." *Nature Communications* 13 (1): 1856.

Zhang, Jing, Donghoon Lee, Vineet Dhiman, Peng Jiang, Jie Xu, Patrick McGillivray, Hongbo Yang, et al. 2020. "An Integrative ENCODE Resource for Cancer Genomics." *Nature Communications* 11 (1): 3696.

Zhao, Yu, Camila Coelho, Amanda L. Hughes, Luciana Lazar-Stefanita, Sandy Yang, Aaron N. Brooks, Roy S. K. Walker, et al. 2023. "Debugging and Consolidating Multiple Synthetic Chromosomes Reveals Combinatorial Genetic Interactions." *Cell* 186 (24): 5220–36.e16.

Zhao, Zhihu, Gholamreza Tavoosidana, Mikael Sjölinder, Anita Göndör, Piero Mariano, Sha Wang, Chandrasekhar Kanduri, et al. 2006. "Circular Chromosome Conformation Capture (4C) Uncovers Extensive Networks of Epigenetically Regulated Intra- and Interchromosomal Interactions." *Nature Genetics* 38 (11): 1341–47.

Zheng, B., M. Sage, E. A. Sheppeard, V. Jurecic, and A. Bradley. 2000. "Engineering Mouse Chromosomes with Cre-loxP: Range, Efficiency, and Somatic Applications." *Molecular and Cellular Biology* 20 (2): 648–55.

Zhou, Sijie, Yi Wu, Yu Zhao, Zhen Zhang, Limin Jiang, Lin Liu, Yan Zhang, Jijun Tang, and Ying-Jin

Yuan. 2022. “Synthetic Genome Rearrangement Reveals Dynamics of Chromosome Evolution Shaped by Hierarchical Chromatin Organization.” *bioRxiv*. https://doi.org/10.1101/2021.07.19.453002.

Zielenski, J., and L. C. Tsui. 1995. “Cystic Fibrosis: Genotypic and Phenotypic Variations.” *Annual Review of Genetics* 29: 777–807.

Zou, Roger S., Alberto Marin-Gonzalez, Yang Liu, Hans B. Liu, Leo Shen, Rachel K. Dveirin, Jay X. J. Luo, Reza Kalhor, and Taekjip Ha. 2022. “Massively Parallel Genomic Perturbations with Multi-Target CRISPR Interrogates Cas9 Activity and DNA Repair at Endogenous Sites.” *Nature Cell Biology* 24 (9): 1433–44.

Zuin, Jessica, Jesse R. Dixon, Michael I. J. A. van der Reijden, Zhen Ye, Petros Kolovos, Rutger W. W. Brouwer, Mariëtte P. C. van de Corput, et al. 2014. “Cohesin and CTCF Differentially Affect Chromatin Architecture and Gene Expression in Human Cells.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (3): 996–1001.

Zuin, Jessica, Gregory Roth, Yinxiu Zhan, Julie Cramard, Josef Redolfi, Ewa Piskadlo, Pia Mach, et al. 2022. “Nonlinear Control of Transcription through Enhancer-Promoter Interactions.” *Nature* 604 (7906): 571–77.